# A Rough-Set-Based Inference Engine for ECG Classification

Sucharita Mitra, *Student Member, IEEE*, Madhuchhanda Mitra, and B. B. Chaudhuri, *Fellow, IEEE*

***Abstract*—In this paper, a rule-based rough-set decision system for the development of a disease inference engine is described. For this purpose, an offline-data-acquisition system of paper electrocardiogram (ECG) records is developed using image-processing techniques. The ECG signals may be corrupted with six types of noise. Therefore, at first, the extracted signals are fed for noise removal. A QRS detector is also developed for the detection of R-R interval of ECG waves. After the detection of this R-R interval, the P and T waves are detected based on a syntactic approach. The isoelectric-level detection and base-line correction are also implemented for accurate computation of different attributes of P, QRS, and T waves. A knowledge base is developed from different medical books and feedbacks of reputed cardiologists regarding ECG interpretation and essential time-domain features of the ECG signal. Finally, a rule-based rough-set decision system is generated for the development of an inference engine for disease identification from these time-domain features.**

***Index Terms*—Decision system, electrocardiogram (ECG), feature extraction, inference engine, knowledge base, rough set, rule based, time domain.**

## I. Introduction

In 1887, Waller recorded the first electrocardiogram (ECG) by capillary electrometer. The method was improved by the introduction of the string galvanometer discovered by Einthoven. The string galvanometer was replaced by electronic amplifiers, and then a direct writing recorder or paper plotter is introduced to get ECG strips.

ECGs are electrical views of the heart, recorded by placing electrodes on the patient's body. The ECG provides the cardiologist with useful information about the rhythm and functioning of the heart. Beats extracted from ECG signals can be categorized in a range of classes. Every ECG consists of three distinct waves denoted by P, QRS, and T (Fig. 1).

In recent years, considerable research has been done to assist cardiologists with their task of diagnosing the ECG recordings. The fields of research range from novelty detectors to fully automated ECG-diagnosing systems. A wide range of techniques have been used for the purpose that include statistical pattern recognition, expert systems, artificial neural networks (ANN), wavelet transform, and fuzzy and neuro-fuzzy systems.

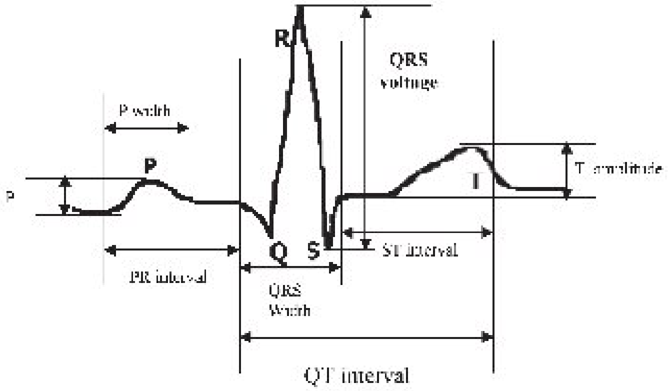


Fig. 1. Different time-plane features that are extracted.

The morphological diagnosis of the ECG is a pattern-recognition task [1]. The ECG interpretation is comprised of two distinct and sequential phases: feature extraction and classification. A set of signal measurements containing information for the characterization of the waveform are obtained by shape-identification methods. These waveform descriptors are used to allocate the ECG to one or more diagnostic classes in the classification phase. These classifiers may be heuristic and use rules-of-thumb or employ syntactic or fuzzy logic as reasoning tools. The classifier may also be statistical with the use of complex and even abstract-signal features as waveform descriptors and different discriminate function models for class allocation.

An approach based on ECG ST-T segment analysis has been proposed for intelligent ischemia-event detection [25]. ST-T trends are processed by means of a Bayesian forecasting approach using the multistate Kalman filter. Hermite functions and self-organizing maps are used for clustering ECG complexes [24]. The automated-analysis technology of ECGs was summarized for existing problems and with possible ways to solve those troubles [27].

More recently, ANN techniques have also been used for signal classification [6]. Learning algorithms for two-phase and three-phase radial-basis function (RBF) networks are proposed for the categorization of high-resolution ECGs. The ECG records are given as a time series and as a set of features extracted from these time series [26]. Bi-group neural network (NN) classifiers are also used to examine independent feature vectors of ECG recordings for each diagnostic class, and the outputs from all classifiers are fused together to produce a composite result [7].

A classifier has been developed based on wavelet transforms for extracting the features, and then, using a RBFNN to classify the arrhythmia [4]. Moreover, fuzzy approaches are

Manuscript received December 29, 2005; revised August 7, 2006. This work was supported by the Council of Scientific and Industrial Research, Government of India.

S. Mitra and M. Mitra are with the Department of Applied Physics, Faculty of Technology, University of Calcutta, Calcutta, India (e-mail: susa68@hotmail.com; mmaphy@caluniv.ac.in).

B. B. Chaudhuri is with the Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Calcutta, India (e-mail: bbc@isical.ac.in).

Color versions for Figs. 3–5 are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIM.2006.884279

proposed for computerized-ECG diagnosis and classification [3], [5]. Fuzzy Adaptive-Resonance Theory MAP has also been employed to classify cardiac arrhythmia [2]. A hybrid neuro-fuzzy system was used for ECG classification of myocardial infarction (MI) [8].

For the past few years, rough-set theory and granular computation have emerged as other soft-computing tools, which in various synergetic combinations with fuzzy logic, ANN and genetic algorithms, provide a stronger framework to achieve tractability, low-cost solution, robustness, and close resemblance with humanlike decision making. To describe different concepts or classes, crude domain knowledge in the form of rules is extracted with the help of rough neural-synergistic integration and encoded as network parameters. For this purpose, an initial knowledge-base network is built for efficient learning. In case of granular computation, each operation is done on granules (clump of similar objects or points), rather than on the individual data points. As a result, the computation time is substantially reduced. As the methodology has matured, several interesting applications of the theory have surfaced to be applied also in medicine. Pawlak [12] used rough-set theory in the Bayes' theorem and showed that it can be applied for generating rule base to identify the presence or absence of a disease. Discrete wavelet transform and rough-set theory have been used for classification of arrhythmia [9].

The main contribution of this paper is the development of a rule-based rough-set decision system to generate an inference engine for ECG classification from different standard time-plane features.

## II. Rough Set—A Tool for Representing and Reasoning About Imprecise or Uncertain Information

The theory of rough sets, introduced by Pawlak [10], [11] in 1982, has emerged as an important mathematical tool for managing uncertainty that arises from inexact, noisy, or incomplete information. It is methodologically significant to the domains of artificial intelligence and cognitive sciences, especially in the representation of reasoning with vague or imprecise knowledge, data classification, rule generation, machine learning, data mining, and knowledge discovery. The theory has substantial influence in many other areas of applications.

A brief description of the rough-set theory and its use as a classifier is given below.

### *A. Mathematical Basics of Rough-Set Theory*

As described by Pawlak [18], the basic concepts of rough-set theory are information system and approximation of sets. An information system is a data table, whose columns are labeled by attributes, rows are labeled by objects of interest, and entries of the table are attribute values. Formally, an information system may be expressed as a pair $S = (U, A)$, where $U$ and $A$ are finite nonempty sets called the universe and the set of attributes, respectively. Every attribute $a \in A$ is associated with a set $V a$, of its values, called the domain of $a$. Any subset $B$ of $A$ determines a binary relation $I(B)$ on $U$, which will be called an indiscernibility relation. An information system may be distinguished as two disjoint classes of attributes, called condition and decision attributes, respectively. Then, the system will be called a decision table and will be denoted by $S = (U, C, D)$, where $C$ and $D$ are disjoint sets of condition and decision attributes, respectively. Thus, the decision table determines decisions, which must be taken when some conditions are satisfied. Let us suppose an information system $S = (U, A)$, $X \subseteq U$, and $B \subseteq A$. The task is to describe the set $X$ in terms of attribute values from $B$. To this end, it defines two operations assigning to every $X \subseteq U$ two sets of $B_*(X)$ and $B^*(X)$, called the $B$-lower and the $B$-upper approximation of $X$, respectively, and are defined as follows:

$$B_*(X) = \bigcup_{x \in U} \{B(x) : B(x) \subseteq X\} \tag{1}$$

$$B^*(X) = \bigcup_{x \in U} \{B(x) : B(x) \cap X \neq \varnothing\}. \tag{2}$$

Hence, the $B$-lower approximation of a set is the union of all $B$ granules that are included in the set, whereas the $B$-upper approximation of a set is the union of all $B$ granules that have a nonempty intersection with the set.

The set $\mathrm{BN}_B(X) = B^*(X) - B_*(X)$ will be referred to as the $B$-boundary region of $X$. If the boundary region of $X$ is the empty set, i.e., $\mathrm{BN}_B(X) = \varnothing$, then $X$ is crisp (exact) with respect to $B$. In the opposite case, i.e., if $\mathrm{BN}_B(X) \neq \varnothing$, $X$ is referred to as rough (inexact) with respect to $B$.

### *B. Rough Set as A Classifier*

Rough-set theory is mainly used for getting the optimal number of appropriate rules that are used for developing a classifier. From every information system, a subset of minimal attributes is generated, which is known as reduct. Different algorithms are available to generate rules from this reduct.

Let us describe the decision rules more precisely. Each decision rule of a decision table determines decisions in terms of conditions.

Let $S = (U, C, D)$ be a decision table. Every $x \in U$ determines a sequence $c_1(x), \ldots, c_n(x), d_1(x), \ldots, d_m(x)$, where $\{c_1, \ldots, c_n\} = C$ (conditions), and $\{d_1, \ldots, d_m = D\}$ (decisions).

The sequence will be called a decision rule induced by $x$ (in $S$) and denoted by $c_1(x), \ldots, c_n(x) \to d_1(x), \ldots, d_m(x)$ or, in short, $C \to_x D$.

The term $\mathrm{supp}_x(C, D) = |C(x) \cap D(x)|$ is called a support of the decision rule $C \to_x D$, and the number $\sigma_x(C, D) = (\mathrm{supp}_x(C, D))/|U|$ will be referred to as the strength of the decision rule $C \to_x D$. Every decision rule $C \to_x D$ is associated with the certainty factor of the decision rule, which is denoted as $\mathrm{cer}_x(C, D)$ and defined as follows:

$$\mathrm{cer}_x(C, D) = \frac{|C(x) \cap D(x)|}{|C(x)|} = \frac{\mathrm{supp}_x(C, D)}{|C(x)|} = \frac{\sigma_x(C, D)}{\pi\,(C(x))} \tag{3}$$

where

$$\pi(C(x)) = \frac{|C(x)|}{|U|}.$$

The certainty factor may be interpreted as a conditional probability that $y$ belongs to $D(x)$ given $y$ belongs to $C(x)$, symbolically defined as $\pi_x(D|C)$. If $\text{cer}_x(C, D) = 1$ then, $C \rightarrow_x D$ will be called a certain decision rule in $S$. If $0 < \text{cer}_x(C, D) < 1$, the decision rule will be referred to as an uncertain decision rule in $S$. Besides, a coverage factor of the decision rule is also used and denoted as $\text{cov}_x(C, D)$ and is defined as

$$\text{cov}_x(C, D) = \frac{|C(x) \cap D(x)|}{|D(x)|} = \frac{\text{supp}_x(C, D)}{|D(x)|} = \frac{\sigma_x(C, D)}{\pi(D(x))} \quad (4)$$

where

$$\pi(D(x)) = \frac{|D(x)|}{|U|}.$$

Similarly, $\text{cov}_x(C, D) = \pi_x(C|D)$. If $C \rightarrow_x D$ is a decision rule then, $D \rightarrow_x C$ will be called an inverse decision rule. The inverse decision rules can be used to give explanations (reasons) for decisions. Decision rules are often represented in a form of "if ... then ..." implications. Thus, any decision table can be transformed in a set of "if ... then ..." rules called a decision algorithm. Using this decision algorithm, the optimal rule set is generated, which are used for development of the rule-based classifier.

### C. Qualitative Comparison With Other Approaches

Rough-set theory addresses the problem where the objects cannot always be assigned to a class crisply. Sometimes, the classes overlap, or it is unclear to which class an object should belong. In such a situation, classifying objects is not a 0–1 problem. This is true for data in the real world, especially in the medical domain.

Rough-set theory is equipped to handle such inconsistent or seemingly conflicting or vague examples of the data. Inconsistencies may occur due to, e.g., transcription errors, subjective determination of attribute values or outcomes, lack of information, etc. The theory of rough sets can handle any finite number of outcome categories and not just dichotomous outcomes.

On the other hand, using rough sets is advantageous over black box-type classification schemes such as NN. With the rough-set approach, the generated rules are visible and available for the user, making the extraction of more information from the rules possible. Since the strength of a rule can also be calculated, this makes the rough set a valuable tool for classification.

A guiding philosophy of rough-set theory is to let the data material speak for itself. As such, very few assumptions are made about the data attributes. It needs only some notion of inequality defined on the data domain. In particular, rough-set theory does not make assumptions about statistical distributions of the data, nor does it need external information such as, e.g., membership functions for fuzzy sets, weighted inputs for NN, or density function in statistical classifications.

The output of a rough-set analysis is usually a collection of if–then rules. An if–then rule is, arguably, as close to a model in natural language as one might expect to obtain and can be read and interpreted by personnel without expertise in the actual model-induction technique. Construction of the knowledge base is commonly perceived as a major bottleneck in building rule-based expert systems. As rough sets can be used to automatically induce if–then rules from empirical data, this offers the possibility to automate, at least in part, the knowledge-acquisition stage for developing such systems.

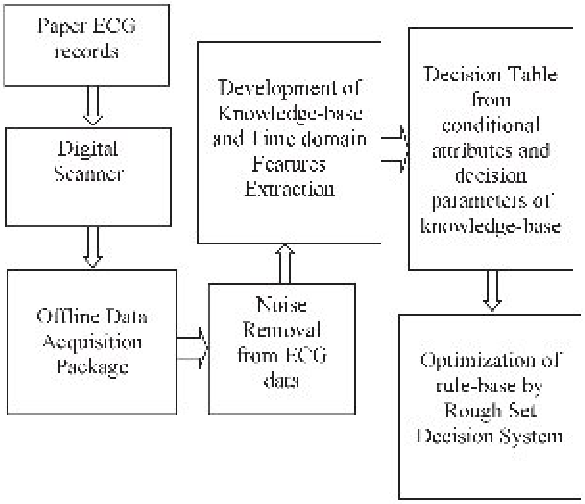

Paper ECG records
Digital Scanner
Offline Data Acquisition Package
Noise Removal from ECG data
Development of Knowledge-base and Time domain Features Extraction
Decision Table from conditional attributes and decision parameters of knowledge-base
Optimization of rule-base by Rough Set Decision System

Fig. 2. Block diagram of the proposed system.

All the aforementioned reasons make rough-set more suitable than the other existing tools in development and optimization of rule base and knowledge base or in task of classification for the data in medical province.

The suitability of rough sets applied in the medical domain has already been investigated by, e.g., Tsumoto [37], [38], who discussed some characteristics of medical reasoning and argued that rough-set representation of diagnostic models is a useful approach for extracting medical knowledge from databases. Tsaptsinos and Bell [39] also argue along similar lines.

Hence, we select a rough-set decision system to optimize the rule base for ECG-data classification to test the utility of this most recent and popular soft-computing tool for getting a low-cost, more robust, and humanlike decision.

## III. Materials and Method

The block diagram of the developed system is given in Fig. 2. The different steps for developing the system are described below in sequential order.

### A. Data-Acquisition System Development

Digital-time databases from offline paper ECG records are developed in this module [GUI based]. For this purpose, a flatbed scanner (HP Scanjet 2300C) is used to capture the image of each ECG signal recorded on a single-channel chart recorder (voltage in millivolts versus time in seconds). These TIF-formatted gray-tone scanned images are then converted into two-tone binary images by binarization technique using a histogram-analysis process. It is known that after scanning, we obtain gray-tone images having pixel-gray value in the range of 0–255. To detect the black or nearly black pixels, a threshold is chosen from the peak nearest to 0-pixel level of the histogram. All pixels above the threshold got a value (say 0), and the rest,

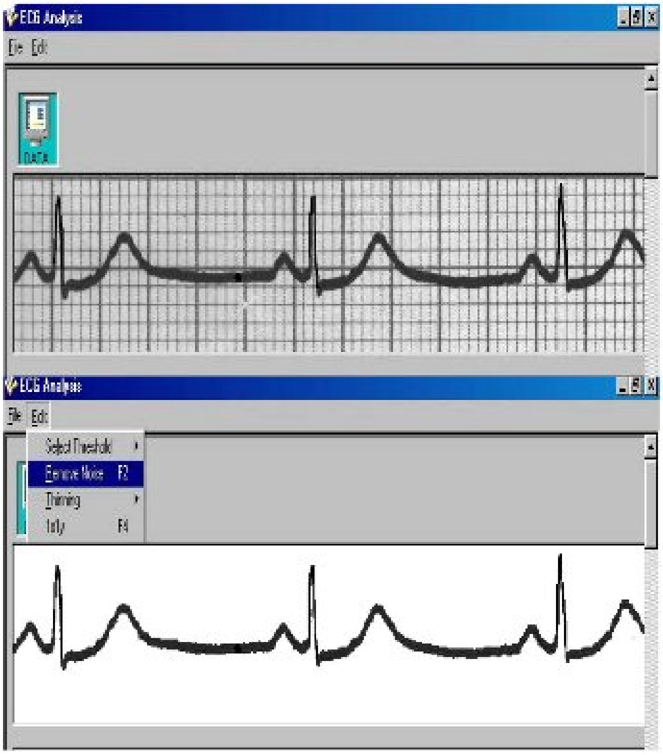


Fig. 3. (Upper) Original paper ECG record image and (lower) ECG signal after removal of background noise.

i.e., the black pixels, were assigned the value 1, resulting in a two-tone or binary image.

Proper selection of the threshold mostly removes the background noise, but some noise dots are still present in a few cases. Therefore, for extracting the region of interest (i.e, the ECG signals), a connected component algorithm [13] (8-neighbor connected) has been used. The resulting image is then ported to the next part of the automated system, where skeletonizaton or thinning of the ECG signal is performed. Thinning of the input images is necessary to avoid repetition of coordinate information in the data set of the normalized database. For this purpose, a thinning algorithm [13] is used. The method consists of successive passes of two basic steps applied to the contour point of the given region. To keep the description brief, the details of the algorithms are not given here. In the next step, the raw database in ASCII format is generated. These data are then sorted by using bubble-sort algorithm and ported to the regeneration domain of the system. The captured pattern is compared with the original waveform with the help of this regeneration module of our system (Figs. 3 and 4). A time (in seconds) versus amplitude (in millivolts) data file is obtained for each of the 12-lead ECG signals after each processing.

The present database contains 100 normal and 100 diseased subjects out of which, 55 patients have acute MI, and the other 45 patients have myocardial ischemia.

In this paper, ECG records are collected mainly from the Institute of Post Graduate Medical Education and Research (IPGMER), Calcutta. Some records are also collected from Peerless Hospital and some other ECG clinics of Calcutta. All the reports are diagnosed by the doctors from their experience and knowledge base. All subjects were in the age group of 40–66 years. Most of them were male.

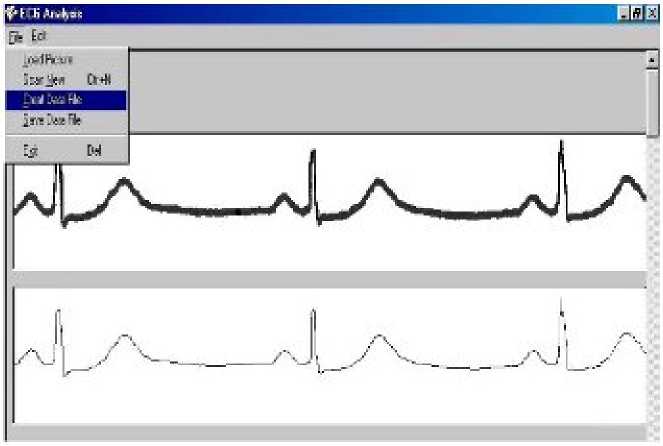


Fig. 4. ECG signal after thinning.

### B. Noise Removal From ECG Signals

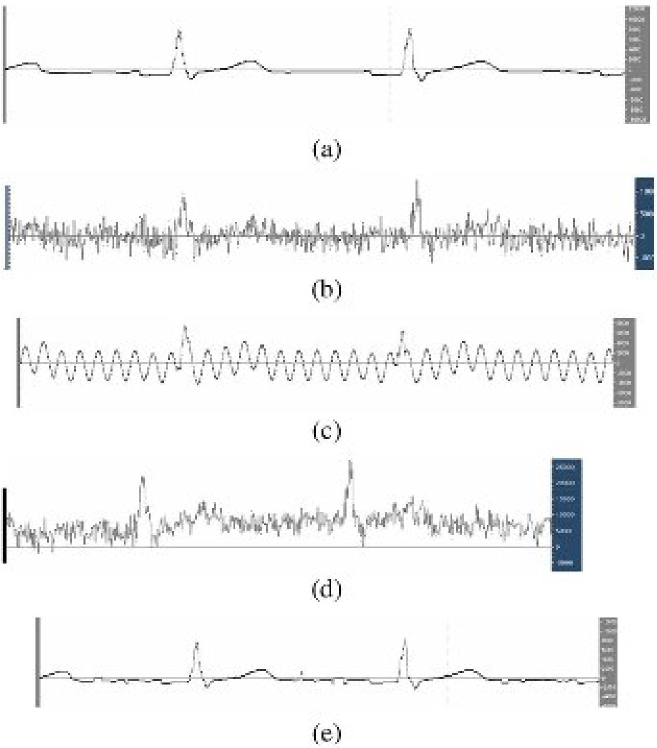
ECG signals may be corrupted by different types of noise [34]. Typical examples of the noise are 1) power-line interference; 2) electrode-contact noise; 3) motion artifacts; 4) muscle contraction (electromyograph, EMG); 5) baseline drift and ECG-amplitude modulation with respiration; and 6) electrosurgical noise.

Our database contained a few ECG data corrupted by power-line interference and baseline shift. All types of noises are simulated up to certain level and removed from the signal by a software package Cool Edit Pro offered by Syntrillium Software Corporation. This was done to get a realistic situation for the algorithm and to get more accuracy in time-domain features detection. All types of noise levels are varied from 10% to 30%, and the generated filters provided satisfactory response in all cases (Fig. 5).

### C. Knowledge-Base Development and Time-Domain Features Extraction

A knowledge base regarding ECG interpretation has been developed from the opinion of reputed cardiologists of different hospitals and clinical centers. For this purpose, we selected 20 doctors and presented to them different sample questions about ECG interpretation. From their feedbacks and after consultation of different medical books [21]–[23], we have selected 12 time-plane features for disease identification, as listed below:

1) heart rate;
2) PR interval;
3) P-wave height;
4) P-wave width;
5) QRS width;
6) QRS voltage;
7) QTc = (QT interval/Sqrt R − R interval)
8) Abnormal Q wave;
9) R-wave Progression;


(a)

(b)

(c)

(d)

(e)

Fig. 5. Original ECG signal, signal corrupted with different noises, and filtered output signal.

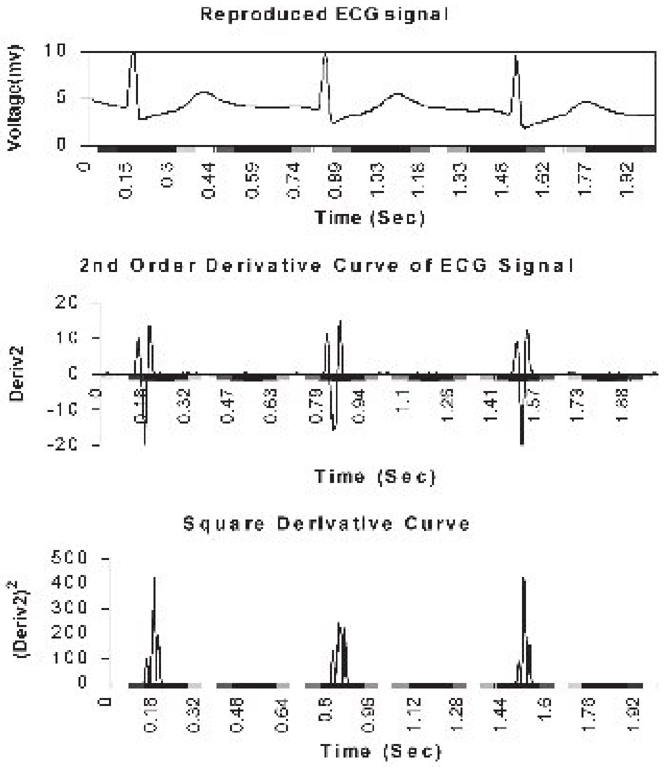


Fig. 6. QRS complex or R-R interval detection.

10) ST segment;
11) Reciprocity in T wave;
12) T wave.

To extract the time-based features from ECG signals, accurate detection of the R-R interval between two consecutive ECG waves is very important. For this purpose, the second-order derivative of the captured signal is computed by using five-point Lagrangian interpolation formulas for differentiation [14] given as follows:

$$f_0' = \frac{1}{12h}(f_{-2} - 8f_{-1} + 8f_1 - f_2) + \frac{h^4}{30} f^v(\xi). \qquad (5)$$

Here, $\xi$ lies between the extreme values of the abscissas involved in the formula. After squaring the values of the second-order derivative, a square-derivative curve having only high-positive peaks of small width at the QRS complex region can be obtained (Fig. 6). A small window of length (say $W$) was taken to detect the area of this curve, and we obtained maximum area at those peak regions. The local maxima of these peak regions are considered as R-peak. For this paper, $W$ is set as ~0.07 s. The system was tested for both noise-free and noisy signals. The levels of all types of noise were increased from 0% to 30%, and still, we achieved 99.4% accuracy in the detection of QRS complexes.

In order to accurately detect the P wave and ST segments, isoelectric line must be correctly identified. Most methods are based upon the assumption that the isoelectric level of the signal lies in the region ~80 ms left of the R-peak, where the first derivative becomes equal.

In particular, let $y_1$, $y_2,\ldots$, $y_n$ be the samples of a beat, $y_1'$, $y_2',\ldots$, $y_{n-1}'$ be their first differences and $y_r$, which is the sample where the R-peak occurs. The isoelectric-level samples $y_b$ are then defined if either of the two following criteria is satisfied:

$$\begin{aligned}
&|y'_{r-j-\mathrm{int}(0.08f)}| = 0, \qquad j = 1, 2, \ldots, 0.01f \text{ or}\\
&|y'_{r-j-\mathrm{int}(0.08f)}| \le |y'_{r-i-\mathrm{int}(0.08f)}|, \quad i, j = 1, 2, \ldots, 0.02f\\
&[\text{For positive slope } j < i, \text{for negative slope j} > i] \qquad (6)
\end{aligned}$$

where $f$ is the sampling frequency. After the isoelectric level is found, and after comparing the current beat with the previous corrected one $(y_{bp}^{p})$, it is easy to align the current beat with the previous one, using the declination of the line connecting the isoelectric levels of the two beats. If we define

$$\gamma = \frac{y_b - y_{b_p}^{p}}{n_b}$$

where $n_b$ is the number of samples between the two baseline points; the alignment procedure becomes

$$y_t \to \gamma y_t. \qquad (7)$$

Other baseline-correction techniques rely on adaptive filtering of the ECG beat and provide reliable results as well. However, the QRS-wave shape is corrupted by the use of such techniques [31].

After detection of baseline, the location of the P wave is determined from the first derivative of the samples.

The R wave can be detected very reliably, and for this reason, it is used as the starting point for ST-segment processing and for T-wave detection. In most algorithms dealing with ST-segment processing, it is assumed that in normal sinus rhythm, the ST

TABLE I
PORTION OF DECISION TABLE

| Heart Rate | PR Interval | P wave height | P wave width | QRS width | QTc | QRS voltage | R Wave Prog | Abn. Q waves | ST Seg-Ment | Reci-Pro-city | T waves | Disease |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| String | String | String | String | String | String | String | String | String | String | String | String | String |
| N | N | N | N | N | N | N | A | P | E | P | – | MI |
| B | N | N | N | A | N | N | A | P | D | P |  | MI |
| N | A | N | N | A | N | N | A | P | E | P | - | MI |
| N | N | N | N | N | N | N | N | Ab | I | Ab | + | N |
| N | N | N | N | N | N | N | N | Ab | I | Ab | + | N |
| N | N | N | N | N | N | N | N | Ab | I | Ab | + | N |
| N | A | N | N | N | N | N | N | Ab | E | Ab |  | ISC |
| N | N | N | N | N | N | N | N | Ab | D | Ab |  | ISC |
| N | N | N | N | N | N | N | N | Ab | E | Ab | -- | ISC |

where N→ Normal, A→ Abnormal, P→ Present, AB→ Absent, B→ Bradycardia,
T→ Tachycardia, E→ Elevated, D→ Depressed, I→ Isoelectric, MI→ Myocardial
Infarction, ISC → Ischemia.

segment begins at 60 ms after the R peak. In case of tachycardia (R-R interval $< 600$ ms), the beginning of the ST segment is marked at 40 ms after the R peak. The ST-segment duration has beat-to-beat variability, but since this is not easily determined, many algorithms assume that ST has a predefined length of 160 ms (this means that the end point is 220 ms after R peak in the normal case and 200 ms, otherwise).

Other algorithms follow the Bazzet formula that links the ST-segment duration with the R-R-interval duration. The aforementioned ST-segment limits are more or less in general agreement with the recommendation of the European ST-T database and with the observations in [15]–[17].

Our algorithm adopted the first assumption, and once getting the beginning, it computes the slope of the ST segments and also detects the zero crossings. Depending on the zero crossings and shape of each wave, a syntactic approach is developed for detection of P, Q, R, S, and T waves. The area of P, QRS, and T waves are also computed. In getting the QRS complex, we achieved 99.4% accuracy, for T waves, the accuracy was 96.7%, and for P waves, the accuracy obtained was 92.2%.

The performance of the developed QRS detector is satisfactory especially in noisy environment, and we obtained a 99.4% accuracy on an average in detection of QRS, whereas the ANN-based QRS detector gives 99.2%, and the bandpass filtering method gives 97.8% accuracy [28]. The wavelet-based QRS detector (Dy WT) has minimum of 2.6% and maximum of 15.4% error rate [29]. An approach based on mathematical morphology detects QRS with a sensitivity of 99.38% and positive predictability of 99.48% [30]. Therefore, our method compares favorably with the other methods. Moreover, our method is also very simple to compute.

### D. Development of Inference Engine

A rule-based rough-set decision system is generated to develop an inference engine for disease identification from the time-domain feature analysis of ECG signals. The most popular and widely used rough-set software toolbox is ROSETTA [19], [20]; the URL for downloading this is http://www.idi.ntnu.no/~aleks/rosetta/rosetta.html. This software supports different options of generating decision tables, reducts, discretization techniques, classification, and decision algorithms. For this reason, we used this software for this paper. Here, learning samples are processed in the following way. First, a knowledge base is acquired for the data set, and in this particular case, they are the time-domain features that are listed above. The knowledge base consists of objects, which are represented using conditional attributes and decision parameters. All the time-domain features acquired their specific attributes according to different conditional attributes and decision parameters of the knowledge base developed from doctors' opinion and different medical books. They are used as the input parameters of the decision table, a portion of which is given in Table I.

Consequently, the acquired data are quantized to convert real attribute values into discretized form, thus allowing further rule processing. Based on the discrete values, attributes are analyzed in terms of discernibility investigation. Sets of attributes that generate partition of object classes are then revealed. These sets are called reducts.

The ROSETTA system supports a variety of quantization as well as reduct and rule generation procedures. However, the details of these lie beyond the scope of this paper. For this paper, the following processing parameters were used.

1) Equal-frequency binning using three intervals is used for discretization.
2) Object-related genetic algorithms producing a set of rules via minimal-attribute subsets that discern object classes, reducts, and rules, are generated upon analysis of all learning patterns.

These processing parameters were chosen during a preliminary research aimed at optimizing the system efficiency and generation ability.

## IV. EXPERIMENTAL RESULTS

In this paper, a total of 23 rules (partly shown in Table II) are generated. Intuitively, a "strong" rule is both accurate and has a high coverage. The accuracy of a rule reflects how trustworthy its consequence is. A portion of the generated rule set and the confusion matrix, generated using standard voting classifiers, are given below in Tables II and III. We consider both left-hand side (LHS) and right-hand side (RHS) coverage factors for the selection of the optimum rule set. For example, rule 1 in Table II gives the decision according to LHS coverage factor

TABLE II
PORTION OF GENERATED RULE SET

| | Rule | LHS Support | RHS Support | RHS Accuracy | LHS Coverage | RHS Coverage |
|---|---|---|---|---|---|---|
| 1 | Abn_Q_waves(P) => Disease(MI) | 27 | 27 | 1.0 | 0.313953 | 1.0 |
| 2 | Heart_Rate(N) AND ST_segments(E) AND Reciprocity(Ab) => Disease(ISC) | 21 | 21 | 1.0 | 0.244186 | 1.0 |
| 3 | Abn_Q_waves(Ab) AND ST_segments(E) => Disease(ISC) | 21 | 21 | 1.0 | 0.244186 | 1.0 |
| 4 | ST_segments(I) => Disease(N) | 38 | 38 | 1.0 | 0.44186 | 1.0 |
| 5 | Reciprocity(P) => Disease(MI) | 23 | 23 | 1.0 | 0.267442 | 0.851852 |
| 6 | PR_interval(N) AND ST_segments(E) AND Reciprocity(Ab) => Disease(ISC) | 17 | 17 | 1.0 | 0.197674 | 0.809524 |
| 7 | R_wave_Prog(N) AND ST_segments(E) => Disease(ISC) | 16 | 16 | 1.0 | 0.186047 | 0.761905 |
| 8 | PR_interval(A) AND R_wave_Prog(A) => Disease(MI) | 12 | 12 | 1.0 | 0.139535 | 0.444444 |
| 9 | Heart_Rate(B) => Disease(MI) | 10 | 10 | 1.0 | 0.116279 | 0.37037 |
| 10 | Heart_Rate(T) => Disease(N) | 14 | 14 | 1.0 | 0.162791 | 0.368421 |

TABLE III
CONFUSION-MATRIX OUTPUT FOR STANDARD VOTING CLASSIFIER

| | | Predicted (Output) | | | |
|---|---|---|---|---|---|
| | | MI ($c_1$) | N ($c_2$) | ISC ($c_3$) | Pr($r_i$) |
| Actual (Input) | MI ($r_1$) | 27 | 0 | 0 | 1.0 |
| | N ($r_2$) | 0 | 38 | 0 | 1.0 |
| | ISC ($r_3$) | 0 | 0 | 21 | 1.0 |
| | Pr ($c_i$) | 1.0 | 1.0 | 1.0 | 1.0 |
| ROC | Class | N | | | |
| | Area | 1 | | | |
| | Std. error | 0 | | | |
| | Thr. (0, 1) | 0.628 | | | |
| | Thr. acc. | 0.628 | | | |

$\Pr(r_i) = c_i / c_1+c_2+c_3$
Where i = 1 to 3

$\Pr(c_i) = r_i / r_1+r_2+r_3$
Where i = 1 to 3

Sensitivity

TABLE IV
RESULT OBTAINED FROM RULE-BASED ROUGH-SET DECISION SYSTEM

| Type of Samples | No. of Trained Samples | No. of Untrained Samples | Accuracy for Trained Samples | Accuracy for Untrained Samples |
|---|---|---|---|---|
| Normal | 38 | 62 | 100% (38/38) | 100% (62/62) |
| Ischemia | 21 | 24 | 100% (21/21) | 95.8% (23/24) |
| MI | 27 | 28 | 100%(27/27) | 100% (28/28) |

that only 31.4% of the patients having ECG, where an abnormal Q wave is present (P), are suffering from the disease MI. From the inverse decision rule, considering RHS coverage factor, it can be concluded that 100% of the patients suffering from MI have ECG records, where an abnormal Q wave is present. The inverse decision rule provides a stronger explanation of the generated decision, which is mentioned at the end of Section II, where the theoretical background of the rough set as classifier is given. Obviously, rule 4, having the highest LHS and RHS coverage factor, will be the strongest. The first seven rules, having high accuracy and coverage factor (both LHS and RHS), are taken for the generation of the rule-based classifier of the disease. Both trained and untrained samples for all the three sets of data (e.g., normal, Ischemia, and MI) are fed to the inference system, and the result obtained is given in Table IV. The training samples have been selected randomly, and the rest are considered as test samples. The numbers given in brackets in Table IV represent the number of properly classified samples versus all tested samples. The confusion matrix predicts the percent accuracy for all the three sets of training data. Table IV supports this prediction. Still, the present system is tested by three types of ECG data samples, and encouraging results are obtained. In the future, the system will be tested on a large number of samples and on other types of diseases.

In recent years, the other published approaches proposed for ECG classification are mainly based on NN and fuzzy logic. The simulation comparison of classification results between the proposed method and such other methods [15], [31], [32], [35], [36] is given in Table V. We use the performance index as the detection or classification sensitivity, which is defined as the ratio of the number of truly detected test samples and the total number of test samples.

## V. CONCLUSION

The suitability of rough-set theory in ECG analysis has been tested in this paper. To do so, an automated offline-data-acquisition package is developed to extract the ECG signals from paper records. The creation of this offline-data-acquisition package is essential because our final goal is to develop a digital ECG database for subjects of different ages and food habits, for rural and urban people, and for normal and diseased subjects in the Indian context. In rural areas, most of the doctors use conventional ECG machines with a paper plotter. Therefore, the easiest way to create digital time databases is to use an appropriate offline-data-acquisition package.

Six different types of noise may corrupt these extracted ECG signals. We use Cool Edit Pro software for simulating different noises and, then, generating the appropriate filters to remove them. A knowledge base about the time-domain features and ECG interpretation is developed from various medical books and from the feedbacks of some reputed cardiologists. These features are extracted from each of the 12-lead ECG signals with the help of syntactic approaches. A rule-based rough-set decision system is developed from these time-domain features to make an inference engine for disease identification. Currently, the system is tested with three types of ECG data, namely normal, Myocardial Ischemia, and Myocardial Infarction. An accuracy of 100% is obtained for both the trained and untrained

TABLE V
SIMULATION COMPARISON BETWEEN THE PROPOSED METHOD AND THE OTHER METHODS

| Algorithm | ECG Data Type | Classification / Detection Sensitivity | |
|---|---|---|---|
| Proposed Algorithm | Normal<br>Ischemia | 100.0%<br>95.8% | |
| | MI | 100.0% | |
| U.R. Acharya et.al [32] | Normal | 85.0% [ by NN] | 95.0% [ by fuzzy] |
| | Ischemia / Dilated cardiomyopathy | 85.0%[by NN] | 90.0% [by fuzzy] |
| N. Maglaveras et al [15] | Ischemia Episode | 85.0% | |
| F. Jager et al [35] | Ischemia Episode | 83.8% | |
| A. Taddei et al [36] | Ischemia Episode | 84.0% | |

dataset for normal and MI, whereas for Ischemia, 100% and 95.8% accuracy is obtained for the trained dataset and for the untrained sample, respectively.

The simulation comparison between the proposed method and a few other methods, where mainly NN and fuzzy sets are used, is also reported here. This paper is unique basically for two reasons. First, the rough set is used to optimize rules for cardiac-disease identification, by which the complexity of NN can be avoided. Second, the time-domain features used by the doctors are selected for input parameters of the classifier to incorporate more humanlike decision-making, whereas in other works, only a few of these features [26], or different characteristic points of the ECG [33], are used as deterministic parameters.

Heart disease is one of the most common causes of death all over the world. Different statistical surveys indicate that heart disease, especially ischemic heart disease (IHD), has become a major health burden also in India, where these surveys show a steady increase of IHD throughout this country.

Therefore, advanced research is needed as a preventive measure against this silent killer. At the primary stage of developing this inference engine, we selected three major classes of ECG data (normal, Ischemia and MI) for this paper. In the future, the system will be tested with a large number, and also other types, of diseases that are fairly common in India.

## REFERENCES

[1] C. Abreu-Lima and J. P. de Sa, "Automatic classifiers for the interpretation of electrocardiograms," *Rev. Port. Cardiol.*, vol. 17, no. 5, pp. 415–428, May 1998.

[2] F. M. Ham and S. Han, "Classification of cardiac arrhythmia using fuzzy ARTMAP," *IEEE Trans. Biomed. Eng.*, vol. 43, no. 4, pp. 425–430, Apr. 1996.

[3] Y. H. Hu, W. J. Tompkins, J. L. Urrusti, and V. X. Afonso, "Applications of artificial neural networks for ECG signal detection and classification," *J. Electrocardiol.*, vol. 26, pp. 66–73, 1993.

[4] A. S. Al-Fahoum and I. Howitt, "Combined wavelet transformation and radial basis neural networks for classifying life-threatening cardiac arrhythmias," *Med. Biol. Eng. Comput.*, vol. 37, no. 5, pp. 566–573, Sep. 1999.

[5] R. Degani, "Computerized electrocardiogram diagnosis: Fuzzy approach," *Methods Inf. Med.*, vol. 31, no. 4, pp. 225–233, Nov. 1992.

[6] G. Bortolan, C. Brohet, and S. Fusaro, "Possibilities of using neural networks for ECG classification," *J. Electrocardiol.*, vol. 29, pp. 10–16, 1996.

[7] C. D. Nugent, J. A. Webb, and N. D. Black, "Feature and classifier fusion for 12-lead ECG classification," *Med. Inform. Internet Med.*, vol. 25, no. 3, pp. 225–235, Jul.–Sep. 2000.

[8] P. Bozzola, G. Bortolan, C. Combi, F. Pinciroli, and C. BroHet, "A hybrid neuro-fuzzy system for ECG classification of myocardial infarction," in *Proc. Comput. Cardiol.*, Indianapolis, IN, 1996, pp. 241–244.

[9] M. J. King, J. S. Han, K. H. Park *et al.*, "Classification of arrhythmia based on discrete wavelet transform and roughset theory," in *Proc. ICCAS*, 2001.

[10] Z. Pawlak, "Rough sets," *Int. J. Inf. Comput. Sci.*, vol. 11, no. 5, pp. 341–356, 1982.

[11] ——, *Rough Sets: Theoretical Aspects of Reasoning About Data*, ser. Series D: System Theory, Knowledge Engineering and Problem Solving, vol. 9. Dordrecht, The Netherlands: Kluwer, 1991.

[12] ——, "Bayes' theorem revised—The rough set view, new frontiers in artificial intelligence," in *Proc. JSAI Workshop*, 2001, vol. 2253, pp. 240–250.

[13] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd ed. Reading, MA: Addison Wesley Longman, 2000, pp. 491–495.

[14] F. B. Hildebrand, *Introduction To Numerical Analysis*. New Delhi, India: Tata McGraw-Hill, pp. 82–84.

[15] N. Maglaveras, T. Stamkopoulos, C. Pappas, and M. Strintzis, "An adaptive back-propagation neural network for real-time Ischemia episodes detection. Development and performance analysis using the European ST-T database," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 7, pp. 805–813, Jul. 1998.

[16] R. Silipo, P. Laguna, C. Marchesi, and R. G. Mark, "ST-T segment change recognition using artificialneural networks and principal component analysis," in *Proc. Comput. Cardiol.*, 1995, pp. 213–216.

[17] F. Jager, R. G. Mark, G. B. Moody, and S. Divjak, "Analysis of transient ST segment changes during ambulatory monitoring using the Karhunen-Loeve transform," in *Proc. Comput. Cardiol.*, 1992, pp. 691–694.

[18] Z. Pawlak, "The rough set view on Bayes' theorem, advances in soft computing," in *Proc. Int. Conf. Fuzzy Systems AFSS*, 2002, vol. 2253, pp. 106–116.

[19] L. Polkowski and A. Skowron, *Rough Sets in Knowledge Discovery*. Wurzburg, Wein: Physica-Verlag, 1998.

[20] A. Øhrn, "Discernibility and rough sets in medicine: Tools and applications," Ph.D. dissertation, Dept. Comput. Inf. Sci., Norwegian Univ. Sci. Technol., Trondheim, Norway, 1999. NTNU Rep. 1999:133, IDI Report.

[21] M. J. Goldman, *Principles of Electrocardiography*, 11th ed. Singapore: Marugen Asia (Pvt.) Ltd.

[22] J. R. Hampton, *The ECG Made Easy*, 5th ed. Edinburgh, U.K.: Churchill Livingstone.

[23] A. L. Goldberger, *Clinical Electrocardiography, A Simlified Approach*, 6th ed. India: Harcourt India Pvt. Ltd

[24] M. Lagerholm and C. Peterson, "Clustering ECG complexes using hermite functions and self organizing maps," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 7, pp. 838–848, Jul. 2000.

[25] A. Bosnjak, G. Bevilacqua, G. Passariello, F. Mora, B. Sanso, and G. Carrault, "An approach to intelligent ischemia monitoring," *Med. Biomed. Eng. Comput.*, vol. 33, no. 6, pp. 749–756, 1995.

[26] H. Al-Nashash, "Cardiac arrhythmia classification using neural networks, technology and healthcare," *Official J. Eur. Soc. for Eng. Med.*, vol. 8, no. 6, pp. 363–372, 2000.

[27] Y. Xiao, H. Chen, and J. Ge, "Automated analysis technology of electrocardiogram," *J. Biomed. Eng.*, vol. 17, no. 3, pp. 339–342, 2000.

[28] X. Yu and X. Xu, "QRS detection based on neural-network," *J. Biomed. Eng.*, vol. 17, no. 1, pp. 59–62, Mar. 2000.

[29] S. Kadamb, R. Murray, and G. F. Boudreaux-Bartels, "Wavelet transform-based QRS complex detector," *IEEE Trans. Biomed. Eng.*, vol. 46, no. 7, pp. 838–848, Jul. 1999.

[30] P. E. Trahanias, “An approach to QRS complex detection using mathematical morphology,” *IEEE Trans. Biomed. Eng.*, vol. 40, no. 2, pp. 201–205, Feb. 1993.

[31] N. Maglaveras, T. Stamkopoulos, K. Diamantaras, C. Pappas, and M. Strintzis, “ECG pattern recognition and classification using non-linear transformations and neural networks: A review,” *Int. J. Med. Inform.*, vol. 82, pp. 191–208, 1998.

[32] U. R. Acharya, P. S. Bhat, S. S. Iyengar, A. Rao, and S. Dua, “Classification of heart rate data using artificial neural network and fuzzy equivalence relation,” *Pattern Recognit.*, vol. 36, no. 1, pp. 61–68, Jan. 2003.

[33] X. M. Huang and Y. H. Zhang, “A new application of rough set to ECG recognition,” in *Proc. Int. Conf. Mach. Learn. and Cybern.*, Nov. 2003, vol. 3, pp. 1729–1734.

[34] G. M. Friesen, T. C. Jennett, M. A. Jadallah, S. L. Yates, S. R. Quint, and H. T. Nagle, “A comparison of the noise sensitivity of nine QRS detection algorithms,” *IEEE Trans. Biomed. Eng.*, vol. 37, no. 1, pp. 85–98, Jan. 1990.

[35] F. Jager, G. B. Moody, A. Taddei, and R. G. Mark, “Performance measure for algorithms to detect transient ischemic ST segment changes,” in *Proc. Comput. Cardiol.*, 1991, pp. 369–372.

[36] A. Taddei, G. Distante, M. Emdin, P. Pisani, G. B. Moody, C. Zeelenberg, and C. Marchesi, “The European ST-T database: Standard for evaluating systems for the analysis of St-T changes in ambulatory electrocardiography,” *Eur. Heart. J.*, vol. 13, no. 9, pp. 1164–1172, 1992.

[37] S. Tsumoto, “Automated induction of medical expert system rules from clinical databases based on rough set theory,” *Inf. Sci.*, vol. 112, no. 1, pp. 67–84, Dec. 1998.

[38] ——, “Extraction of experts’ decision rules from clinical databases based on rough set theory,” *J. Intell. Data Anal.*, vol. 2, no. 3, pp. 548–552, 1998.

[39] D. Tsaptsinos and M. G. Bell, “Medical knowledge mining using rough set theory,” in *Proc. Int. Conf. Neural Netw. and Expert Syst. Med. and Healthcare*, Plymouth, U.K., Aug. 1994.

**Sucharita Mitra** (S’02) received the B.Sc. degree in physics (with honors), the M.Sc. degree in electronic science, and the Ph.D. degree in technology of measurement and instrumentation from the University of Calcutta, Calcutta, India, in 1995, 1997, and 2005, respectively.

She is a Senior Research Fellow of Council of Scientific and Industrial Research, Government of India, and is currently engaged as a Guest Lecturer with the Department of Applied Physics, University of Calcutta. Earlier, she was associated with Computer Vision and Pattern Recognition Unit of Indian Statistical Institute, Calcutta, as a Project Personnel. Her research interest includes biomedical signal and image processing, pattern recognition, document processing, etc. She has published about 20 research papers in reputed journals and international conferences.

**Madhuchhanda Mitra** received the B.Sc. degree in physics (with honors), and the B.Tech., M.Tech., and Ph.D. degree in measurement and instrumentation from the University of Calcutta, Calcutta, India.

She has 12 years of teaching in the University College of Technology, University of Calcutta. Her special fields of interest include biomedical-signal processing, process instruments, microprocessors, and microcontrollers. She is currently a Reader with the Instrumentation Engineering Division, Department of Applied Physics. She has about 50 technical papers published in national and international journals and conference proceedings.

Dr. Mitra is the recipient of Best Paper Award for three of her papers. She is a recipient of “Griffith Memorial Award” of the University of Calcutta.

**B. B. Chaudhuri** (S’79–M’80–SM’87–F’01) received the B.Sc. (with honors), B.Tech., and M.Tech. degrees from Calcutta University, Calcutta, India, in 1969, 1972, and 1974, respectively, and the Ph.D. degree from Indian Institute of Technology, Kanpur, in 1980.

He worked as Postdoc Fellow at Queen’s University, London, U.K., during 1981–1982. He joined the Indian Statistical Institute in 1978, where he served as the Project Coordinator and Head of the National Nodal Center for Knowledge-Based Computing funded by Department of Electronics (DoE) and United Nations Development Programme (UNDP). He was a Visiting Scientist at the Germany Science Foundation, Munich, and as a Guest Faculty at the Technical University of Hannover, Hannover, Germany during 1986–1988. Currently, he is the Head of Computer Vision and Pattern Recognition Unit of the Indian Statistical Institute. His research interests include pattern recognition, image processing, computer vision, natural language processing (NLP), information retrieval, speech processing, digital-document processing, and Optical Character Recognition (OCR). He pioneered the first Indian-language Bharati-braille system for the blind, a successful Bangla speech-synthesis system, as well as the first workable OCR for Bangla, Devnagari, Assamese, and Oriya scripts. In the NLP area, a robust Indian-language spell-checker, morphological processor, multiword-expression detector, and statistical analyzer were pioneered by him. Some of his technologies have been transferred to industry for commercialization. He has published about 300 research papers in reputed international journals, conference proceedings, and edited books. Also, he has authored three books entitled *Two-Tone Image Processing and Recognition* (Wiley Eastern, 1993), *Object-Oriented Programming: Fundamentals and Applications* (Prentice-Hall, 1998) and *Computer and Software Technology Dictionary* (Ananda, 2002).

Dr. Chaudhuri received a Leverhulme fellowship in 1981–1982, the Sir J. C. Bose Memorial Award in 1986, the M. N. Saha Memorial Award (twice) in 1989 and 1991, the Homi Bhabha Fellowship award for OCR of the Indian languages and computer communication for the blind in 1992, the Dr. Vikram Sarabhai Research Award for outstanding achievements in the fields of electronics, informatics, and telematics in 1995, the C. Achuta Menon Prize in 1996, the Homi Bhabha Award: Applied Sciences in 2003, and the Ram Lal Wadhwa Gold Medal in 2005. He is also a recipient of the prestigious Jawaharlal Nehru Fellowship to conduct research on document processing during 2004–2006. He is a member secretary of the Indian section of International Academy of Sciences, Fellow of National Academy of Sciences in India, Fellow of the Institution of Electronics and Telecommunication Engineering, and Fellow of the Indian National Academy of Engineering. He was elected as a Fellow of the International Association of Pattern Recognition in 1998. Currently, he is serving as an Associate Editor of *Pattern Recognition, VIVEK, International Journal of Pattern Recognition, and Artificial Intelligence, International Journal of Computer Vision,* and *International Journal of Document Analysis and Recognition.* He served as a Guest Editor of special issues of several journals.