Journal of Advances in Computer Research
Quarterly ISSN: 2008-6148
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 3, No. 3, August 2012), Pages: 35-52
www.jacr.iausari.ac.ir

A New Approach to Detect Congestive Heart Failure Using Symbolic Dynamics Analysis of Electrocardiogram Signal

Chandrakar Kamath

Ex-Professor, Electronics and Communication Dept.,Manipal Institute of Technology, Manipal-576104
chandrakar.kamath@gmail.com

Received: 2012/05/14; Accepted: 2012/06/28

Abstract

The aim of this study is to show that the measures derived from Electrocardiogram (ECG) signals many a time perform better than the same measures obtained from heart rate (HR) signals. A comparison was made to investigate how far the nonlinear symbolic dynamics approach helps to characterize the nonlinear properties of ECG signals and HR signals, and thereby discriminate between normal and congestive heart failure (CHF) subjects. The symbolic dynamics calculations performed on normal and CHF ECG and HR signals showed significant differences in the symbol-sequence histogram statistics and complexity measures (modified Shannon entropy (MSE) and multi-valued Lempel-Ziv complexity (MLZC)) of symbol sequences between the two groups. The ability of these complexity measures to discriminate normal from CHF subjects was evaluated using receiver operating characteristic (ROC) plots. It is found that MSE and MLZC measures obtained from ECG signals performed better than the same measures derived from HR signals of the same subjects.

Keywords: *Congestive heart failure, Complexity measures, Electrocardiogram signal, Heart rate signal, Modified Shannon entropy, Multi-valued Lempel-Ziv complexity, Symbolic dynamics*

1. Introduction

Despite numerous recent advances in the understanding of the pathophysiology of congestive heart failure (CHF) and improvements in its therapy, the mortality rate has remained high [1]. As a consequence the development of new methods and measures of mortality risk in CHF, including sudden cardiac death, is still a major challenge in contemporary cardiology. Besides this, there is a need to reach remote and underserved communities with life saving healthcare. A reliable automated diagnostic system combined with high-speed communication can resolve this issue. This work is an attempt to develop such an automated system to discriminate between normal and congestive heart failure subjects.

The rest of this paper is organized as follows. In Section 2, the related works are discussed. In Section 3, motivation of doing this work and in Section 4, the data used and the proposed framework are explained. In Section 5, the measures of complexity are presented. In Section 6, the results of the application of the new method are discussed. Finally, some concluding remarks are mentioned in Section 7.

2. Related Work

Cardiovascular system is characterized by a high complexity, partly because of its continuous interactions with other physiological systems [2]-[3]. Further, it has been found that this complexity breakdowns with cardiac diseases and also, aging [4]. Cardiac diseases often manifest themselves in characteristic changes in the ECG as well as HR signals. As an implication the complex dynamics hidden in the generation of ECG/heart beats cannot be quantified or characterized using traditional methods of data analysis in time and frequency domains [5]. The classical nonlinear methods suffer from the disadvantage of dimensionality [6]-[7]. Further, there are not enough samples in the time series to arrive at a reasonable estimate of the nonlinear measures [8]. From this point of view it is advisable to resort to methods which can quantify system dynamics even for short time series, like the symbolic dynamics.

For the past few decades, the application of symbolic analysis has found several diverse fields like, astrophysics, geomagnetism, geophysics, classical mechanics, chemistry, medicine and biology, mechanical systems, fluid flow, plasma physics, robotics, communication, and linguistics [9]. To be specific, in medicine, various implementations of symbolic sequences have been used to characterize electroencephalography (EEG) signals to understand the interaction between brain structures during seizures [10]. Under mechanical systems, symbolic methods were applied to combustion data from internal combustion engines to study the onset of combustion instabilities [11] and in multiphase flow data-symbolization were found to be useful in characterizing and monitoring fluidized-bed measurement signals [12]. Symbolic dynamics, as an approach to investigate complex systems, has found profound use in the analysis of HR signals [13]-[17]. Kurths *et al.* [13] concluded that the traditional methods of data analysis in time and frequency domains were insufficient to characterize the HRV. They applied methods of nonlinear dynamics based on symbolic dynamics to analyze the HRV time series. They found that the renormalized entropy together with the parameters in the frequency domain were promising in quantifying individual risk. Porta *et al*. [14] applied symbolic dynamics to beat-to-beat HRV series from 24h Holter recordings. Using a uniform quantization procedure the HRV series was transformed into a sequence of six symbols and the symbols were grouped in patterns to characterize physiological conditions. The indexes derived from symbolic dynamics were found to be capable of discriminating pathological from healthy populations. Changes in autonomic modulation during the progression of CHF in rat model were assessed by Tobaldini *et al*. using spectral and symbolic analyses. Their study revealed that the symbolic analysis was found to be more suitable than spectral analysis to describe the alterations of heart rate dynamics [15]. The efficacy of the measures of complexity based on symbolic dynamics has been confirmed in the assessment of risk of patients after myocardial infarction and the architecture of human cancellous bones [16]. Voss *et al*. employ symbolic dynamics to investigate the complexity of the dynamical aspects of the HRV series [17]. By comparison with other nonlinear methods they conclude that symbolic dynamics has a close connection to physiological aspects and that it relatively easy to interpret. They found that symbolic analysis can separate structures of nonlinear dynamics in the HRV series more successfully than the conventional methods in time and frequency domains. In all the above studies and many more, the thrust has been on HRV time series. Although Shannon entropy and Lempel-Ziv complexity measure have been widely used in the

literature, not much research is found in the context of CHF. Based on symbolic representations and their probability distributions, Parlitz *et al*. used different biomarkers, including Shannon entropy, to distinguish CHF patients from control group [18]. With Shannon entropy as the biomarker the correct classification was at 80%. Voss *et al*. proposed a novel method using compression entropy, a complexity measure based on Lempel-Ziv algorithm, for the analysis of heart rate dynamics in CHF patients [19]. They found compression entropy to be useful to detect differences in heart rate dynamics before the onset of ventricular tachyarrhythmia.

3. Motivation

The prime advantages of symbolic dynamics are the following: If the fluctuations of the two data series are governed by different dynamics then the evolution of the symbolic sequences is not related. The resulting symbolic sequences histograms give a reconstruction of their respective histories and provide a visual representation of the dynamic patterns. In addition, they may be used as a basis to build statistics to compare the regions that show different dynamical properties and indicate which patterns are predominant. Moreover, symbolization has been successfully applied to a number of noisy nonlinear processes [20]. Thus methods of symbolic dynamics are useful approaches for classifying the underlying dynamics of a time series. Parameters of time domain and frequency domain often leave these dynamics out of consideration. Fruitful applications of symbolic methods are preferred in situations where robustness to noise, speed, and/or cost is important [9]. The process of symbolization can be used to represent any possible variation over time, depending on the number of symbols and the sequence lengths used. This is a very powerful property because it does not make any assumptions about the nature of the signals/patterns (e.g., it works equally well for both linear and nonlinear phenomena).

However, there is hardly any literature where symbolic dynamics is applied for the analysis of raw ECG signals. The disadvantages of most of the methods used for the analysis of HR signal are (1) misrecognitions of RR intervals of lengths zero, RR intervals less than 200 ms (human refractory time) and pauses, i.e. the interval when heart does not pump; (2) removal of artefacts (e.g. double recognition, i.e. R-peak and T-wave recognized as two beats); and (3) the required corrections for ectopic beats. These difficulties make the analysis complicated and time-consuming. Further, some of the pathologies such as the left bundle branch block and the right bundle branch block cannot be detected using only the heart rate variability features [21]. On the other hand, ECG signal is more susceptible to noise than HR signal. However, symbolic dynamics takes care of this noise as mentioned above. In this contribution symbolic dynamics is employed to classify (or: distinguish between) both the ECG and HR signals obtained from standard Holter recordings from MIT-BIH database into normal and CHF subjects using modified Shannon entropy (MSE) and multi-valued Lempel-Ziv complexity (MLZC) as complexity measures. Receiver operating characteristic (ROC) plots were used to evaluate the ability of these complexity measures to discriminate normal from CHF subjects.

4. The Proposed Framework and materials

First the ECG and the corresponding HR data used are discussed followed by the symbolic dynamics of the time series.

a. Analyzed data

In this work two data sets of signals from the benchmark PhysioNet database [22] are used. The first data set includes 18 ECG records from MIT-BIH normal sinus rhythm (NSR) database (nsrdb) and ECG records of 15 subjects with severe CHF (NYHA class 3-4) from BIDMC CHF database (chfdb). The NSR database includes long term ECG recordings of 5 men, aged 26 to 45 years, and 13 women, aged 20 to 50 years. The CHF database includes long term ECGs (about 20 hours each) of 11 men, aged 22 to 71 years, and 4 women, aged 54 to 63 years. From each record the modified limb lead II was only considered for analysis. The resolution is 200 samples per mV. The sampling frequency of normal sinus rhythm signal is 128 Hz and that of CHF signal is 250 Hz. Since the sampling frequency does influence upon the calculated parameters it is necessary to have the same sampling frequency for all the records. For this reason ECG signals from normal database are first re-sampled at 250 Hz. Then each record is divided into segments of equal time duration (20 sec), with 5000 samples/segment in both normal sinus rhythm and CHF database. A total of 3510 segments from NSR and a total of 2925 segments from CHF data bases are analyzed. It is to be noted that the above re-sampling will have no effect on the timing of R-peaks and the derived RR interval signals.

The second data set includes HR/RR interval signals of the same NSR and CHF subjects as in data set one. All the normal and CHF HR records are passed through a square filter to eliminate artifacts, premature beats and outliers, if any. Each record was then divided into segments, with 5500 samples/segment, in both the groups.

b. Symbolic dynamics analysis

First the two common types of symbolic transformations are dealt with and then their advantages and implications. Then the specific transformation as applied to ECG and HR time series are discussed. Next, the construction of symbol sequences, plotting symbol-sequence histograms, applying symbol-sequence statistics and finally employ measures of complexity to decide on the nature of time series are presented.

c. Static and Dynamic transformations

Symbolic dynamics/time series analysis or symbolization, as an approach to investigate complex dynamical processes, facilitates the analysis of dynamic aspects of the signal of interest. The concept of symbolic dynamic analysis is based on coarse-graining of the dynamics of the time series [17]. That is the range of original observations or the range of some transform of the original observations such as the first difference between the consecutive values, is partitioned into a finite number of discrete regions, *n*, and each region is associated with a specific symbolic value so that each observation or the difference between successive values is uniquely mapped to a particular symbol depending on the region into which it falls. The former mapping is called static transformation and the latter dynamic transformation. Static transformation with more number of partitions is preferred where one is concerned about observing details which are small compared to the overall range. On the other hand dynamic

transformation is preferred when the observed process appears nonstationary or has long time scale variations. Thus the original observations are transformed into a series of same length but the elements are only a few different symbols (letters from the same alphabet), the transformation being termed symbolization. A good criterion to symbolize the data is to define the partitions such that (1) the individual occurrence of each symbol is equiprobable with all other symbols or (2) the measurement range covered by each region is equal. This is done to bring out ready differences between stochastic and deterministic structure in the data. The transformations into symbols have to be chosen context dependent [23]. This way the study of dynamics simplifies to the description of symbol sequences. Some detailed information is lost in the process but the coarse and robust properties of the dynamic behavior is preserved and can be analyzed [23].

d. Symbolic Dynamics and ECG and HR time series

In this study, static transformation approach for the symbolic dynamics [12] is employed. In the literature a symbolic dynamic representation using two symbols with one quantization level or four symbols with three non-uniform quantization levels, as applied to HR time series is common [24]. Sometimes the thresholds used in these quantization approaches are related to mean or median of the time series. But in the non-stationary signals the mean or the median change abruptly [25]. This problem can be remedied by using static transformation with uniform quantization levels. This also fulfills the requirement of a MLZC which demands uniform quantization. In this study six symbols (n=6) with five uniform quantization levels as shown in the eqn. (1) below are used.

$$S_i = \begin{cases} 0, & if \quad x_{\min} \le x_i < x_{\min} + d \\ 1, & if \quad x_{\min} + d \le x_i < x_{\min} + 2d \\ 2, & if \quad x_{\min} + 2d \le x_i < x_{\min} + 3d \\ 3, & if \quad x_{\max} - 3d \le x_i < x_{\max} - 2d \\ 4, & if \quad x_{\max} - 2d \le x_i < x_{\max} - d \\ 5, & if \quad x_{\max} - d \le x_i \le x_{\max} \end{cases} \tag{1}$$

where, x_{min} and x_{max} are respectively, minimum and maximum values of the time series x_i. The distance, d, between partitions is given by $d = (x_{max} - x_{min})/n$.

After symbolization the next step in the identification of characteristic temporal patterns is the construction of symbol sequences of specific length L, termed words, from the symbol series by gathering groups of symbols in the temporal order. L is called the word length. This sequencing process involves definition of a template of finite length L that can be moved along the symbol series one symbol at a time, each step revealing a new sequence/word. If each possible new sequence is identified by a unique identifier the resulting series will be a new time series, termed word-sequence series. The next step is to evaluate the relative frequency of occurrence of all possible words. A simple way to keep track word-sequence frequencies is to assign a unique value, called symbolic code, to each word by computing the corresponding base-10 value for each base-n word, where, n is the number of partitions. For example, with number of partitions n=2, and word length L=3, a sequence '*101*' will have a sequence code of 5. The next step is to plot symbol-sequence frequencies as a function of symbolic code, the

plot being termed symbol-sequence histogram. Because of the above rule of thumb for partitioning, for a truly random data the relative frequency of all possible symbolic codes will be equal. This implies that any significant deviation from this equiprobable feature is an indication of deterministic characteristic of the given data, the more the deviation the more is the data deterministic and time correlated.

e. Determining Optimum Symbol-sequence Length

One approach that is useful for selecting an appropriate sequence/word length involves employing MSE explained in Section 5 below. It is empirically found that this value decreases, reaches a minimum and then increases, as sequence length is increased from 1. This sequence length corresponding to minimum reflects the symbol sequence transformation that best distinguishes the data from a random sequence [11]. Sequences that are too short lose some deterministic information while those that are too long reflect noise and deplete data for reliable statistics. Thus the sequence length L, for which MSE is minimum corresponds to almost an optimal length. In this study, empirically it is observed that word lengths of three (L=3) is a suitable choice for both the normal and CHF groups as explained in Section 6 below. There are several quantities (statistics and complexity measures) that properly characterize such symbol strings. In this work Euclidean norm (T statistic) and a modified χ^2 statistic are used to compare the histograms. In particular, the frequency distribution (relative frequencies) of six symbol and length 3 words, i.e. substrings which consist of three adjacent symbols from the alphabet {0, 1, 2, 3, 4, 5} leading to a maximum of 126 (6^3) different words/bins are investigated. The symbol-sequence histogram for each case is plotted and then the pattern classification is performed. This is a compromise between retaining important dynamical information and of having a robust statistics to estimate probability distribution.

f. Symbol-sequence Statistics

In addition to providing a visual representation of the dynamic patterns, symbol-sequence histograms provide the basis for quantitative statistics. As mentioned above Euclidean norm (T statistic) and a modified χ^2 statistic are employed to compare the histograms. The Euclidean norm is defined as [12]

$$T_{AB} = \sqrt{\sum_i (A_i - B_i)^2} \tag{2}$$

and the modified χ^2 statistic is defined as [13]

$$\chi^2_{AB} = \sum_i \frac{(A_i - B_i)^2}{(A_i + B_i)} \tag{3}$$

where A_i and B_i are the individual sequence probabilities for sequence i for histograms A and B. It is seen that both the statistics are obtained by differencing the frequencies of the individual sequences for the different histograms. When the frequency differences are large, the resulting statistics will also be large. Thus, large values for the statistics imply that the dynamic patterns in the data set are completely different. The Euclidean norm is based on the idea that each symbol histogram can be considered as a vector in multi-dimensional space, where the number of dimensions corresponds to the possible unique sequences. Consequently, the magnitude of the vector difference between the histograms must provide a good comparison of the histograms. A larger distance between histograms implies that the dynamics in the data set are very different. The

modified χ^2 statistic has been derived from the standard χ^2 statistic with the univariate frequencies replaced by sequence frequencies.

g. Time-irreversibility

Time reversibility of a time series refers to the invariance of the statistical properties under time reversed conditions. It is important to note that data nonstationarity will result in time irreversibility [26]-[28]. It is shown that the time irreversibility, although not an absolute test for nonstationarity, the degree of time reversibility can serve as a good indicator of nonstationarity [27]. The level of time irreversibility is used as an indicator to classify the ECG records. Symbol-sequence histograms are useful for quantifying the time irreversibility because the relative frequencies will shift when the data are observed backwards in time. Both *T* statistic and modified χ^2 statistic can be used to characterize time irreversibility in a given time series by observing the difference in symbol sequence histograms for the forward-time and reverse-time realizations using the same Equation (2) and (3) with *A* and *B* representing histogram frequencies of forward and reverse time analyses, respectively. The statistics (*T* statistic and a modified χ^2 statistic) quantify the level of time-irreversibility.

5. Measures of Complexity

The first measure of complexity is the MSE given below [11]. A larger value implies higher complexity and a smaller value implies a lower complexity. The MSE (MSE) defined as

$$H_s = \frac{1}{log\, N_{obs}} \sum_i p_i \, log\, p_i \qquad (4)$$

where p_i is the normalized probability of the i^{th} symbol sequence, and N_{obs} is the number of possible sequences which are actually observed in the data. Note that the normalization is with respect to Shannon entropy for a completely random process (one in which all sequences are equiprobable). The advantage of this normalization is to bring down the bias on the statistics due to finite size of the data sets. This implies that the MSE will converge to 1 as the data approaches true randomness and for non-random data this value will be $0 \leq H_s \leq 1.0$ and a lower H_s implies more deterministic structure.

The second measure of complexity is the MLZC. The LZC algorithm was proposed by Lempel and Ziv to evaluate the randomness of finite sequences. It is rather a simple-to-compute nonparametric measure of complexity suitable for finite length one-dimensional signals related to the number of distinct substrings and the rate of their recurrence. Larger values of LZC imply higher complexity data. Since LZC analyzes finite symbol-sequences it is essential that the given signal must first be coarse-grained. As the symbol sequence using binary coarse-graining method is likely to lose some important information of the dynamical system, in this study, a multi-valued coarse-graining method (with six symbols, n=6 as given in Equation (1) above) is used. This symbolic string is scanned from left to right and a complexity counter $c(N)$ is incremented by one unit every time a new subsequence pattern is encountered in the scanning process, and the immediate next symbol is regarded as the beginning of the next subsequence pattern. The LZC can be estimated using the following algorithm [26].

1. Let *P* denote the original string sequence i.e. $P = \{s_1, s_2, s_{3,}, \ldots\}$, with s_i defined as in Equation (1). Let *S* and *Q* denote two subsequences of *P* and *SQ* be concatenation of *S*

and *Q*. Also, let *SQπ* be a sequence derived from *SQ* after its last character is deleted (*π* implying deletion of last character in the sequence) and *υ(SQπ)* denote the vocabulary of all different subsequences of *SQπ*.

2. At the beginning, the complexity counter *c(N)=1*, $S=s_1$, $Q=s_2$, $SQ=s_1,s_2$, and therefore, $SQ\pi=s_1$.

3. In general, with $S=s_1, s_2, s_3,\ldots,s_r$ and $Q=s_{r+1}$, $SQ\pi= s_1, s_2, s_3,\ldots,s_r$. If *Q* belongs to *υ(SQπ)* then *Q* is subsequence of *SQπ* and not a new sequence.

4. With *S* intact, change *Q* to s_{r+1}, s_{r+2} and check if *Q* belongs to *υ(SQπ)* or not.

5. Keep repeating previous steps until *Q* does not belong to *υ(SQπ)*. Now $Q=s_{r+1}, s_{r+2},\ldots,s_{r+i}$ is not a subsequence of $SQ\pi=s_1, s_2,..,s_{r+i-1}$. So increase *c(N)* by 1.

6. Thereafter, *S* is renewed to $S=s_1, s_2,\ldots, s_{r+i}$ and *Q* to $Q=s_{s+i+1}$.

7. Repeat the previous steps until *Q* is the last character. At this point in time, the number of subsequences in *P* is *c(N)*, which corresponds to measure of complexity.

To arrive at a measure of complexity independent of sequence length, *c(N)* must be normalized. If the length of the sequence is *n* and the number of different symbols is α, it has been shown that the upper bound of *c(N)* is [26]

$$c(N) < \frac{N}{(1-\varepsilon_N)log_\alpha(N)} \tag{5}$$

where ε_N is a small quantity and $\varepsilon_N \rightarrow 0$ ($N\rightarrow\infty$). In general, $N/log_\alpha(N)$ is the upper limit of *c(N)*, i.e.,

$$\lim_{N\rightarrow\infty} c(N) = b(N) = \frac{N}{log_\alpha(N)} \tag{6}$$

For a coarse-graining method with six symbols, α=6, $b(N)=N/log_2(N)$ and *c(N)* can be normalized by *b(N)* as

$$C(N) = \frac{c(N)}{b(N)} \tag{7}$$

C(N), the normalized LZC, reflects the arising rate of new patterns along with the sequence and thus captures the temporal structure of the sequence. A larger value of LZC means that the chance of generating a new pattern is greater, so the sequence is more complex, and vice versa.

Statistical analysis and Receiver operating characteristic (ROC) plots

As mentioned above *T* statistic and χ^2 statistic are used to evaluate the statistical differences between the estimated MSE and MLZC for normal CHF subjects. If significant differences between groups are found, then the ability of the non-linear analysis method to discriminate normal from CHF subjects is evaluated using receiver operating characteristic (ROC) plots. ROC curves are obtained by plotting sensitivity values (which represent the proportion of the patients with diagnosis of CHF who test positive) along the y axis against the corresponding (1-specificity) values (which represent the proportion of the correctly identified normal subjects) for all the available cutoff points along the x axis. Accuracy is a related parameter that quantifies the total number of subjects (both normal and CHF) precisely classified. The area under ROC curve (AUC) measures this discrimination, that is, the ability of the test to correctly classify those with and without the disease. The optimum threshold is the cut-off point in which the highest accuracy (minimal false negative and false positive results) is obtained. This can be determined from the ROC curve as the closest value to the left top

point (corresponding to 100% sensitivity and 100% specificity). An AROC value of 0.5 indicates that the test results are better than those obtained by chance, where as a value of 1.0 indicates a perfectly sensitive and specific test. A rough guide to classify the precision of a diagnostic test based on AROC is as follows: If the AROC is between 0.9 and 1.0, then the results are treated to be excellent; If the AROC is between 0.8 and 0.89, then the results are treated to be good; the results are fair for values between 0.7 and 0.79; the results are poor for values between 0.6 and 0.69; If the AROC is between 0.5 and 0.59, then the outcome is treated to be bad.

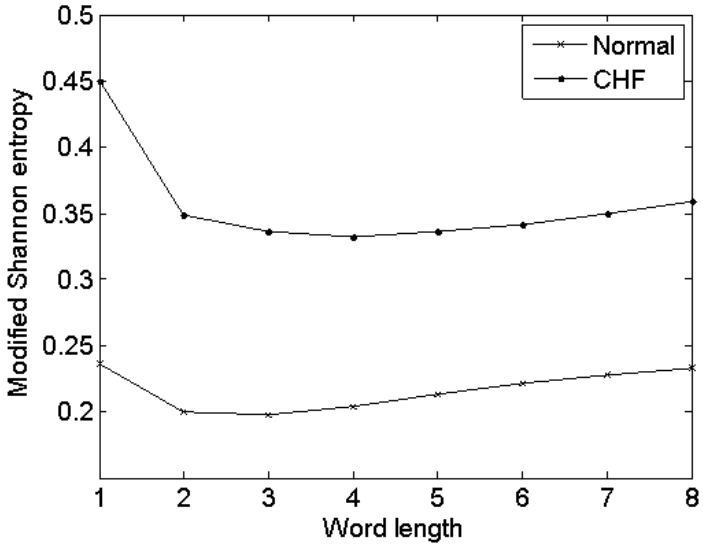

Figure 1. Modified Shannon entropy as a function of word length for Normal and CHF subjects with 6 symbols.

6. Results and discussion

The results of raw ECG time series from data set one and then the results of the corresponding HR time series from data set two are presented. The ECG records of the NSR and CHF databases are pre-processed, grouped, and segmented as mentioned in Section 4.a above. Symbolic dynamics analysis is then applied to the segments from both the groups to decide whether a particular segment belongs to normal, or CHF group. Static transformation as given in Equation (1) is applied on each segment to arrive at a symbol string with a range of six possible symbols {0, 1, 2, 3, 4, 5} (hex symbolization). The order that the regions are visited by the evolving dynamics generates a symbol sequence that characterizes physiological conditions. The resulting symbol sequence is then grouped in patterns, words, as explained in Sections 4.d and 4.e, above. The optimum length of the words is determined as explained below.

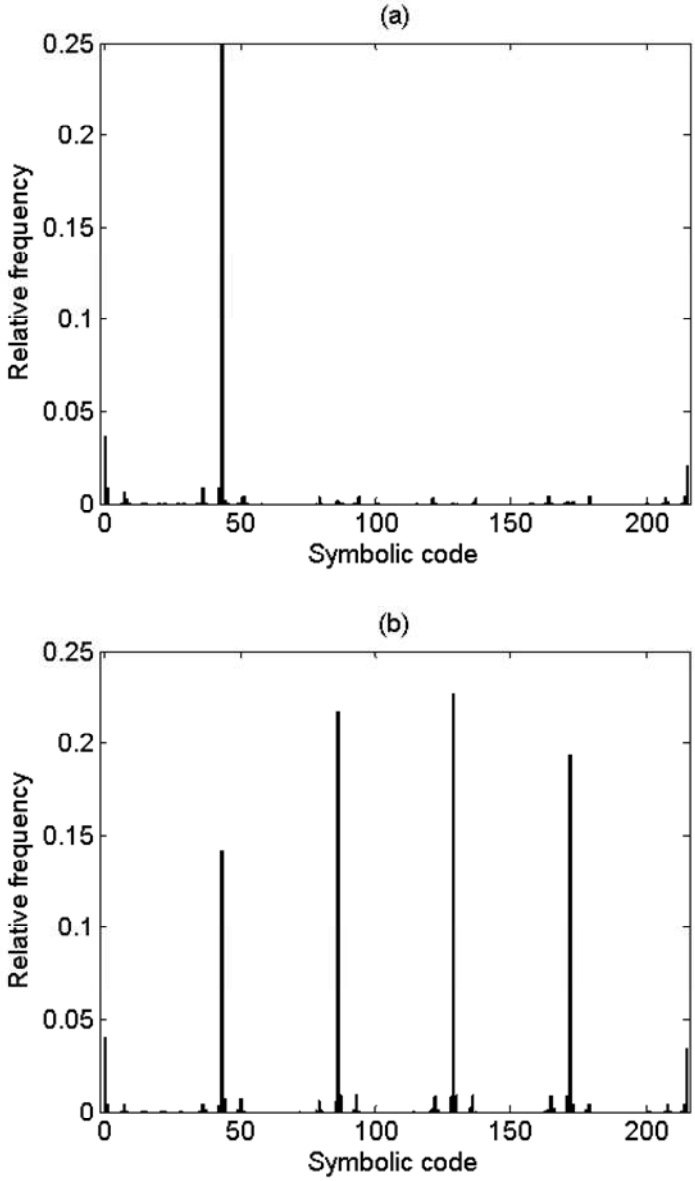

Figure2. Relative frequency distribution of symbol-sequences for (a) Normal group and (b) CHF group. Six symbols of word length 3 were used.

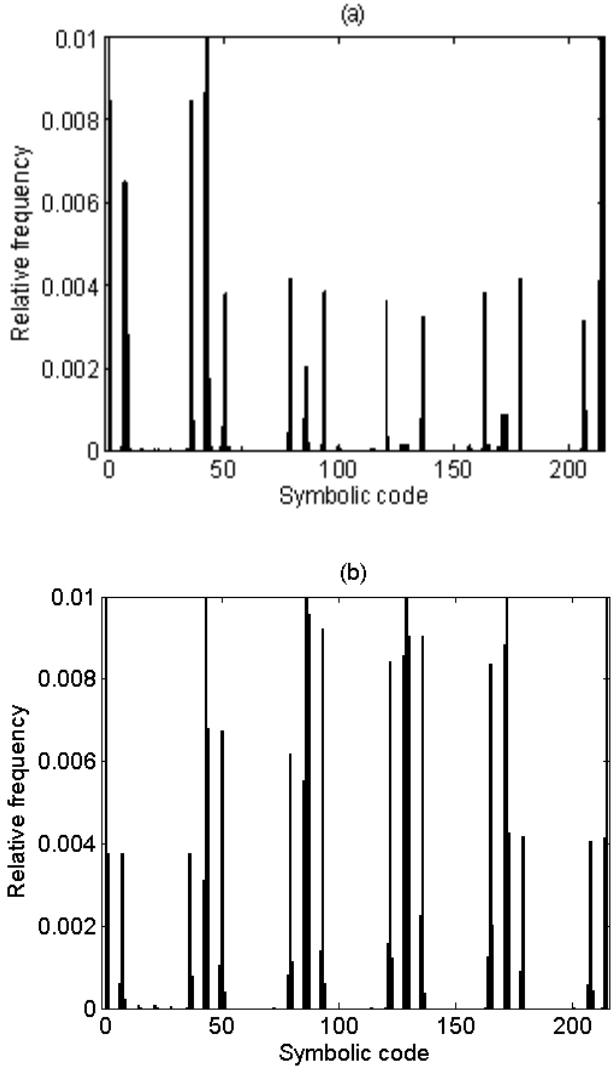

Figure3. Relative frequency distribution of symbol-sequences for (a) Normal group and (b) CHF group (Fig. 2 exaggerated to visualize lower amplitude bins). Six symbols of word length 3 were used.

a. Determining Optimum Symbol-sequence Length

As mentioned above, one approach that is useful for selecting an appropriate sequence length involves plotting MSE vs. sequence length and observing the minimum which reflects the symbol sequence transformation that best distinguishes the data from a random sequence. Thus the sequence length *L*, for which MSE is minimum, corresponds to almost an optimal length. Such plots of MSE vs. sequence/word length for hex partitions (with 6 symbols) are shown in Fig. 1 for normal and CHF groups. It is found that for normal group *L*=2 or 3 and for CHF group *L*=3 or 4. In this work a word length of three i.e., *L*=3 is chosen as a suitable value for both the normal and CHF groups.

b. Characterizing and comparing Symbol-sequence histograms of word length 3

From the same symbol strings, words of length 3 are built. A sequence code is then assigned for each of the words by using equivalent base-10 value for each of the base-n word of length 3, where, *n* is the number of partitions (in this study n=6). The average

relative frequencies of length 3 words are then computed over all the segments of the respective normal and CHF groups and symbol-sequence histograms are plotted for each of the two groups. Figs. 2(a) and 2(b) compare these word histograms for normal, and CHF subjects. The same histograms are shown exaggerated in Figs. 3(a) and 3(b) for better visualization and comparison of the lower amplitude bins. The relative frequency distribution of patterns for the two cases is found to be distinctly different. This indicates that there is a difference in the dynamics governing the two data series. Comparison among the two histograms shows that in the case of normal group, some symbolic sequences and their time-reversed versions, like, 8(012) and 78(210), 137(345) and 207(543), 1(001) and 36(100), 7(011) and 42(110), 44(112) and 79(211), 51(123) and 121(321), 94(234) and 164(432), 179(435) and 214(554), etc. exhibit some kind of dominance compared to other words. Note that the bin values are expressed as decimal (equivalent hex value). Among these sequences the first four bins {8(012) and 78(210), 137(345) and 207(543)} exhibit maximum dominance. Such sequences appear to occur because of non-stationary dynamics in the system, a characteristic of normal subjects. In the CHF group the most prominent bins include {43(111), 86(222), 129(333) and 172(444)}, which are absent in the normal subject histogram. Further, the most predominant four bins of normal subjects {8(012) and 78(210), 137(345) and 207(543)} are absent in the CHF histogram. Such presence and absence of particular patterns are typical of CHF subjects. However, all other paired bins are present, but with comparatively lower dominance. Besides these lower dominant bins there are additional lower dominant paired bins {50(122) and 85(221), 93(233) and 128(332), 122(322) and 87(223), 130(334) and 165(433), 171(443) and 136(344)}. The only predominant bins common to both normal and CHF groups are 0(000) and 215(555). Thus it is found that the symbolic sequence histograms are significantly different for each class. This implies that the dynamics governing the evolution of the ECG time series for normal and CHF subjects is completely different.

For persons with cardiac risk, the distribution of length-3 words (with 6 symbols) is concentrated on about 8 bins (out of 216 bins) where as for healthy persons it is characterized by more number of bins.

c. Symbol-sequence Statistics for the Forward-time and Reverse-time realization

Symbol sequence statistics (*T* statistic and the modified χ^2 statistic) were applied to average relative frequencies of the histograms of both normal and CHF groups and the results are tabulated in Table 1. Both the *T* statistic and the modified χ^2 statistic between the two groups (normal and CHF) are found to be large. This implies that there is a large difference in the dynamics governing the two data series of different groups.

As mentioned earlier, a heartbeat or cycle influences up to 6-10 cycles downstream [9] and this implies time irreversibility of the ECG time series. Using the same symbolic analysis, time reversal property studies are performed on both normal and CHF groups. Symbol-sequence statistics (*T* statistic and the modified χ^2 statistic) were applied to average relative frequencies of the histograms for the forward-time and reverse-time realizations of both normal and CHF groups and the results are tabulated in Table 1. Both the *T* statistic and the modified χ^2 statistic in the case of normal group are found to be larger than those in the case of CHF group. This implies that the level of time irreversibility is larger in the normal group than that observed in the CHF group. In other words, there is loss of time irreversibility in the CHF group while, prevalence of irreversibility in the normal.

d. Modified Shannon entropy

As explained in Section 3 the MSE is evaluated on relative frequency distribution of words for both the groups. A comparison statistics (mean ± SD) is shown in Table 2. For the normal group the MSE is 0.2163 ± 0.01414 and for CHF group, it is 0.4631 ± 0.02466, respectively, implying that normal subjects are more informative than CHF subjects.

Table1. T statistic and modified χ2 statistic between ECG signals of normal and CHF groups in the forward direction and their respective time-reversed versions.

Groups	*T* and χ^2 statistics
Normal and CHF	T = 0.8103 χ^2 = 1.2658
Normal and its reversed version	T = 0.0059 χ^2 = 0.0065
CHF and its reversed version	T = 0.0037 χ^2 = 0.0035

Table2. Modified Shannon entropy and Multi-valued LZC (mean ± SD) for ECG signals of normal and CHF groups.

Group	H_s (p < 0.0001)	MLZC (p < 0.0001)
Normal	0.2163 ± 0.01414	0.06575 ± 0.0049
CHF	0.4631 ± 0.02466	0.1488 ± 0.01483

e. Multi-valued Lempel-Ziv Complexity

As explained in Section 3 the MLZC is evaluated on hex symbolization of both the groups. A comparison statistics (mean ± SD) is also shown in Table 2. For the normal group the MLZC is 0.06575 ± 0.0049 and for CHF group, it is 0.1488 ± 0.01483, respectively, implying that normal subjects have a decreased complexity of temporal patterns compared to CHF subjects.

f. Receiver Operating Characteristic (ROC) plots

The ability of the MSE and MLZC to discriminate between normal and CHF subjects, in which significant differences were found, is evaluated using ROC plots. Figs. 4(a) and 4(b) show ROC curves for the two cases respectively. Table 3 summarizes the results. The value for the area under the ROC curve can be interpreted as follows: an area of 0.9943 (in the case of MLZC, for example) means that a randomly selected individual from the normal group has a MLZC value smaller than that of a randomly chosen individual from CHF group in 99.43% of the time. A rough guide to classify the accuracy of a diagnostic test is related to the area under ROC curve. With values between 0.90 and 1.00 the precision is considered to be excellent, for values between 0.80 and 0.90 it is good, for the range 0.70-0.79 it is fair, it is poor for the range of values between 0.60-0.69, bad for 0.50-0.59 and fail for the values below 0.49. Thus the results obtained with both MSE and MLZC are considered excellent (AUC=0.9991 and 0.9943, respectively). MLZC showed a sensitivity of 100.0%, selectivity of 97.5%, positive predictivity of 86.0% and an accuracy of 97.9% while the MSE showed better results with a sensitivity of 100.0%, selectivity of 99.2%, positive predictivity of 94.9% and an accuracy of 99.3%.

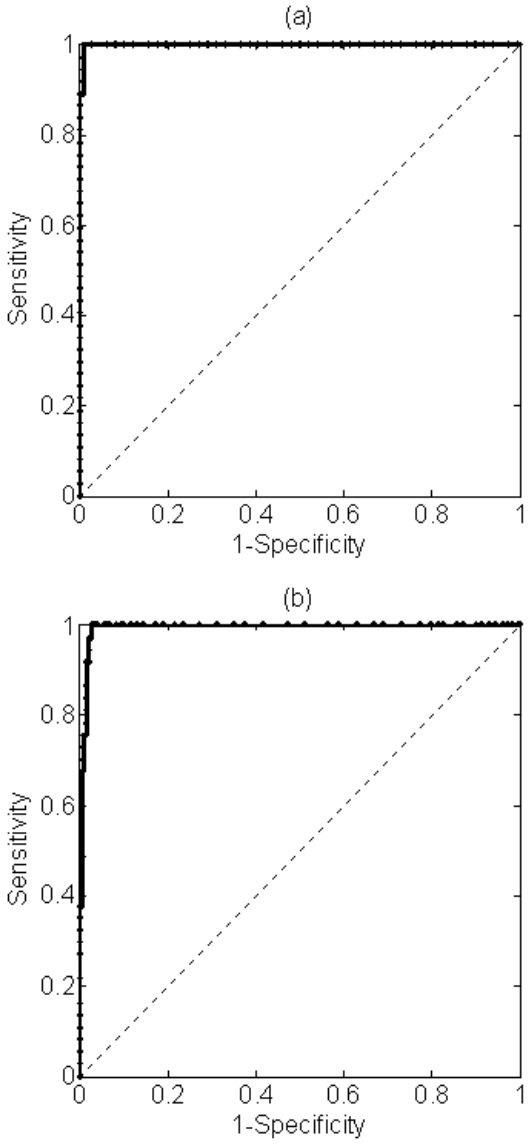

Figure4. ROC plot for discriminating ECG signals of normal and CHF subjects using (a) Modified Shannon entropy and (b) Multi-valued LZC.

Now the results of analyzing HR/RR interval time series from data set two, of the same normal and CHF subjects from data set one, are presented. The HR signals are pre-processed, grouped, and segmented as mentioned in Section 4.a. The same analysis is then applied to the segments from both the groups to determine whether a particular segment belongs to normal, or CHF group. Since the aim of this study is to show that the measures derived from ECG signals sometimes (as shown in this study) can perform better than the same measures obtained from HR signals, only the symbol-sequence statistics (T statistic and the modified χ^2 statistic) are skipped. However, all other results, including comparison statistics of MSE and MLZC are presented.

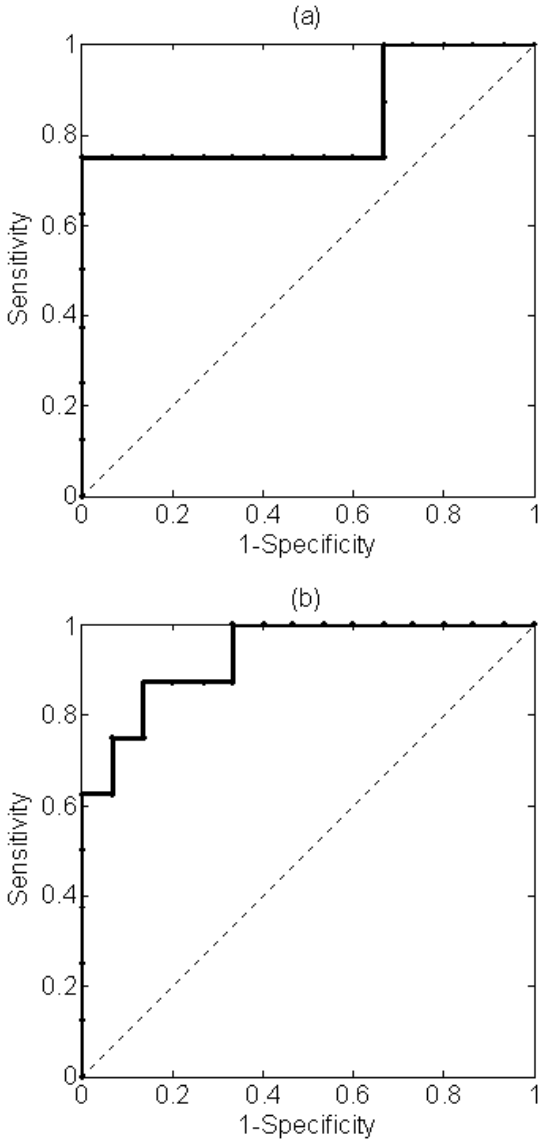

Figure5. ROC plot for discriminating HR signals of normal and CHF subjects using (a) Modified Shannon entropy and (b) Multi-valued LZC.

A comparison statistics of MSE and MLZC (mean ± SD) for HR signals of normal and CHF groups from data set two, are tabulated in Table 4. For the normal group the MSE is 0.6512 ± 0.04472 and for CHF group, it is 0.3576±0.1923, respectively, implying that normal subjects are less informative than CHF subjects. For the normal group the MLZC is 0.3613 ± 0.02753 and for CHF group, MLZC is 0.1147 ± 0.08131, implying that normal subjects have an increased complexity of temporal patterns compared to CHF subjects.

Table3. ROC results for Modified Shannon entropy and Multi-valued LZC between ECG signals of normal and CHF groups

Parameter	AUC	Sensitivity %	Selectivity %	Predictivity % (Positive)	Accuracy %
H_s	0.9991	100	99.2	94.9	99.3
MLZC	0.9943	100	97.5	86.0	97.9

Figs. 5(a) and 5(b) show ROC curves for the MSE and MLZC cases, respectively, for HR signals of normal and CHF groups. Table 5 summarizes the results. Thus the results obtained with both MSE are considered good (AUC=0.8333) while those with MLZC are considered excellent (AUC=0.9333). MSE showed a sensitivity of 75.0%, selectivity of 100.0%, positive predictivity of 100.0% and an accuracy of 91.3% while the MLZC showed better results with a sensitivity of 87.5%, selectivity of 86.7%, positive predictivity of 92.9% and an accuracy of 87.0%. Comparing Tables 3 and 5, it is found that AUC values for MSE and MLZC of ECG signals are higher than the corresponding measures of HR signals.

The important findings of this study are: ECG signals of normal subjects are more deterministic and have decreased complexity of temporal patterns than ECG signals of CHF subjects. ECG signals, both in normal and CHF, exhibit time irreversibility which implies that the signals are nonstationary and the generating cardiac systems are nonlinear. This is in agreement with previous finding [29]. On the other hand, HR signals of normal subjects are less deterministic and have increased complexity of temporal patterns than HR signals of CHF subjects. In the present study there is clear evidence that, many a time, measures derived from ECG signals do perform better than the same measures obtained from HR signals in distinguishing CHF from normal subjects.

Table4. Modified Shannon entropy and Multi-valued LZC (mean ± SD) for HR signals of normal and CHF groups.

Group	H_s (p < 0.0001)	MLZC (p < 0.0001)
Normal	0.6512 ± 0.04472	0.3613 ± 0.02753
CHF	0.3576±0.1923	0.1147 ± 0.08131

Table5. ROC results for Modified Shannon entropy and Multi-valued LZC between HR signals of normal and CHF groups

Parameter	AUC	Sensitivity %	Selectivity %	Predictivity % (Positive)	Accuracy %
H_s	**0.8333**	75	100.0	100.0	91.3
MLZC	0.9333	87.5	86.7	92.9	87.0

Another chief finding of this study is: Persons with cardiac risk show the distribution of length-3 words (with 6 symbols) to be concentrated on about 8 bins (out of 216 bins) where as healthy persons show the distribution being characterized by more number of bins.

7. Conclusion

A new approach to classification of ECG and HR signals using nonlinear symbolic dynamic analysis has been presented. The relative frequency distribution in symbol-sequence histograms reveals significant differences among the normal and CHF classes. The MSE reveals increased randomness and decreased deterministic structure in CHF group compared to normal group (lower randomness and more deterministic structure). MLZC shows decreased complexity in CHF group compared to normal group. Although this nonlinear analysis cannot be used as an exact diagnostic tool, our findings

show the possibility to analyze and compare the cardiac dynamic behavior in normal and CHF patients using MSE and MLZC. Nonlinear dynamics suggests that CHF can be a dynamical disease which is characterized by changes in qualitative dynamics of the related physiological processes. Nevertheless, the presented results of this study show the effectiveness of symbolic dynamics in ECG and HR signal classification into normal and CHF groups. More importantly, the present study shows clear evidence that, many a time, measures derived from ECG signals do perform better than the same measures obtained from HR signals in distinguishing CHF from normal subjects.

8. References

[1] H. Gavin Richard Sandercock, and A. David Brodie, "The Role of Heart Rate Variability in Prognosis for Different Modes of Death in Chronic Heart Failure," Pacing and Clinical Electrophysiology, vol. 79, no. 8, 2006, pp. 892-904.

[2] M. Ferrario, M.G. Signorini, and G. Magenes, "Estimation of long-term correlations from foetal heart rate variability signal for the identification of pathological fetuses," Proc. 29th Ann. Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBS'07), 2007, pp. 295-298.

[3] S. Truebner, I. Cygankiewicz, R. Schroeder, M. Baumert, M. Vallverd_u, P. Caminal, A.Bd. Luna, and A. Voss, "Compression entropy contributes to risk stratification in patients with cardiomyopathy," Biomed. Tech., vol. 51, no. 2, 2006, pp. 77-82.

[4] M. Costa, A.L. Goldberger, and C.-K. Peng, "Multiscale entropy analysis of biological signals," Phys. Rev. E, 71(021906), 2005, pp. 1-18.

[5] P.K. Stein, P.P. Domitrovich, H.V. Huikuri, and R,E. Kleiger, "Traditional and nonlinear heart rate variability are each independently associated with mortality after myocardial infarction," J. Cardiovasc. Electrophysiol., vol. 16, no. 1, 2005, 13-20.

[6] A.L. Goldberger, "Is the normal heartbeat chaotic or homeostatic?," News Physiol Sci., vol. 6, 1991, pp. 87-91.

[7] A.L. Goldberger, "Fractal mechanisms in the electrophysiology of the heart," IEEE Eng Med Biol., vol. 11, no. 2, 1992, pp. 47-52.

[8] N. Wessel, H. Malberg, R. Bauernschmitt, and J. Kurths, "Nonlinear methods of cardiovascular physics and their clinical applicability," International Journal of Bifurcation and Chaos, vol. 17, no. 10, 2007, pp. 3325-3371.

[9] C.S. Daw, C.E.A. Finney, and E.R. Tracy, "A review of symbolic analysis of experimental data," Review of Scientific Instruments, vol. 74, no. 2, 2003, pp. 915-930.

[10] J.H. Xu, Z.R. Liu, and R., Liu, "The measures of sequence complexity for EEG studies," Chaos, vol. 4, no. 11, 1994, pp. 2111-2119.

[11] C.S. Daw, "Observing and modeling nonlinear dynamics in an internal combustion engine," Phys. Rev. Lett., vol. 57, no. 3, 1998, pp. 2811-2819.

[12] C.S. Daw, C.E.A. Finney, K. Nguyen, and J.S. Halow, "Symbol statistics: a new tool for understanding multiphase flow phenomena," International Mechanical Engineering congress and Exposition, 1998, pp. 405-411.

[13] J. Kurths, A. Voss, P. Saparin, A. Witt, H.J. Kleiner, and N. Wessel, "Quantitative analysis of heart rate variability," Chaos, vol. 5, 1995, pp. 88-94.

[14] A. Porta, G. D'Addio, G.D. Pinna, R. Maestri, T. Gnecchi-Ruscone, R. Furlan, N. Montano, S. Guzzetti, and A. Malliani, "Symbolic analysis of 24h Holter heart period variability series: Comparison between normal and heart failure patients," Computers in Cardiology, vol. 32, 2005, pp. 575-578.

[15] K. Tobaldini, A. Porta, S.G. Wei, Z.H. Zhang, J. Francis, K.R. Casali, R.M. Weiss, R.B. Felder, and N. Montano, "Symbolic analysis detects alterations of cardiac autonomic modulation in congestive heart failure rats," Auton Neurosci., vol. 150, no. 1-2, 2009, pp. 21-26.

[16] N. Wessel, U. Schwarz, P.I. Saparin, and J. Kurths, "Symbolic dynamics for medical data analysis," Attractors, Signals, and Synergetics, W. Klonowski (Ed.), Proceedings of EUROATTRACTOR2000, Frontiers on Nonlinear Dynamics, vol. 1, 2002, pp. 45-61.

[17] A. Voss, J. Kurths, H.J. Kleiner, A. Witt, N. Wessel, P. Saparin, K.J. Osterziel, R. Schurath, and R. Dietz, "The application of methods of nonlinear dynamics for the improved and predictive

recognition of patients threatened by sudden cardiac death," Cardiovasc. Res., vol. 31, 1996, pp. 419-433.

[18] U. Parlitza, S. Berg, S. Luthera, A. Schirdewane, J. Kurthsf, and N. Wesself, "Classifying Cardiac Biosignals using Ordinal Pattern Statistics and Symbolic Dynamics," Computers in Biology and Medicine, vol. 43, no. 3, 2012, pp. 319-327.

[19] A. Voss, M. Baumert, V. Baier, S. Trübner, and A. Schirdewan, "Analysis of heart rate dynamics before ventricular tachyarrhythmia and in patients with DCM based on the compression entropy," 11th Congress of the International society of Holter and noninvasive electrocardiology, 2005, pp. 428-430.

[20] X.Z. Tang, E.R. Tracy, A.D. Boozer, A. deBrauw, and R. Brown, "Symbol sequence statistics in noisy chaotic signal reconstruction," Physical Review E., vol. 51, no. 5, 1995, pp.3871-3889.

[21] M. Tavassoli, M.M. Ebadzadeh, and H. Malek. "Classification of cardiac arrhythmia with respect to ECG and HRV signal by genetic programming," Canadian Journal on Artificial Intelligence, Machine Learning and Pattern Recognition, vol. 3, no. 1, 2012, pp. 1-8.

[22] A.L. Goldberger, L.A.N. Amaral, L. Glass, J.M. Hausdorff, P.Ch. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, and E.H. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals," Circulation, vol. 101, 2000, pp. e215--e220. (Available at http://circ.ahajournals.org/ cgi/content/full/101/23/e215).

[23] N. Wessel, C. Ziehmann, J. Kurths, U. Meyerfeldt, A. Schirdewan, and A. Voss, "Short-term forecasting of life-threatening cardiac arrhythmias based on symbolic dynamics and finite-time growth rates," http://www.ncbi.nlm.nih.gov/pubmed/11046317 vol. 61, no. 1, 2000, pp. 733-739.

[24] U. Parlitz, S. Berg, S. Luther, A.J. Schirdewan, and N. Wessel, "Classifying cardiac biosignals using ordinal pattern statistics and symbolic dynamics," Computers in Biology and Medicine, vol. 42, no. 3, 2012, pp. 319-327.

[25] R.A. Stepien, "New method for analysis of nonstationary signals," *Nonlinear Biomedical Physics*, vol. **5**, no. 3, 2011, pp. 1-8.

[26] C. Gómez, and R. Hornero, "Entropy and Complexity Analyses in Alzheimer's disease: An MEG Study," The Open Biomedical Engineering Journal, vol. 4, 2010, 223-235.

[27] C. Diks, J.C. van Houwelingen, F. Takens, and J. DeGoede, "Reversibility as a criterion for discriminating time series," Physics Letters A, vol. 201, 1995, pp. 221-228.

[28] C.E.A. Finney, K. Nguyen, C.S. Daw, J.S. and Halow, "Symbol-sequence statistics for monitoring fluidization," *Proceedings of the ASME Heat Transfer Division, 1998, pp.* 405-411.

[29] M. Costa, A.L. Goldberger, and C.-K. Peng, "Broken asymmetry of the human heartbeat: loss of time irreversibility in aging and disease," Phys. Rev. Lett., 2005, pp. 198102-1-198102-4.