Journal of Multimedia and Information System VOL. 3, NO. 2, June 2016(pp. 1-8): ISSN 2383-7632(Online)
http://dx.doi.org/10.9717/JMIS.2016.3.2.1

R-to-R Extraction and Preprocessing Procedure for an Automated Diagnosis of Various Diseases from ECG Data

Vincentius Timothy[1, 2, a], Ary Setijadi Prihatmanto[1, b], Kyung-Hyune Rhee[*]

Abstract

In this paper, we propose a method to automatically diagnose various diseases. The input data consists of electrocardiograph (ECG) recordings. We extract R-to-R interval (RRI) signals from ECG recordings, which are preprocessed to remove trends and ectopic beats, and to keep the signal stationary. After that, we perform some prospective analysis to extract time-domain parameters, frequency-domain parameters, and nonlinear parameters of the signal. Those parameters are unique for each disease and can be used as the statistical symptoms for each disease. Then, we perform feature selection to improve the performance of the diagnosis classifier. We utilize the selected features to diagnose various diseases using machine learning. We subsequently measure the performance of the machine learning classifier to make sure that it will not misdiagnose the diseases. The first two steps, which are R-to-R extraction and preprocessing, have been successfully implemented with satisfactory results.

Key Words: Automated diagnosis, ECG recordings, Heart rate variability, Machine learning.

I. INTRODUCTION

Heart rate variability (HRV) refers to the variation in time between consecutive heart beats. This variation is under the control of autonomic nervous system (ANS), which is divided into sympathetic and parasympathetic branches. There have been numerous studies describing the relations of ANS to HRV. Generally speaking, the sympathetic branch tends to increase heart rate (HR) and decrease HRV, whereas the parasympathetic branch tends to decrease HR and increase HRV [1]. A cardiacally-healthy person tends to have both well-functioning sympathetic and parasympathetic branches. As a result, they tend to have higher variability of HRV than the less cardiacally-healthy person.

Main areas of applications of HRV analysis include the risk stratification of sudden cardiac death after acute myocardial infarction. HRV analysis is also generally accepted to provide an early warning sign of diabetic neuropathy. Besides these main applications, HRV has been studied with relation to several cardiovascular diseases, renal failures, physical exercise, occupational and psychosocial stress, gender, age, drugs, alcohol, smoking, and sleep [2].

A number of studies report that each cardiac-related disease has unique HRV in terms of time-domain parameters, frequency-domain parameters, and nonlinear parameters. These parameters have become the de facto numerical/statistical symptoms for each disease. Due to the repetitive and patterned nature of HRV analysis for diagnosis of various diseases, their automation has become increasingly prevalent in our lives. Linear Discriminant Analysis (LDA) has been developed to detect acute mental stress due to university examination with a total classification accuracy, a sensitivity, and a specificity rate of 90%, 86%, and 95%, respectively [3]. Support Vector Machine (SVM) – Radial Basis Function (RBF) and Neural Network (NN) have been developed to detect arrhythmia with an equal average accuracy, sensitivity and specificity of 98.9% [4].

Manuscript received June 25,2016; Revised July 5, 2016; Accepted July 28, 2016. (ID No. JMIS-2016-0006)
Corresponding Author (*): Kyung-Hyune Rhee, Department of IT Convergence and Application Engineering, Pukyong National University, South Korea, +82 51-629-6245, khrhee@pknu.ac.kr
[1]Department of Electrical Engineering, Institut Teknologi Bandung, Bandung, Indonesia
[2]Department of IT Convergence and Application Engineering, Pukyong National University, Busan, South Korea
E-mail: [a] vincentius_timothy@hotmail.com, [b] asetijadi@lskk.ee.itb.ac.id

Table 1. Summary of proposed HRV analysis parameters.

Parameter	Unit	Description	References
Time-domain analyses: statistical methods			
Mean RR	ms	The mean of RR intervals	[5]
SDNN	ms	Standard deviation of RR intervals	[5]
RMSSD	ms	Square root of the mean squared differences between successive RR intervals	[5]
NN50	count	Number of successive RR interval pairs that differ more than 50 ms	[5]
pNN50	%	NN50 divided by the total number of RR intervals	[5]
Time-domain analyses: geometrical methods			
HRV triangular index	-	The integral of the RR interval histogram divided by the height of the histogram	[5]
TINN	ms	Baseline width of the RR interval histogram	[5]
Frequency-domain analyses			
VLF, LF, HF power	ms^2	Absolute powers of very low frequency band (0-0.04 Hz), low frequency band (0.04-0.15 Hz), and high frequency band (0.15-0.4 Hz), respectively	[5]
LF/HF ratio	-	Ratio between LF and HF band power	[5]
Nonlinear analyses: Poincaré plot			
SD1, SD2	ms	Short-term and long-term variability standard deviation, respectively	[9,10]
Nonlinear analyses: correlation dimension			
ApEn(0.2), ApEn(r_{max}), ApEn(r_{chon})	-	Approximate entropy where the tolerance value r is chosen to be r=0.2×SDNN, in the interval [0.1×SDNN, 0.9×SDNN] which maximizes ApEn, and computed according to the following formula proposed by Chon[12], respectively	[11,12]
SampEn	-	Sample entropy	[11]
Nonlinear analyses: correlation dimension			
D_2	-	Correlation dimension	[13]
Nonlinear analyses: detrended fluctuation analyses			
α_1, α_2	-	Short-term and long-term fluctuations, respectively	[14.15]
Nonlinear analyses: recurrence plot			
l_{mean}	beats	Mean line length of diagonal lines	[16-18]
l_{max}	beats	Maximum line length of diagonal lines	[16-18]
REC	%	Recurrence rate (percentage of recurrence points)	[16-18]
DET	%	Determinism (percentage of recurrence points which form diagonal lines)	[16-18]
ShEn	-	Shannon entropy of diagonal line lengths' probability distribution	[16-18]

This paper proposes a method to automatically diagnose various diseases. The input data is raw electrocardiogram (ECG) recordings, and the output is multilabel classification of various diseases. The rest of the paper is organized as follows: in Section 2 we briefly review the time-domain parameters, frequency-domain parameters, and nonlinear parameters of HRV analysis. The proposed method is described in Section 3, while Section 4 contains the implementation and discussion of the proposed method. Section 5 contains the conclusion, and finally, future work is presented in Section 6.

II. BACKGROUND

This section describes briefly the prospective analyses that are done to HRV extracted from ECG recordings. The prospective analyses include time-domain parameter analysis, frequency-domain parameter analysis, and nonlinear parameter analysis. The computations as well as the notations used are mainly based on the guidelines given in [5]. A summary of the analysis parameters is given in Table 1.

Time-domain analyses are computationally simple and they are applied directly to the series of successive RR interval values. Also, they do not require stationarity in the same manner as most frequency domain and nonlinear analyses do. There are two methods in time-domain analysis: statistical methods and geometric methods. Statistical methods are based on various moments of the RR intervals and the delta RR intervals. Geometric methods convert the RR interval data into a geometric pattern. The geometric techniques generally have better performance on poorly edited data [6]. The main limitation of time domain analyses is their lack of discrimination between effects of the sympathetic and parasympathetic autonomic branches [7].

The main idea behind the frequency-domain analyses of HRV is the observation that HRV is composed of certain well-defined rhythms, which are related to different regulatory mechanisms of cardiovascular control [7]. Frequency-domain analyses consist of first calculating the power spectral density (PSD) of the RR intervals. Secondly, the PSD is broken into separate frequency bands: very low frequency (0-0.04 Hz), low frequency (0.04-0.15 Hz), and high frequency (0.15-0.4 Hz). Thirdly, the power in each band is calculated by integrating the PSD within the band limits [6]. It is believed that the power in the low frequency band associates with the combination of sympathetic and parasympathetic autonomic branches, while the power in the high frequency band only associates with parasympathetic autonomic branch. Because the high frequency component of HRV is centered around respiratory frequency, respiration should always be considered in HRV analysis [2]. The respiratory frequency can be estimated from the ECG signals – more precisely, from the R-wave amplitudes [8].

Nonlinear analyses of HRV are employed because linear analyses such as time- and frequency-domain analyses are insufficient to describe the complexity of the heart. Therefore, various nonlinear analyses are applied to HRV to fully capture the characteristics of beat-to-beat variability. Nonlinear properties of HRV were analyzed by the following methods: Poincaré plot [9,10], approximate and sample entropy [11,12], correlation dimension [13], detrended fluctuation analysis [14,15], and recurrence plot [16-18]. It is important to note that nonlinear analyses tend to reveal more information about HRV characteristics than time- or frequency-domain analyses.

III. PROPOSED METHOD

The overall procedures for automated diagnosis of various diseases are shown in Figure 1. The input data is raw ECG recordings, from which the R-to-R intervals (RRI) of HRV are extracted. The requirements of the RRI signal to be processed are freedom from ectopic beats (abnormal beats that are due to unusual impulses), stationarity (absent of low frequency trends), and evenly-sampled RRI [19]. To achieve these three requirements, the RRI signal is initially preprocessed.

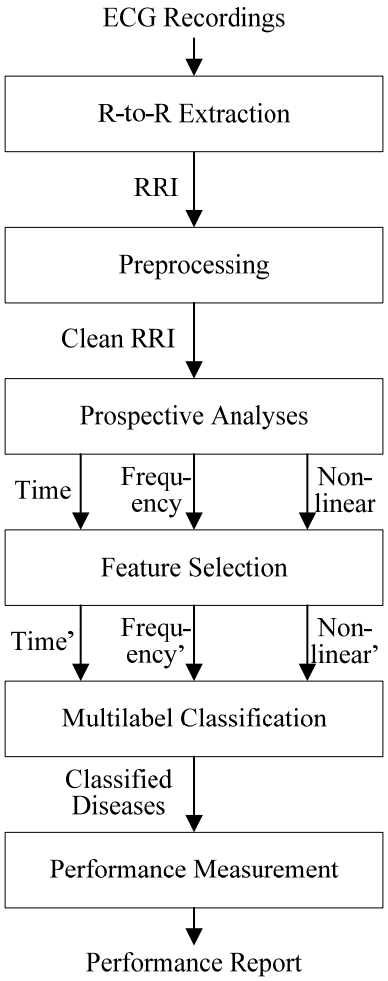


Fig. 1. Proposed method for automated diagnosis of various diseases.

After that, prospective analyses are performed on the preprocessed RRI signal to extract time-domain parameters, frequency-domain parameters, and nonlinear parameters of RRI signal, as explained in Section 2. Those parameters are unique for each disease and can be used as the statistical symptoms for each disease. It is possible to utilize all HRV parameters reported in Table 1 for performing diagnosis on various diseases, however this may decrease the performance of the classifier, particularly because of the curse of dimensionality [20]. Therefore, feature selection is performed to obtain only the essentials of time-domain, frequency-domain, and nonlinear parameters.

Table 2. Summary of the details of the proposed methods.

No	Procedures	Functionality	Inputs	Outputs
1	R-to-R extraction	Converts raw ECG recordings to R-to-R intervals (RRI)	Raw ECG recordings	Unpreprocessed RRI signal
2	Preprocessing	Preprocesses RRI signal to meet the following requirements: no ectopic beats, stationary, and evenly-sampled	Unpreprocessed RRI signal	Preprocessed RRI signal
3	Prospective analyses	Performs prospective analyses on preprocessed RRI signal	Preprocessed RRI signal	Time-domain, frequency-domain, and nonlinear parameters
4	Feature selection	Selects only the essentials of time-domain, frequency-domain, and nonlinear parameters	Time-domain, frequency-domain, and nonlinear parameters	Selected time-domain, frequency-domain, and nonlinear parameters
5	Multilabel classification	Performs diagnosis on various diseases	Selected time-domain, frequency-domain, and nonlinear parameters	Multilabel classification of various diseases
6	Performance measurement	Evaluates the performance of multilabel classifier	Multilabel classification of various diseases	Accuracy, sensitivity, and specificity

The selected features (i.e. selected parameters) are then utilized to diagnose various diseases using machine learning techniques. Machine learning techniques suitable for this task are multilabel classification techniques, such as Artificial Neural Networks (ANN), random forest, Support Vector Machine (SVM), and Linear Discriminant Analysis (LDA). In order to evaluate the classifier, common measures are computed for binary classification performance measurement [21] for each class. The classification performance is described in terms of statistical accuracy, sensitivity, and specificity.

The details of the proposed method are summarized in Table 2.

IV. IMPLEMENTATION AND DISCUSSION

The proposed method above is implemented inside MATLAB® R2016a environment (The MathWorks, Inc.). The proposed method is still a work in progress; therefore, not all of the procedures have been implemented. The procedures that have been implemented are R-to-R extraction and preprocessing.

4.1. Input data

The input data used to test the system is the standard MIT-BIH arrhythmia database. A record consists of three files; the header file, the annotation file, and the data file. The header file is a short text file that describes the signals, including the name or URL of the signal file, storage format, number and type of signals, sampling frequency, calibration data, digitizer characteristics, record duration, and starting time. The annotation file contains sets of labels, each of which describes a feature of one or more signals at a specified time in the record. The data file contains digitized samples of one or more signals.

4.2. R-to-R extraction

By nature, the raw ECG data is non-stationary; that is, there are low-frequency trends in raw ECG data. Therefore, the first step is to remove the low-frequency component [22]. This process is done by applying Fast Fourier Transform (FFT), removing low-frequencies, and restoring the ECG signal by applying Inverse FFT (IFFT).

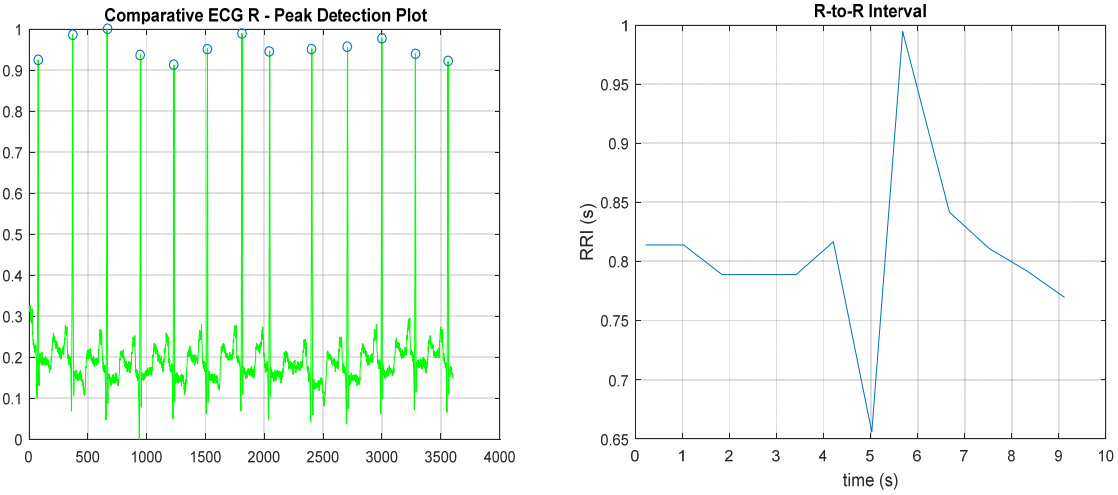
The second step is to find the local maxima. To do that, apply windowed filter which detects the maximum in its window only and ignores all other values. Only the significant values of the windowed signal should be preserved. To do this, use threshold filter. In order to refine the result and ensure that all the peaks are detected, adjust the filter window size and repeat filtering.

Using MITDB 100 data, the R-peaks detection and extraction works as expected. The result for R-peaks detection and R-to-R extraction for the first 10 seconds is shown in Figure 2.

4.3. Preprocessing

In preprocessing, the signal is required to have no ectopic beats, to be stationary, and to be evenly-sampled. Therefore, there are three subprocedures that needs to be done in this step. In order to have the signal with no ectopic beats, two methods are used: ectopic beats detection based on thresholding and ectopic beats correction based on linear interpolation. In order to have stationary signal, the method of detrending based on smoothness prior approach is used.

Journal of Multimedia and Information System VOL. 3, NO. 2, June 2016(pp. 1-8): ISSN 2383-7632(Online)
http://dx.doi.org/10.9717/JMIS.2016.3.2.1



(a) (b)

Fig. 2. The R-to-R extraction procedure: (a) detected R-peaks, and (b) extracted RRI signals.

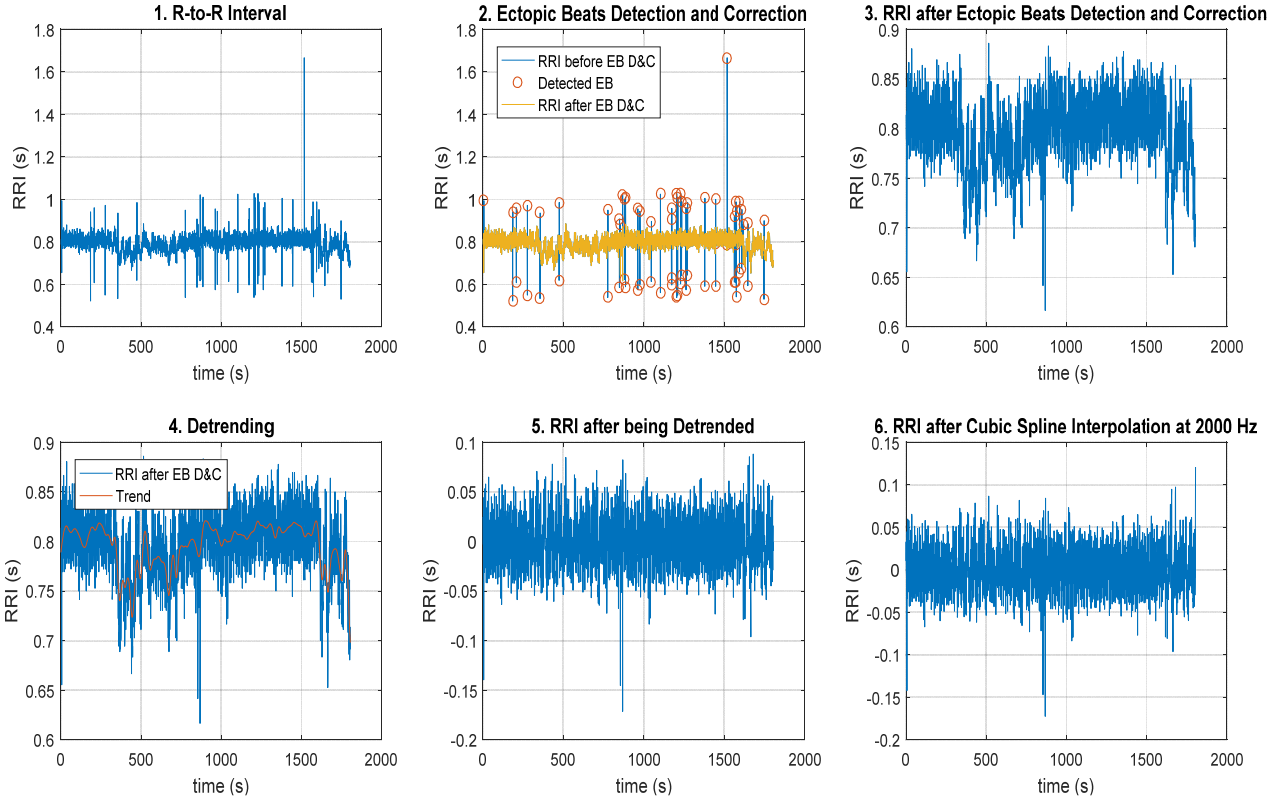


Fig. 3. The preprocessing procedure, step-by-step.

In order to have an evenly-sampled signal, the method of resampling based on cubic sampling interpolation is used.

Using MITDB 100 data, the preprocessing method works as expected. The step-by-step procedure of preprocessing is shown in Figure 3.

V. CONCLUSION

Due to the repetitive and patterned nature of HRV analysis for the diagnosis of various diseases, their automation has become increasingly prevalent in our lives. In this paper, a new method was proposed for automated diagnosis of various diseases, as explained in Section 3. The first two steps, which are R-to-R extraction and preprocessing, have been successfully implemented with satisfactory results.

VI. FUTURE WORK

There is still a need to explore in detail various alternatives for statistical analysis methods, feature selection methods, multilabel classification methods, and performance measurement methods. After that, these methods would need to be integrated and implemented to form one large system capable of automated diagnosis of various diseases.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIP)(No.NRF-2014R1A2A1A11052981).

REFERENCES

[1] G. Berntson, J.B. Jr., D. Eckberg, P. Grossman, P. Kaufmann, M. Malik, H. Nagaraja, S. Porges, J. Saul, P. Stone, and M.V.D. Molen, "Heart rate variability: origins, methods, and interpretive caveats," *Psychophysiology* 34, pp. 623–648, 1997.

[2] M.P. Tarvainen, J.-P. Niskanen, J.A. Lipponen, P.O. Ranta-Aho, and P.A. Karjalainen, "Kubios HRV-heart rate variability analysis software," *Comput. Methods Prog. Biomed.,* Vol. 113, No. 1, pp.210–220, 2014.

[3] P. Melillo, M. Bracale, and L. Pecchia, "Nonlinear Heart Rate Variability features for real-life stress detection. Case study: students under stress due to university examination," *Biomed. Eng. Online* 10, pp.1–13, 2011.

[4] F.A. Elhaj, N. Salim, A.R. Harris, T.T. Swee, and T. Ahmeda, "Arrhythmia recognition and classification using combined linear and nonlinear features of ECG signals," *Comput. Methods Prog. Biomed.,* Vol. 127. No. 1, pp.52-63, 2016.

[5] Task force of the European society of cardiology and the North American society of pacing and electrophysiology, "Heart rate variability – standards of measurement, physiological interpretation, and clinical use," *Circulation* Vol. 93, No. 5, pp.1043–1065, 1996.

[6] A.H. Khandoker, C. Karmakar, M. Brennan, A. Voss, and M. Palaniswami, "Poincaré Plot Methods for Heart Rate Variability Analysis," *Springer*, 2013.

[7] T. Kuusela, "Methodological Aspects of Heart Rate Variability Analysis". In: M.V. Kamath (Ed), M.A. Watanabe (Ed), A.R.M. Upton (Ed), "Heart Rate Variability (HRV) Signal Analysis: Clinical Applications," *CRC Press*, pp. 9-42, 2013.

[8] G. Moody, R. Mark, A. Zoccola, and S. Mantero, "Derivation of respiratory signals from multi-lead ECGs," *Comput. Cardiol.* 12, pp. 113–116, 1985.

[9] P. Melillo, R. Fusco, M. Sansone, M. Bracale, and L. Pecchia, "Discrimination power of long-term heart rate variability measures for chronic heart failure detection," *Medical and Biological Engineering and Computing*, Vol. 49, No. 1, pp. 67-74, 2011.

[10] M. Brennan, M. Palaniswami, and P. Kamen, "Do existing measures of Poincare plot geometry reflect nonlinear features of heart rate variability?," *IEEE Transactions on Biomedical Engineering*, Vol. 48, No 11, pp. 1342-1347, 2001.

[11] J.S. Richman and J.R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal of Physiology-Heart and Circulatory Physiology*, Vol. 278, No 6, H2039-H2049, 2000.

[12] K.H. Chon, C.G. Scully and S. Lu, "Approximate Entropy for all Signals Is the Recommended Threshold Value r Appropriate?," *IEEE Engineering in Medicine and Biology Magazine*, Vol. 28, pp. 18-23, 2009.

[13] R. Carvajal, N. Wessel, M. Vallverdú, P. Caminal, and A. Voss, "Correlation dimension analysis of heart rate variability in patients with dilated cardiomyopathy," *Computer Methods and Programs in Biomedicine*, Vol. 78, pp. 133-140, 2005.

[14] T. Penzel, J.W. Kantelhardt, L. Grote, J.H. Peter, and A. Bunde, "Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea," *IEEE Transactions on Biomedical Engineering*, Vol. 50, pp. 1143-1151, 2003.

[15] C.K. Peng, S. Havlin, H.E. Stanley, and A.L. Goldberger, "Quantification of Scaling Exponents and Crossover Phenomena in Nonstationary Heartbeat Time-Series," *Chaos*, Vol. 5, pp. 82-87, 1995.

[16] L.L. Trulla, A. Giuliani, J.P. Zbilut, and C.L. Webber, "Recurrence quantification analysis of the logistic equation with transients," *Physics Letters A*, Vol. 223, pp. 255-260, 1996.

[17] C.L. Webber and J.P. Zbilut, "Dynamical Assessment of Physiological Systems and States Using Recurrence Plot Strategies," *Journal of Applied Physiology*, Vol. 76, pp. 965-973, 1994.

[18] J.P. Zbilut, N. Thomasson, and C.L. Webber, "Recurrence quantification analysis as a tool for nonlinear exploration of nonstationary cardiac signals," *Medical Engineering & Physics*, Vol. 24, pp. 53-60, 2002.

[19] J.T. Ramshur, "Design, Evaluation, and Application of Heart Rate Variability Analysis Software (HRVAS)," *University of Memphis*, 2010.

[20] A.K. Jain, R.P.W. Duin, and M. Jianchang, "Statistical pattern recognition: a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, pp. 4-37, 2000.

[21] M. Sokolova, and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf Process Manage*, Vol. 45, pp. 427-437, 2009.

[22] Michael Mukovoz, "ECG processing — R-peaks detection," Jun. 2016; http://librow.com/cases/case-2

Authors

Vincentius Timothy has received his B.Sc. degree in electrical engineering from Institut Teknologi Bandung, Indonesia, in 2015. He is currently pursuing master dual-degree both in electrical engineering from Institut Teknologi Bandung, Indonesia, and in IT convergence and application engineering from Pukyong National University, South Korea. His main research interests include intelligent healthcare systems, artificial intelligence, machine learning, pattern recognition, signal processing, image processing, VLSI circuits, and mixed-signal electronics.

Ary Setijadi Prihatmanto was born in Bandung, Indonesia, in August 1972. He received his Bachelor and Master degree in electrical engineering from ITB, and doctorate degree in informatics from Johannes Kepler University of Linz. His research interests include dualism of computer vision and computer graphics, human-computer interface, brain-computer interface, game theory on intelligent system and its applications

Kyung-Hyune Rhee received his M.S. and Ph.D. degrees from Korea Advanced Institute of Science and Technology (KAIST), Daejon, Korea in 1985 and 1992, respectively. He worked as a senior researcher in Electronic and Telecommunications Research Institute (ETRI), Daejon, Korea from 1985 to 1993. He also worked as a visiting scholar in the University of Adelaide in Australia, the University of Tokyo in Japan, the University of California at Irvine in USA, and Kyushu University in Japan. He has served as a Chairman of Division of Information and Communication Technology, Colombo Plan Staff College for Technician Education in Manila, the Philippines. He is currently a professor in the Department of IT Convergence and Application Engineering, Pukyong National University, Busan, Korea. His research interests center on multimedia security and analysis, key management protocols and mobile ad-hoc and VANET communication security.