Automated Detection of Paroxysmal Atrial Fibrillation Using an Information-Based Similarity Approach

Xingran Cui [1],* (ID) **, Emily Chang [2], Wen-Hung Yang [3], Bernard C. Jiang [4], Albert C. Yang [5] and Chung-Kang Peng [5],*** (ID)

[1] School of Biological Sciences & Medical Engineering, Southeast University, Nanjing 210000, China
[2] Departments of Computer Science and Biology, Emory University, Atlanta, GA 30322, USA; eyjchang@gmail.com
[3] Department of Industrial Engineering and Management, Yuan Ze University, Taoyuan 320, Taiwan; iewhyang@gmail.com
[4] Industrial Management Department, National Taiwan University of Science and Technology, Taipei 100, Taiwan; iebjiang777@gmail.com
[5] Center for Dynamical Biomarkers, Beth Israel Deaconess Medical Center/Harvard Medical School, Boston, MA 02215, USA; accyang@gmail.com
* Correspondence: cuixr@seu.edu.cn (X.C.); cpeng@bidmc.harvard.edu (C.-K.P.); Tel.: +86-25-8379-2349 (X.C.); +1-617-667-7122 (C.-K.P.)

Abstract: Atrial fibrillation (AF) is an abnormal rhythm of the heart, which can increase heart-related complications. Paroxysmal AF episodes occur intermittently with varying duration. Human-based diagnosis of paroxysmal AF with a longer-term electrocardiogram recording is time-consuming. Here we present a fully automated ensemble model for AF episode detection based on RR-interval time series, applying a novel approach of information-based similarity analysis and ensemble scheme. By mapping RR-interval time series to binary symbolic sequences and comparing the rank-frequency patterns of m-bit words, the dissimilarity between AF and normal sinus rhythms (NSR) were quantified. To achieve high detection specificity and sensitivity, and low variance, a weighted variation of bagging with multiple AF and NSR templates was applied. By performing dissimilarity comparisons between unknown RR-interval time series and multiple templates, paroxysmal AF episodes were detected. Based on our results, optimal AF detection parameters are symbolic word length $m = 9$ and observation window $n = 150$, achieving 97.04% sensitivity, 97.96% specificity, and 97.78% overall accuracy. Sensitivity, specificity, and overall accuracy vary little despite changes in m and n parameters. This study provides quantitative information to enhance the categorization of AF and normal cardiac rhythms.

Keywords: paroxysmal atrial fibrillation; RR-interval time series; symbolic sequence; information-based similarity index; ensemble model

---

1. Introduction

Atrial fibrillation (AF), the most common sustained cardiac arrhythmia, is an abnormal heart rhythm characterized by rapid and irregular beating of the atria [1]. The disease is associated with an increased risk of heart failure, dementia, stroke and other heart-related complications [2]. Paroxysmal AF (PAF), also termed intermittent AF, is defined as an episode of AF that terminates spontaneously or with intervention in less than seven days [3]. The frequency of PAF is uncertain, because previous studies have suggested that a majority of these episodes are asymptomatic [4,5], including some that may last more than 48 h [4]. Experienced clinicians can identify AF patterns by visual inspection

of the electrocardiogram (ECG) chart. However, due to the paroxysmal nature of the onset and termination of PAF in certain patients, human-based diagnosis of AF is usually time consuming when using a longer-term ECG recording such as a Holter or event recorder. Therefore, an automated, computerized AF detector may provide timely diagnosis and have substantial clinical utility.

It is challenging to implement diagnostic ECG waveform criteria for AF into a computerized algorithm, partly due to the difficulty of quantifying P-waves (and their absence) and that cardiac inter-beat intervals, i.e., RR intervals, follow no repetitive patterns. One feasible approach is to identify the presence of irregular ventricular rhythm during AF episodes based on analysis of RR intervals [6–25]. Moody and Mark [11] showed that the Markov process model for the AF detection is equivalent to determining the arithmetic mean of a series scores based on the RR interval sequence. Tateno and Glass applied standard density histograms of the RR and ΔRR intervals to detect the onset and termination of AF using standard coefficients of variation and the Kolmogorov–Smirnov test [12]. Kikillus, et al. [13] and Babaeizadeh, et al. [14] applied Markov modeling technique to identify AF. Lian, et al. [15] developed an AF detector with its basis centered on the Map of RR intervals versus change of RR intervals. Huang, et al. [16] utilized a histogram of ΔRR_n and standard deviation analysis. Parvaresh, et al. [17] evaluated three classifiers for AF screening by using autoregressive modeling. Dash, et al. [18] proposed the randomness-variability-complexity approach. Lee, et al. [19] introduced the time-varying coherence approach. Petrenas, et al. [20] applied the low-complexity approach. Zhou, et al. [21,22] proposed the symbolic dynamics approach. Additionally, the entropy and heart rate dynamics approaches were also used in many studies [23–26]. Other studies of AF detections are based on the analysis of the probability density or autocorrelation function of the RR interval series during AF [27–29].

Studies have demonstrated that physiologic systems generate complex fluctuations in their output signals that reflect the underlying dynamics [30–32]. We have previously proposed a novel information-based similarity (IBS) index to detect and quantify the repetitive appearance of certain basic patterns that are embedded in the human heart rate time series using tools from physics and statistical linguistics [33–36]. Human cardiac dynamics are driven by the complex nonlinear interactions of two competing forces: Sympathetic stimulation increases and parasympathetic stimulation decreases heart rate. For this type of intrinsically noisy system, it may be useful to simplify the dynamics via mapping the output to binary sequences, where the increase and decrease of the inter-beat intervals are denoted by 1 and 0, respectively. The resulting binary sequence retains important features of the dynamics generated by the underlying control system, but is tractable enough to be analyzed as a symbolic sequence [33–38] Therefore, analysis of symbolic sequences derived from RR intervals may reveal hidden physiological properties of AF. We hypothesize that symbolic patterns mapped from fluctuations of the RR time series may contain important information representing the underlying dynamics, which can be used to discriminate AF and non-AF ventricular rhythms.

Therefore, here we present a study based on a public ECG database on PhysioNet (http://physionet.org) [39–41]. We aim to develop a computerized AF detector based on quantifying the dissimilarity of AF and normal RR-interval time series using an information-based approach. To achieve high detection specificity and sensitivity, and low variance, we designed an ensemble AF detection model. A weighted variation of bagging with multiple AF and normal sinus rhythm (NSR) templates was applied. With respect to the setting of parameters and selection of training data, this study provides quantitative information to enhance the categorization of AF and normal RR time series.

2. Materials and Methods

2.1. Data

The ECG signals used in this study were taken from PhysioNet MIT-BIH Atrial Fibrillation Database (AFDB: http://www.physionet.org/physiobank/database/afdb/) and MIT-BIH Normal

Sinus Rhythm Database (NSRDB: https://www.physionet.org/physiobank/database/nsrdb/). The AF database consists of 25 ECG recordings (10 h in duration) of patients with Paroxysmal AF, whereas the NSR database consists of 18 long-term ECG recordings from healthy subjects who had no significant arrhythmias [39–41].

Figure 1 illustrates a typical tracing of RR interval time series for a PAF patient from the MIT-BIH AFDB Database. Visual inspection suggests that the cardiac rhythm changes dramatically with the AF onset, and the amplitude of the RR-interval fluctuations in AF episodes is substantially higher than that in non-AF periods.
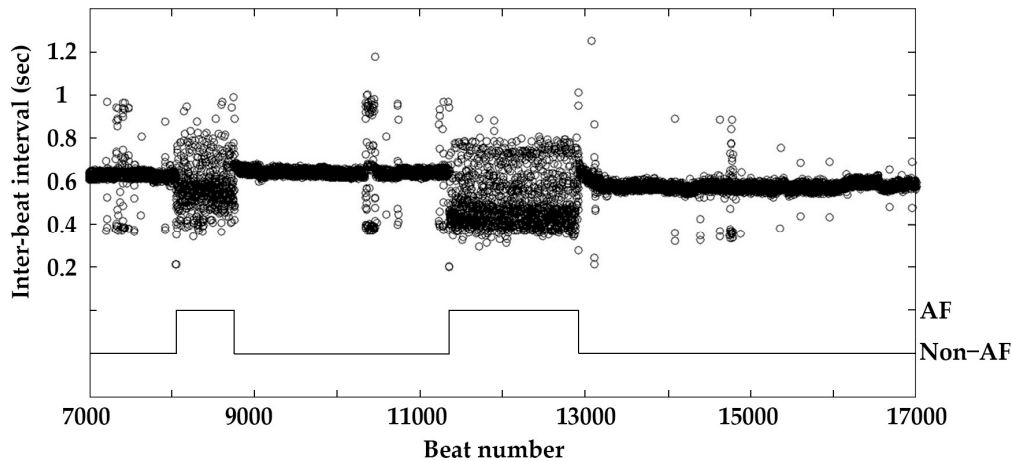

Figure 1. Representative inter-beat (RR) interval time series derived from an electrocardiographic recording of a patient with paroxysmal atrial fibrillation. The dark circles represent consecutive RR intervals and the solid line indicates the presence/absence of AF episodes as reported in the annotations of PhysioNet database. During an episode of atrial fibrillation, the line is set to "AF"; otherwise it is set to "Non-AF", which means a rhythm that is not atrial fibrillation.

2.2. Information-Based Similarity Index

We have previously proposed an algorithm to measure the distance or dissimilarity between two symbolic sequences [33–36]. The algorithm is based on measuring differences in the occurrence of repetitive patterns between two symbolic sequences. In this study, the RR-interval time series was mapped to a binary symbolic sequence, where an increase in the RR-interval was represented by '1' and no change or a decrease in the RR-interval was represented by '0'. We map $m + 1$ successive intervals to a binary sequence of length m, called an m-bit "word". Each m-bit word, therefore, represents a unique pattern of fluctuations in a given RR-interval time series. By shifting one data point at a time, the algorithm produces a collection of m-bit words over the whole time series (total of 2^m possible words). Therefore, it is plausible that the occurrence of these m-bit words reflects the underlying dynamics of the original RR time series. Different patterns of dynamics thus produce different distributions of these m-bit words.

Figure 2 illustrates this mapping procedure using 6-bit words ($m = 6$) from a part of the RR-interval time series. For $m = 6$, there are a total of 64 (=2^6) possible words. The first binary word (100100) shown in Figure 2 is equivalent to decimal number of 36 ($1 \times 2^5 + 1 \times 2^2 = 36$), so as (001001) and (010010) are termed 9 and 18 respectively.
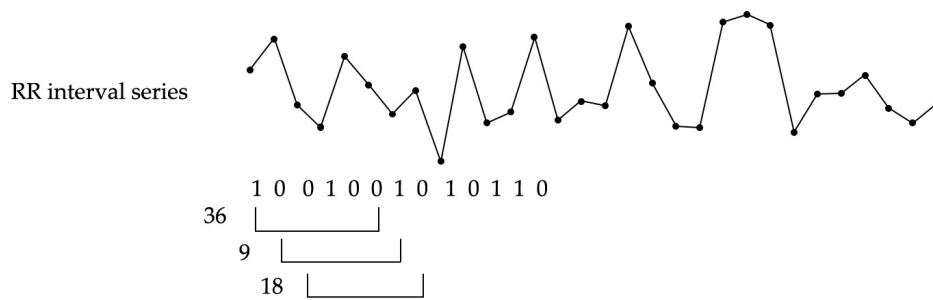
Figure 2. Schematic illustration of the mapping procedure for 6-bit words ($m = 6$).

These m-bit words are then sorted according to their frequency of occurrence. The rank-frequency of any given m-bit word may differ between the two sequences mapped from two RR interval time series. We then plot the rank order of each m-bit word in the first symbolic sequence against its rank order in the second symbolic sequence (Figure 3). Each data point on the graph represent a binary word with its rank on first symbolic sequence (horizontal axis) plotted against that on second symbolic sequence on vertical axis. The diagonal line of identity indicates equal rank order for both signal series. If two symbolic sequences are similar in their rank order, the data points will be located near this diagonal line (Figure 3c,d), comparisons between RR time series for either two AF patients or two healthy subjects). The average deviation of the plotted points from the dashed diagonal line is, therefore, a measure of the distance between two symbolic sequences. Greater distance indicates less similarity (Figure 3e, comparison between AF and normal RR time series), and, vice versa.
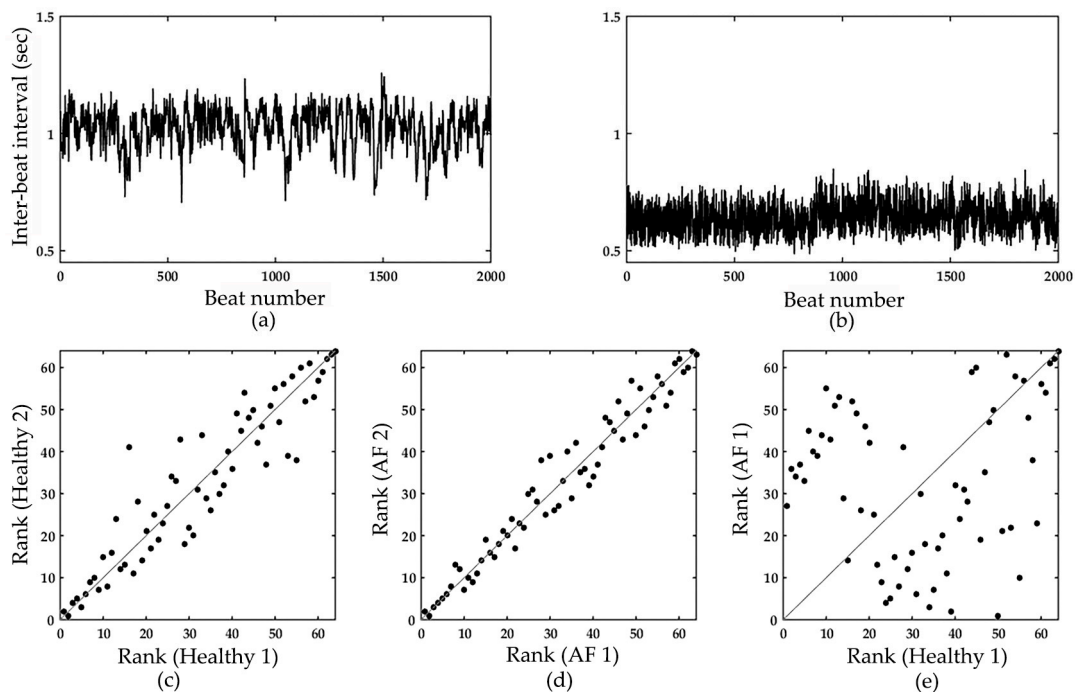
Figure 3. Representative inter-beat time series for a healthy subject (**a**) and an AF patient (**b**); (**c**) Rank order comparison of the time series for two healthy subjects; (**d**) Rank order comparison of the time series for two AF patients; (**e**) Rank order comparison of the time series in (**a,b**). The results in (**c**–**e**) are for the case $m = 6$.

Let $d_r(\psi_1, \psi_2)$ denote a dissimilarity value between 0 and 1 of two symbolic sequences, s_k denote an m-bit word, and L denote the number of unique m-bit words. Let R and p denote the word's rank and probability, respectively. Let F denote the weight of the word, where F is computed using Shannon's

entropy and normalized with the normalization factor Z. The degree of dissimilarity between two symbolic sequences can be defined as [16–18]:

$$d_r(\psi_1, \psi_2) \; = \; \frac{1}{L} \sum_{k=1}^{L} |R_1(s_k) - R_2(s_k)| F(s_k) \tag{1}$$

$$F(s_k) \; = \; [-p_1(s_k)\log p_1(s_k) - p_2(s_k)\log p_2(s_k)]/Z \tag{2}$$

$$Z \; = \; \sum_{k=1}^{L} [-p_1(s_k)\log p_1(s_k) - p_2(s_k)\log p_2(s_k)] \tag{3}$$

The sum is divided by the value L to keep $d_r(\psi_1, \psi_2)$ in the range [0, 1]. A bigger dissimilarity value corresponds to a higher degree of dissimilarity. Therefore, if for an unknown RR-interval series, $d_N > d_{AF}$, where d_N is the dissimilarity value between unknown series and normal RR series, and d_{AF} is the dissimilarity value between unknown series and AF RR series, the unknown series is more similar to AF, and vice versa.

2.3. Ensemble Model of Automated AF Detector

2.3.1. Overall Algorithm of the Ensemble Model

The development of this proposed ensemble AF detector follows five key steps:

1. Retrieving sets of AF and NSR RR-interval series from PhysioNet ECG data;
2. randomly setting aside a percentage of AF and NSR sets as training data and the rest as testing data. In AF training data, only AF segments were picked according to annotations provided on the PhysioNet database. This procedure was repeated five times such that five datasets (i.e., datasets 1–5) with different training and testing data could be generated;
3. extracting RR-interval increment signatures (i.e., the rank-frequency of m-bit word) of the desired observation window length from training and testing data, respectively;
4. building templates to represent AF and NSR signature patterns;
5. designing an ensemble classifier, which is composed of various pairs of AF and NSR templates;
6. comparing the information-based dissimilarity index between an unknown RR-interval time series and the templates; and
7. tuning ensemble parameters to achieve our twin aims of high detection accuracy and low detection variance.

2.3.2. Weighted Dissimilarity Index

Since cardiac patterns can be quite variable between different subjects and even within the same subject, the single template approach of averaging multiple and lengthy ECG signals together could result in dilution of significant patterns and data. We thus propose creating multiple AF and NSR templates from shorter observation-window segments and incorporating the ensemble method to obtain better predictive performance for detecting AF. Furthermore, we propose using a weighted variation of bootstrap aggregating (bagging) to perform weighted voting when comparing an unknown RR-interval time series with AF and NSR templates.

In the current study, we randomly generated 10 AF templates and 15 NSR templates in each dataset. More NSR templates were created due to the greater variability of NSR beat patterns between healthy subjects. Each template was created from a set of observation-window segments extracted from the original RR time series of one subject. To assess accuracy, we tested AF RR time series that were not part of the training data. We analyzed these series and compared their diagnoses to the ground truth provided by annotations from the PhysioNet database. Among the 10 AF and 15 NSR templates, all possible template pair combinations were compared against each testing RR-interval

time series segment for a total of 150 dissimilarity comparisons. Each pair of dissimilarities was normalized such that $\widetilde{d}_{AF} + \widetilde{d}_N = 1$. Weighted sums of the dissimilarity values were then calculated for AF and NSR comparisons and averaged to produce a weighted average AF and NSR dissimilarity values, i.e., weighted dissimilarity index:

$$D_N = \frac{1}{T} \sum_{i=1}^{T} \widetilde{d}_N(i) \tag{4}$$

$$D_{AF} = \frac{1}{T} \sum_{i=1}^{T} \widetilde{d}_{AF}(i) \tag{5}$$

where $\widetilde{d}_N$ and $\widetilde{d}_{AF}$ represent the normalized values of d_N and d_{AF}, D_N and D_{AF} represent the weighted average of $\widetilde{d}_N$ and $\widetilde{d}_{AF}$. T represents the total number of dissimilarity pair comparisons. In this study, $T = 150$.

2.3.3. Parameter Tuning

In addition to the *m* and *n* parameters, we include a tuning parameter Δ to optimize our final results. The final step in our computation is to compare weighted averages of AF and NSR dissimilarity values, where the smaller dissimilarity index determines the predicted diagnosis of a given test segment. While AF detection accuracy remains high, with optimum accuracies between 98–100%, the NSR accuracy can fall as low as 70% to 80%. When observing the normalized dissimilarity values, many of the false positive diagnoses during NSR testing had very close AF and NSR dissimilarity values that were close to a 50/50 weighting (e.g., a 0.49 weighted AF dissimilarity index compared to a 0.51 NSR weighted dissimilarity index). Although the dissimilarities were close in value, our greater/less than comparison scheme caused many NSR segments to be falsely diagnosed as AF. Using a bias factor to adjust decision boundary is a well-established statistical method when training data may exhibit imbalanced distribution [42]. As a result, we implemented a tuning parameter Δ to shift the dissimilarity comparison boundary and yield more accurate results, where if $D_N > D_{AF} + \Delta$, then the segment is AF. Otherwise, the segment is non-AF. For each *m*-bit word and observation window *n* combination, we tested Δ values between 0.00 and 0.19 to find the optimum sensitivity and specificity combination.

Thus, in our ensemble model, we use three parameters: the word length *m*, observation window *n*, and a bias parameter Δ. We experimented with *m* from 4 to 12, n from 50 to 200, and Δ of different values. We aimed to find the best parameter setting(s) and sensitivity of different parameter settings. To evaluate the predictive performance of our ensemble model, the sensitivity (SEN), specificity (SPE) and overall accuracy (ACC) were calculated, and repeated cross-validation were performed. SEN is defined as (True Positive)/(True Positive + False Negative), SPE is defined as (True Negative)/(True Negative + False Positive), and ACC is defined as (True Negative + True Positive)/(True Negative + True Positive + False Negative + False Positive).

3. Results

3.1. Overall Performance of the Ensemble Model

Our ensemble model achieved great performance. Table 1 summarizes the results of SEN, SPE and ACC of the prediction models with different cross-validation datasets. Optimal AF-detection parameters are *m* = 9 and observation window of 150, achieving 97.04% sensitivity, 97.96% specificity, and 97.78% overall accuracy. SEN, SPE, and overall ACC vary little despite changes in word length *m*, observation window *n* and cross-validation datasets (see Table 1). The performance of the ensemble model (*m* = 9, *n* = 150) at changing the tuning parameter Δ is shown in Table 2. The ensemble model had best performance with Δ = 0.08, achieving 96.30% sensitivity, 98.71% specificity and 98.41% accuracy.

Table 1. Detection performances (%) of changing *m* and *n* parameters.

m	Repeated Cross-Validation	Observation Window Size (*n*)											
		50			100			150			200		
		SEN	SPE	ACC	SEN	SPE	ACC	SEN	SPE	ACC	SEN	SPE	ACC
4	Dataset 1	86.34	86.46	86.44	93.06	93.49	93.43	94.74	94.14	94.21	96.62	96.28	96.32
	Dataset 2	85.71	87.09	86.15	91.17	92.00	91.66	94.12	93.63	93.90	95.02	95.11	95.08
	Dataset 3	83.60	86.08	85.59	90.61	89.05	89.99	95.28	96.02	95.71	90.33	92.50	91.87
	Dataset 4	86.49	86.03	86.55	92.67	93.02	92.80	93.06	93.22	93.19	92.49	93.44	93.06
	Dataset 5	85.09	84.66	84.86	91.93	90.95	91.25	95.19	96.79	96.20	93.88	95.38	94.91
	Average	85.45	86.06	85.92	91.89	91.70	91.83	94.48	94.76	94.64	93.67	94.54	94.25
5	Dataset 1	86.34	85.15	85.34	93.06	93.06	93.06	95.32	94.73	94.81	96.62	95.92	96.01
	Dataset 2	86.45	87.85	87.09	89.85	90.70	90.11	94.49	94.05	94.20	91.38	92.12	92.03
	Dataset 3	87.57	86.54	86.97	92.33	93.37	92.90	95.02	96.12	95.86	95.77	95.26	95.50
	Dataset 4	85.23	84.83	85.01	92.28	92.55	92.44	93.66	94.52	94.20	96.13	94.94	95.36
	Dataset 5	88.00	88.52	88.43	93.50	92.71	93.25	94.47	94.71	94.55	95.32	95.88	95.72
	Average	86.72	86.58	86.57	92.20	92.48	92.35	94.59	94.83	94.72	95.04	94.82	94.92
6	Dataset 1	87.39	88.11	88.00	93.27	93.31	93.30	96.49	95.16	95.32	96.14	96.99	96.88
	Dataset 2	87.33	86.40	86.94	95.44	96.04	95.81	94.81	95.29	95.15	94.65	95.08	94.89
	Dataset 3	90.88	91.98	91.60	94.18	94.05	94.07	96.06	96.57	96.44	96.77	96.51	96.60
	Dataset 4	86.47	85.90	86.07	91.46	92.69	92.50	95.90	96.47	96.39	95.02	96.33	96.19
	Dataset 5	87.89	87.66	87.72	91.07	92.33	92.26	96.32	95.90	96.02	94.94	95.38	95.26
	Average	87.99	88.01	88.07	93.08	93.68	93.59	95.92	95.88	95.86	95.50	96.06	95.96
7	Dataset 1	89.59	89.47	89.49	94.69	94.03	94.13	96.78	95.65	95.79	97.10	97.06	97.07
	Dataset 2	88.04	87.23	87.40	91.56	92.86	92.42	96.13	96.60	96.41	94.44	93.95	94.10
	Dataset 3	90.14	91.29	90.25	95.22	95.70	95.58	97.46	98.39	98.18	95.45	95.02	95.13
	Dataset 4	87.37	87.05	87.10	94.31	94.57	94.51	95.08	95.23	95.17	95.48	95.66	95.62
	Dataset 5	90.51	90.26	90.29	93.06	92.91	92.96	97.58	98.41	98.33	94.87	96.81	96.67
	Average	89.13	88.86	88.91	93.77	94.01	93.92	96.61	96.86	96.78	95.47	95.70	95.72
8	Dataset 1	90.54	89.69	89.83	95.51	95.36	95.38	97.08	97.10	97.08	97.10	97.92	97.82
	Dataset 2	89.64	89.22	89.30	95.08	95.27	95.22	95.20	99.04	97.27	97.15	98.58	98.40
	Dataset 3	89.01	90.85	90.24	96.49	96.44	96.45	96.37	97.11	97.66	95.48	95.70	95.61
	Dataset 4	88.46	90.80	90.03	95.68	96.02	95.97	96.68	98.00	97.38	95.20	96.49	96.33
	Dataset 5	90.25	91.02	90.79	96.91	96.77	96.80	95.51	96.78	96.42	96.68	97.00	96.92
	Average	89.58	90.32	90.04	95.93	95.97	95.96	96.17	97.61	97.16	96.32	97.14	97.02
9	Dataset 1	90.35	89.2	89.38	95.71	95.18	95.26	98.25	97.96	98.01	97.58	97.85	97.82
	Dataset 2	92.09	88.25	89.06	97.47	97.18	97.01	97.36	97.79	97.70	93.92	95.06	94.77
	Dataset 3	89.53	90.87	90.35	94.72	97.09	96.39	96.30	98.71	98.41	93.99	97.02	96.68
	Dataset 4	87.60	89.33	89.02	94.76	98.10	97.06	96.95	97.28	97.11	95.05	96.44	96.11
	Dataset 5	86.44	85.98	86.60	94.65	96.28	95.79	96.33	98.05	97.66	97.59	97.99	97.83
	Average	89.20	88.73	88.88	95.46	96.77	96.30	97.04	97.96	97.78	95.63	96.87	96.64
10	Dataset 1	89.11	89.73	89.63	96.33	95.11	95.29	97.95	96.77	96.94	97.58	97.99	97.94
	Dataset 2	89.45	90.92	90.84	95.73	95.88	95.84	94.69	95.66	95.28	95.95	97.93	96.64
	Dataset 3	88.15	87.32	88.03	95.33	94.25	95.02	96.26	98.31	97.81	92.55	95.05	93.13
	Dataset 4	85.06	87.28	86.37	96.85	96.20	96.46	96.64	97.53	97.19	97.26	95.05	95.81
	Dataset 5	86.40	85.91	86.39	97.01	98.33	98.17	96.95	98.03	97.72	95.28	96.33	95.65
	Average	87.27	87.86	87.91	96.25	95.95	96.16	96.14	97.38	97.00	96.03	95.47	95.31
11	Dataset 1	88.78	88.45	88.66	95.51	94.75	94.86	97.66	97.53	97.55	97.58	97.78	97.88
	Dataset 2	84.54	83.97	84.12	93.77	93.92	93.89	97.80	97.12	97.25	95.55	96.78	96.42
	Dataset 3	85.37	84.90	84.78	95.76	94.88	95.05	98.12	97.57	97.68	95.92	95.40	95.51
	Dataset 4	87.77	87.93	87.87	95.15	94.79	94.90	98.04	97.60	97.81	97.99	97.29	97.45
	Dataset 5	89.76	89.45	89.52	95.30	95.96	95.88	95.88	95.26	95.34	96.64	95.90	96.07
	Average	87.24	86.94	86.99	95.10	94.86	94.92	97.50	97.02	97.13	96.74	96.63	96.67
12	Dataset 1	87.93	85.78	86.61	93.67	94.57	94.44	96.78	97.04	96.99	97.10	96.92	96.94
	Dataset 2	86.10	86.82	86.67	96.36	95.33	95.45	95.45	96.23	96.06	97.13	97.64	97.55
	Dataset 3	84.12	85.37	85.04	95.66	95.06	95.21	95.91	95.77	95.81	95.48	95.35	95.40
	Dataset 4	84.87	83.52	83.66	92.50	93.02	92.91	96.86	95.98	96.30	95.52	96.06	96.13
	Dataset 5	85.09	87.70	87.22	93.48	94.36	94.15	97.19	97.07	97.10	95.04	96.17	95.93
	Average	85.62	85.84	85.84	94.33	94.47	94.43	96.44	96.42	96.45	96.05	96.43	96.39

Table 2. Performance of the ensemble model (*m*= 9, *n*= 150) at changing the tuning parameter Δ.

Tuning Parameter Δ	Detection Performance (%)		
	SEN	SPE	ACC
0.00	99.51	96.34	96.73
0.01	99.35	96.67	97.00
0.02	99.19	97.04	97.31
0.03	98.72	97.36	97.53
0.04	98.32	97.74	97.81
0.05	97.77	98.05	98.02
0.06	97.31	98.24	98.13
0.07	96.86	98.59	98.38
0.08	96.30	98.71	98.41
0.09	95.40	98.76	98.34
0.10	94.11	98.80	98.22
0.11	92.99	98.84	98.12
0.12	91.60	98.88	97.98
0.13	90.39	98.92	97.86
0.14	89.10	98.96	97.73
0.15	87.87	99.05	97.66
0.16	86.54	99.24	97.67
0.17	83.03	99.59	97.54
0.18	76.75	99.71	96.86
0.19	70.90	99.76	96.18

3.2. Continuous Behavior of the Detector

Here we present a graphic description of the testing results using the proposed ensemble detection model for the case of *m* = 9, the observation window *n* = 150, and Δ = 0.1 (see Figure 4). The representative testing data in Figure 4 was taken from PhysioNet MIT-BIH AFDB database, record number 04908. The upper panel in Figure 4a displays the testing data, i.e., 10-h raw RR-interval time series recorded from a PAF patient. The lower panel in Figure 4a displays the testing results: D_{AF} (the black curve) represents the weighted average dissimilarity index between testing RR-interval time series and AF templates and D_N (the red curve) represents weighted average dissimilarity index between testing RR-interval time series and NSR templates. $D_{AF} + 0.1 > D_N$ indicates the dynamic patterns of the testing RR intervals are similar to normal beats, otherwise, the testing RR intervals are similar to AF beats. The black step line indicates the ground truth of the AF episodes (marked as AF or non-AF on y axis) as reported in the annotations of PhysioNet database, and the detection results of testing data are shown in the red step line, achieving 94.40% sensitivity, 99.59% specificity, and 97.01% overall accuracy. Moreover, we enlarge a representative AF segment (see the red rectangle in Figure 4a) to show the detailed fluctuations of AF episode (see Figure 4b). In addition to normal beats, our detection model can successfully distinguish AF episode from non-AF fluctuations, e.g., the segment of beat number 18,800–22,100 in testing RR-interval time series in Figure 4a, which might be caused by R peak detection failure or other problems in this segment. Compared with AF signal, the time series inside of the blue rectangle in Figure 4a was enlarged to show the signal details, see Figure 4c.

In case of sporadic AF episodes of very short duration (e.g., less than 30 s), performance of the proposed detection model is not satisfying. Here we take record 04043 from MIT-BIH AFDB database as representative example (see Figure 5). The testing RR-interval time series are shown in the upper panel. With *m* = 9, *n* = 150, and Δ = 0.05, the weighted average dissimilarity index D_{AF} (the black curve) and D_N (the red curve) are calculated and shown in the lower panel. $D_{AF} + 0.05 > D_N$ indicates the testing RR intervals are similar to normal beats, otherwise, they are similar to AF beats. The black step line indicates the ground truth, and the red step line indicates detection results (AF or non-AF). The detection performance for this record is 89.55% sensitivity, 67.43% specificity and 72.05% accuracy. The solid and hollow triangles marked on the x axis show some examples of false positive and false negative brief segments.
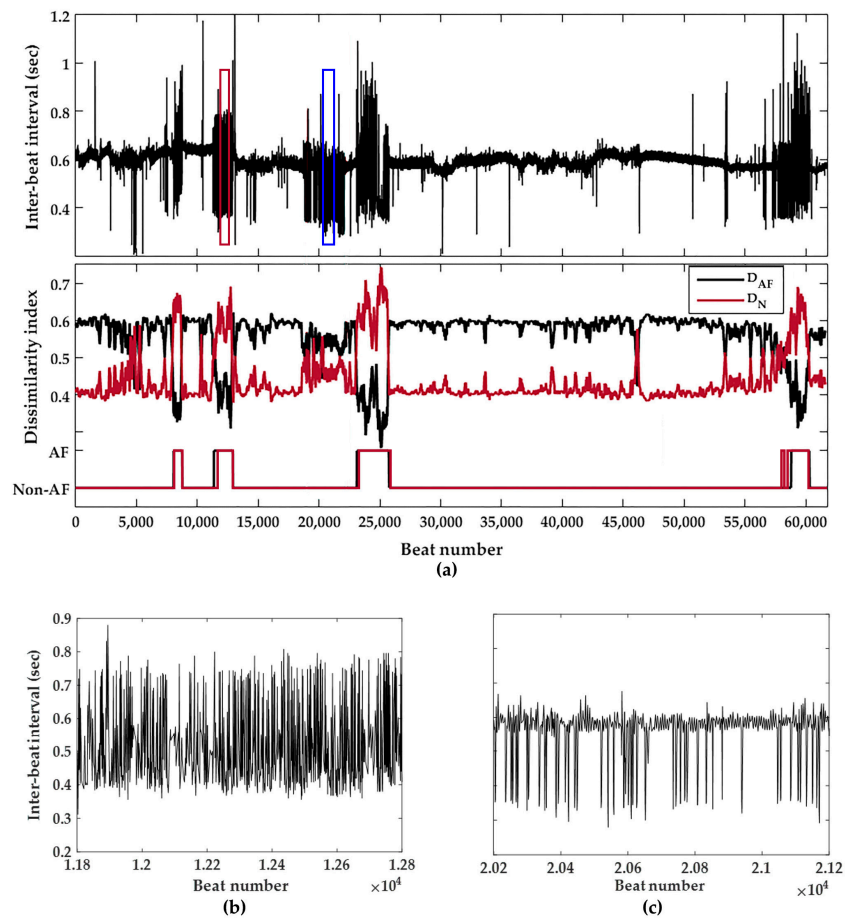
Figure 4. (**a**) Graphic illustration of detection results for a testing data (record 04908, *m* = 9, observation window *n* = 150, Δ = 0.1); (**b**) enlarged AF segment derived from (**a**); (**c**) Enlarged signal segment of neither AF nor normal beats to compare with (**b**).
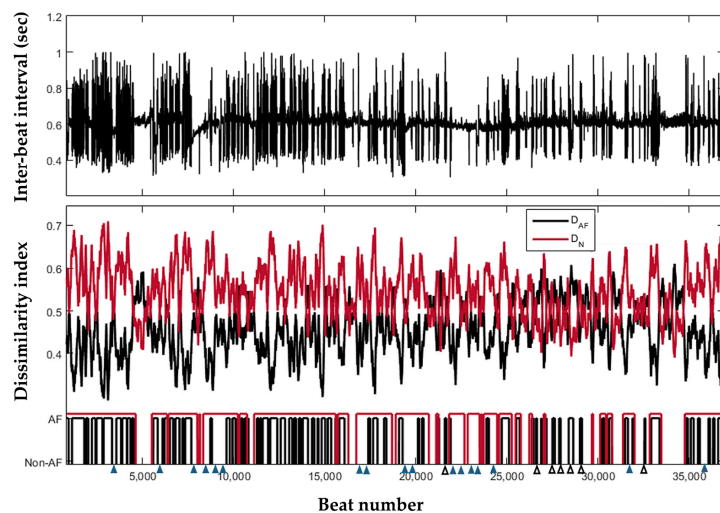

Figure 5. Graphic illustration of detection results for a testing data (record 04043, *m* = 9, *n* = 150, Δ = 0.05).

4. Discussion

4.1. Main Findings

In this study, we present a fully automated ensemble model for AF episode detection based on RR-interval time series, applying a novel approach—information-based similarity index and ensemble scheme. By mapping RR time series to binary sequences and comparing the rank-frequency patterns of *m*-bit word, this study provides quantitative information to enhance the categorization of AF and normal cardiac rhythms. In addition, using a weighted variation of bagging with multiple AF and NSR templates, we can obtain results with low variance and high accuracy. By performing dissimilarity comparisons across multiple templates, we are able to account for RR-interval increment variations between different subjects. Based on our results, optimal AF-detection parameters are symbolic word length $m = 9$ and observation window $n = 150$, achieving 97.04% sensitivity, 97.96% specificity, and 97.78% overall accuracy. Sensitivity, specificity, and overall accuracy vary little despite changes in *m* and *n* parameters. Our findings indicate that the information-based similarity index is relatively reliable in distinguishing AF episode within considerably long time ECG recordings.

4.2. Advantages of Ensemble Model

AF and NSR patterns may not be consistent among different subjects. In addition, not only do different patients exhibit different patterns, but ECG fluctuates from the same subject throughout different activities, such as during wakefulness versus during sleep. This first observation leads us to design multiple templates for each target class to sufficiently represent AF and NSR patterns. Second, a final class detection, AF or NSR, should be a joint decision made by a committee (or an ensemble), with each member consisting of an AF and NSR template pair. The final decision is made through voting by the committee members. When a committee member strongly endorses one class over the other, that member's vote should be weighted higher compared to the vote of a "lukewarm" committee member.

This observation motivated us to design a weighted voting scheme, which enforces high confidence votes but discounts low confidence votes. The empirical study shows that such an ensemble scheme reduces detection noise and hence leads to lower detection variance.

4.3. Comparison with Published Works

Many algorithms have been developed to detect AF based in RR interval variability [11–25]. Here, we compare the performance of the proposed detector with the existing algorithms published in the last 10 years, which used the same databases (i.e., the same records from MIT-BIH AFDB and the same reference annotations), and using the same evaluation metrics. Table 3 summarizes the comparison results. More complete investigations are available in [21,22,43] The proposed method achieve 97.04% sensitivity, 97.96% specificity, and 97.78% overall accuracy, which is better than most of the previous methods, except the method proposed by Zhou, et al. (97.37% sensitivity, 98.44% specificity, and 97.99% accuracy) [22], and the method proposed by Petrėnas, et al. (97.12% sensitivity and 98.28% specificity) [20].

Most of the detectors in Table 3 employ a window length of 127/128 beats, i.e., [15,16,18,19,21,22], however, detectors with a 128-beat window tend to miss brief clinical episodes. It is important to also consider the ability to detect brief AF episodes when evaluating detector performance. Table 3 also compared the shortest length of the detected AF episodes. The proposed method investigated the performance of the detector from 50 beats to 200 beats. With short detection window $n = 50$ beats, it achieving 89.58% sensitivity, 90.32% specificity, and 90.04% accuracy (see Table 1). The proposed detection process involves the estimation of probabilities, and a shorter window implies increased statistical uncertainty. Petrėnas, et al. [20], Lee et al. [19], Lake, et al. [23] and Lian et al. [15] reported on performance for shorter windows. Specially, Lake, et al. [23] worked on short window of 12 beats (91% sensitivity and 94% specificity). Petrėnas, et al. [20] used a even shorter window of only 8 beats.

Table 3. Comparison of detector performance on the MIT-BIH Atrial Fibrillation Database (AFDB).

Method	Year	Database	Length of Detected Episodes (Beats)		Best Performance (%)		
			Shortest	Best	SEN	SPE	ACC
Proposed method	2017	AFDB	50	150	97.04	97.96	97.78
Zhou, et al. [22]	2015	AFDB	127	127	97.37	98.44	97.99
Petrėnas, et al. [20]	2015	AFDB	8	60	97.12	98.28	-
Zhou, et al. [21]	2014	AFDB	127	127	96.89	98.25	97.67
Lee, et al. [19]	2013	AFDB *	12	128	98.22	97.67	97.91
Huang, et al. [16]	2011	AFDB	128	128	96.1	98.1	-
Lake, et al. [23]	2011	AFDB	12	12	91	94	-
Lian, et al. [15]	2011	AFDB	32	128	95.9	95.4	-
Parvaresh, et al. [17]	2011	AFDB *	15 s	15 s	96.14	93.20	-
Dash, et al. [18]	2009	AFDB **	128	128	94.4	95.1	-
Babaeizadeh, et al. [14]	2009	AFDB *	-	-	92	-	-
Couceiro, et al. [44]	2008	AFDB *	12	100	93.8	96.09	-

* Records "00735" and "03665" omitted; ** Records "04936" and "05091" omitted.

Figure 6 shows the computation time according to word length m and observation window n, with Δ changing from 0.00 to 0.19 in each computation. Compared to the observation window, the word length has more influence to the computation time. The computation time is between 6.09 and 6.42 ms with $m = 4$, and between 23.51 and 28.72 ms with $m = 12$. With the optimal AF-detection parameters ($m = 9$, $n = 150$), the computation time is 9.12 ms (programs run in MATLAB R2015a on Intel(R) Core(TM) i7-6700k CPU @ 4.00GHz processor, Lenovo, Beijing, China). This shows that our algorithm can be realizable in real time for practical applications, and it is faster than many other algorithms: 20 ~30 ms with observation seg = 128, and 3 ~4 ms with seg = 12 in Lee et al. [19], 5.2 s with seg = 128 in Lake and Moorman [23], 200 ms with seg = 128 in Dash et al. [18], and 3 s with seg = 100 in Tateno and Glass [12].
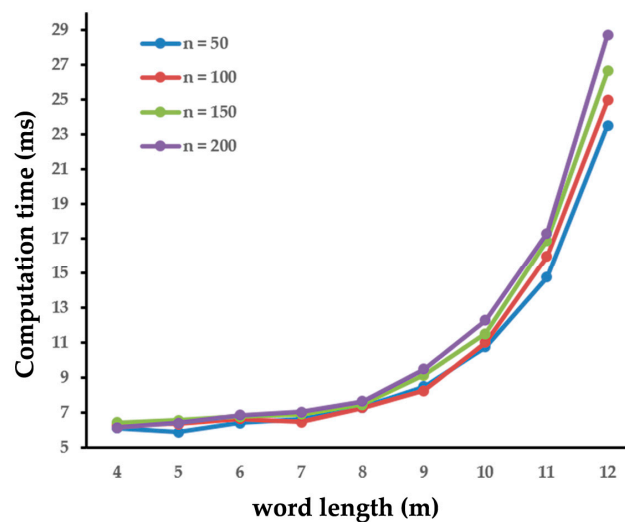

Figure 6. Computation time according to m and n (the computation times are the average values of 100 trials).

4.4. Study Limitations and Future Work

The selection of the observation window places a lower boundary on the length of AF episode that can be detected. This approach is suitable for AF episodes that are prolonged, and the detection results reveal limited performance for sporadic AF episodes of very short duration (e.g., less than 30 s). From a diagnostic point of view, using multiple and complementary methods to detect AF episode may be helpful. Our new approach complements conventional approaches of AF detection, since our

algorithm is based on a completely different concept from other approaches. This algorithm may also be easily adapted to other physiological and physical time series data, provided that a meaningful symbolic mapping rule can be defined.

We propose the implementation of a confidence-based voting scheme due to the disproportionate weighting that may be given to certain dissimilarity values. For example, if a dissimilarity computation for a given test segment in one case yields a 0.90 AF dissimilarity and a 0.95 NSR dissimilarity, the values suggest that the test segment is quite dissimilar from both templates. However, normalization of the dissimilarity values results in a normalized value of 0.47 AF dissimilarity and 0.53 NSR dissimilarity. In a second case where a test segment yields a 0.15 AF dissimilarity and a 0.10 NSR dissimilarity, the test segment is much more similar to each segment and thus its weighting may be more significant. However, the dissimilarity values would be normalized to approximately 0.60 AF and 0.40 NSR, which are not too different from that of the first case. In a future study, we aim to test whether dissimilarity values such as those in the first case should be regarded with less confidence than dissimilarity values of the second case.

In addition to confidence-based voting, to further improve and validate our results, we will also perform cross validation with different sets of training data from the AFDB and NSRDB to validate our current results.

Author Contributions: Xingran Cui, Chung-Kang Peng and Albert C. Yang conceived and designed the study; Wen-Hung Yang and Bernard C. Jiang performed the experiments; Xingran Cui, Wen-Hung Yang, Emily Chang analyzed the data; Xingran Cui wrote the paper. All authors were involved with the editing of the publication, approval of the data, reported findings, and final version of the publication.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Iwasaki, Y.-K.; Nishida, K.; Kato, T.; Nattel, S. Atrial Fibrillation Pathophysiology. *Circulation* **2011**, *124*, 2264–2274. [CrossRef] [PubMed]
2. Munger, T.M.; Wu, L.Q.; Shen, W.K. Atrial fibrillation. *J. Biomed. Res.* **2014**, *28*, 1–17. [PubMed]
3. January, C.T.; Wann, L.S.; Alpert, J.S.; Calkins, H.; Cigarroa, J.E.; Cleveland, J.C., Jr.; Conti, J.B.; Ellinor, P.T.; Ezekowitz, M.D.; Field, M.E.; et al. 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society. *J. Am. Coll. Cardiol.* **2014**, *64*. [CrossRef]
4. Israel, C.W.; Grönefeld, G.; Ehrlich, J.R.; Li, Y.G.; Hohnloser, S.H. Long-term risk of recurrent atrial fibrillation as documented by an implantable monitoring device: Implications for optimal patient care. *J. Am. Coll. Cardiol.* **2004**, *43*, 47–52. [CrossRef] [PubMed]
5. Page, R.L.; Wilkinson, W.E.; Clair, W.K.; McCarthy, E.A.; Pritchett, E.L. Asymptomatic arrhythmias in patients with symptomatic paroxysmal atrial fibrillation and paroxysmal supraventricular tachycardia. *Circulation* **1994**, *89*, 224–227. [CrossRef] [PubMed]
6. Andresen, D.; Bruggemann, T. Heart rate variability preceding onset of atrial fibrillation. *J. Cardiovasc. Electrophysiol.* **1998**, *9*, S26–S29. [PubMed]
7. Glass, L.; Tateno, K. A method for detection of atrial fibrillation using the RR intervals. *Comput. Cardiol.* **2000**, *27*, 391–394.
8. Langley, P.; Bourke, J.P.; Murray, A. Frequency analysis of atrial fibrillation. *Comput. Cardiol.* **2000**, *27*, 65–68.
9. Stridh, M.; Sornmo, L. Spatiotemporal QRST cancellation techniques for analysis of atrial fibrillation. *IEEE Trans. Biomed. Eng.* **2001**, *48*, 105–111. [CrossRef] [PubMed]
10. Fischer, R.; Klein, G.; Widiger, B.; Hoy, L.; Zywietz, C. Discrimination between atrial flutter and atrial fibrillation by computing a flutter index. *Comput. Cardiol.* **2005**, *32*, 81–84.
11. Moody, G.; Mark, R. A new method for detecting atrial fibrillation using R-R intervals. *Comput. Cardiol.* **1983**, *10*, 227–230.

12. Tateno, K.; Glass, L. Automatic detection of atrial fibrillation using the coefficient of variation and density histograms of RR and deltaRR intervals. *Med. Biol. Eng. Comput.* **2001**, *39*, 664–671. [CrossRef] [PubMed]

13. Kikillus, N.; Hammer, G.; Lentz, N.; Stockwald, F.; Bolz, A. Three different algorithms for identifying patients suffering from atrial fibrillation during atrial fibrillation free phases of the ECG. In Proceedings of the Computers in Cardiology, Durham, NC, USA, 30 September–3 October 2007; Volume 34, pp. 801–804.

14. Babaeizadeh, S.; Gregg, R.E.; Helfenbein, E.D.; Lindauer, J.M.; Zhou, S.H. Improvements in atrial fibrillation detection for real-time monitoring. *J. Electrocardiol.* **2009**, *42*, 522–526. [CrossRef] [PubMed]

15. Lian, J.; Wang, L.; Muessig, D. A simple method to detect atrial fibrillation using RR intervals. *Am. J. Cardiol.* **2011**, *107*, 1494–1497. [CrossRef] [PubMed]

16. Huang, C.; Ye, S.; Chen, H.; Li, D.; He, F.; Tu, Y. A novel method for detection of the transition between atrial fibrillation and sinus rhythm. *IEEE Trans. Biomed. Eng.* **2011**, *58*, 1113–1119. [CrossRef] [PubMed]

17. Parvaresh, S.; Ayatollahi, A. Automatic atrial fibrillation detection using autoregressive modeling. In Proceedings of the 2011 International Conference on Biomedical Engineering and Technology, Kuala Lumpur, Malaysia, 4–5 June 2011; pp. 105–108.

18. Dash, S.; Chon, K.; Lu, S.; Raeder, E. Automatic real time detection of atrial fibrillation. *Ann. Biomed. Eng.* **2009**, *37*, 1701–1709. [CrossRef] [PubMed]

19. Lee, J.; Nam, Y.; McManus, D.D.; Chon, K.H. Time-Varying Coherence Function for Atrial Fibrillation Detection. *IEEE Trans. Biomed. Eng.* **2013**, *60*, 2783–2793. [PubMed]

20. Petrenase, A.; Marozas, V.; Sönmo, L. Low-complexity detection of atrial fibrillation in continuous long term monitoring. *Comput. Biol. Med.* **2015**, *60*, 2783–2793. [CrossRef] [PubMed]

21. Zhou, X.; Ding, H.; Ung, B.; Pickwell-MacPherson, E.; Zhang, Y. Automatic Online Detection of Atrial Fibrillation Based on Symbolic Dynamics and Shannon Entropy. *Biomed. Eng. Online* **2014**, *13*, 1–18. [CrossRef] [PubMed]

22. Zhou, X.; Ding, H.; Wu, W.; Zhang, Y. A Real-Time Atrial Fibrillation Detection Algorithm Based on the Instantaneous State of Heart Rate. *PLoS ONE* **2015**, *10*, e0136544. [CrossRef] [PubMed]

23. Lake, D.; Moorman, J. Accurate estimation of entropy in very short physiological time series: The problem of atrial fibrillation detection in implanted ventricular devices. *Am. J. Physiol. Heart Circ. Physiol.* **2011**, *300*, H319–H325. [CrossRef] [PubMed]

24. Carrara, M.; Carozzi, L.; Moss, T.J.; de Pasquale, M.; Cerutti, S.; Lake, D.E.; Moorman, J.R.; Ferrario, M. Classification of cardiac rhythm using heart rate dynamical measures: Validation in MIT-BIH databases. *J. Electrocardiol.* **2015**, *48*, 943–946. [CrossRef] [PubMed]

25. Carrara, M.; Carozzi, L.; Moss, T.J.; de Pasquale, M.; Cerutti, S.; Ferrario, M.; Lake, D.E.; Moorman, J.R. Heart rate dynamics distinguish among atrial fibrillation, normal sinus rhythm and sinus rhythm with frequent ectopy. *Physiol. Meas.* **2015**, *36*, 1873–1888. [CrossRef] [PubMed]

26. Masè, M.; Disertori, M.; Marini, M.; Ravelli, F. Characterization of rate and regularity of ventricular response during atrial tachyarrhythmias. Insight on atrial and nodal determinants. *Physiol. Meas.* **2017**, *38*, 800–818. [CrossRef] [PubMed]

27. Lian, J.; Mussig, D.; Lang, V. Computer modeling of ventricular rhythm during atrial fibrillation and ventricular pacing. *IEEE Trans. Biomed. Eng.* **2006**, *53*, 1512–1520. [CrossRef] [PubMed]

28. Lerma, C.; Trine, K.M.; Guevara, M.; Glass, L. Stochastic aspects of cardiac arrhythmias. *J. Stat. Phys.* **2007**, *128*, 347–374. [CrossRef]

29. Zeng, W.; Glass, L. Statistical properties of heartbeat intervals during atrial fibrillation. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.* **1996**, *54*, 1779–1784. [CrossRef]

30. Costa, M.; Goldberger, A.L.; Peng, C.-K. Multiscale entropy analysis of complex physiologic time series. *Phys. Rev. Lett.* **2002**, *89*, 068102. [CrossRef] [PubMed]

31. Costa, M.; Goldberger, A.L.; Peng, C.-K. Multiscale entropy analysis of biological signals. *Phys. Rev. E* **2005**, *71*. [CrossRef] [PubMed]

32. Peng, C.-K.; Costa, M.; Goldberger, A.L. Adaptive data analysis of complex fluctuations in physiologic time series. *Adv. Adapt. Data Anal.* **2009**, *1*, 61–70. [CrossRef] [PubMed]

33. Yang, A.C.; Hseu, S.S.; Yien, H.W.; Goldberger, A.L.; Peng, C.K. Linguistic analysis of the human heartbeat using frequency and rank order statistics. *Phys. Rev. Lett.* **2003**, *90*, 108103. [CrossRef] [PubMed]

34. Yang, A.C.; Goldberger, A.L.; Peng, C.K. Genomic classification using an information-based similarity index: application to the SARS coronavirus. *J. Comput. Biol.* **2005**, *12*, 1103–1116. [PubMed]

35. Yang, A.C.; Peng, C.K.; Yien, H.W.; Goldberger, A.L. Information categorization approach to literary authorship disputes. *Phys. A Stat. Mech. Appl.* **2003**, *329*, 473–483. [CrossRef]

36. Peng, C.K.; Yang, A.C.C.; Goldberger, A.L. Statistical physics approach to categorize biologic signals: From heart rate dynamics to DNA sequences. *Chaos* **2007**, *17*, 015115. [CrossRef] [PubMed]

37. Christini, D.J.; Glass, L. Introduction: Mapping and control of complex cardiac arrhythmias. *Chaos* **2002**, *12*, 732–739. [CrossRef] [PubMed]

38. Hall, K.; Christini, D.J.; Tremblay, M.; Collins, J.J.; Glass, L.; Billette, J. Dynamic control of cardiac alternans. *Phys. Rev. Lett.* **1997**, *78*, 4518–4521. [CrossRef]

39. Goldberger, A.L.; Amaral, L.A.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.-K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **2000**, *101*, E215–E220. [CrossRef] [PubMed]

40. Moody, G.B.; Mark, R.G.; Goldberger, A.L. PhysioNet: A research resource for studies of complex physiologic and biomedical signals. *Comput. Cardiol.* **2000**, *27*, 179–182. [PubMed]

41. Costa, M.; Moody, G.B.; Henry, I.; Goldberger, A.L. PhysioNet: An NIH research resource for complex signals. *J. Electrocardiol.* **2003**, *36*, 139–144. [CrossRef] [PubMed]

42. Wu, G.; Chang, E.Y. KBA: Kernel Boundary Alignment Considering Imbalanced Data Distribution. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, P786–P795. [CrossRef]

43. Larburu, N.; Lopetegi, T.; Romero, I. Comparative study of algorithms for Atrial Fibrillation detection. In Proceedings of the Conference on Computing in Cardiology, Hangzhou China, 18–21 September 2011; Volume 38, pp. 265–268.

44. Couceiro, R.; Carvalho, P.; Henriques, J.; Antunes, M.; Harris, M.; Habetha, J. Detection of atrial fibrillation using model-based ECG analysis. In Proceedings of the 19th International Conference on Pattern Recognition, Tampa, FL, USA, 8–11 December 2008; pp. 1–5.