Guest Editorial: IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2014) Special Issue

Bioinformatics and biomedicine research is fundamental to our understanding of complex biological systems, impacting the science and technology of fields ranging from agricultural and environmental sciences to pharmaceutical and medical sciences. This type of research requires close collaboration among multidisciplinary teams of researchers in computer science, statistics, physics, engineering, life sciences and medical sciences, and their interfaces. The IEEE International Conference on Bioinformatics and Biomedicine (BIBM) aims to provide an open and interactive forum to catalyze the cross-fertilization of ideas from these disciplines and to bridge our knowledge gaps.

The IEEE BIBM 2014 was held in Belfast, U.K., from November 18–22, 2014. The scientific program highlights five themes to provide breadth, depth, and synergy for research collaboration: 1) genomics and molecular structure, function, and evolution; 2) computational systems biology; 3) medical informatics and translational bioinformatics; 4) cross-cutting computational methods and bioinformatics infrastructures; and 5) healthcare informatics. IEEE BIBM 2014 received 291 research paper submissions from 1049 authors and coauthors at 39 countries. The 288 Program Committee members from 31 countries accepted 111 papers this year, of which 56 (19.2%) are regular research papers and 55 (18.9%) are short papers. Based on the PC review recommendation, 9 papers are selected from the IEEE BIBM 2014 conference for this special issue, each paper has been extended significantly based on the conference papers.

The first paper by Yuan Ling *et al.* presents a framework and system for matching medical findings (i.e., symptoms, signs, test results, etc., but collectively referred to as symptoms) with suitable medications, by combining NLP and statistical text extraction with an ILP-based matching algorithm. A core structure is a matrix of weights between symptom-medication pairs, where weights are determined by to types of co-occurrences of terms.

The paper by Rahman *et al.* addresses the problem of using machine learning classifiers to distinguishing between hypertrophic cardiomyopathy (HCM) and implantable cardioverter defibrillator (ICD) patients based on their heartbeats from electrocardiograms (ECG). A study of performance as the number of features is decreased based on the information gain criterion is also performed.

Xu *et al.* present a novel automatic classification of the HEp-2 cell images from IIF with fractal dimension features, together with morphological descriptor and pixel difference descriptor. The method was applied to the MIVIA dataset, and SVM was also used. Results showed that the fractal descriptor combining with morphological descriptor and pixel difference descriptor performed better.

Drug repositioning and related areas are gaining popularity among the machine learning community, the paper by Fang and He present a novel algorithm in prioritizing disease-causing genes based on network diffusion and rank concordance. The goal of their method is to dealing with isolated genes, in addition to using well-curated PPI network database, it is also important to apply the proposed method on predicated databases such as Michigan Molecular Interactions (MiMI), Human Protein-Protein Interaction Prediction Database (PIPs), Online Predicted Human Interaction Database (OPHID), Known and Predicted Protein-Protein Interactions (STRING).

It is of great biological significance to develop reliable computational methods for the identification of PPIs. Zhu *et al.* develop a new approach Leave-One-Out Logistic Metric Embedding (LOO-LME) for assessing the reliability of interactions in a PPI network. The method appears to compare well to previously published methods.

Yu *et al.* describes an efficient algorithm for motif finding in large DNA datasets. This is used to locate transcription factor binding sites in next-gen datasets. Their algorithm (MCES) uses MapReduce to mine emerging substrings distributedly. Their algorithm runs faster than a number of existing algorithms.

The 7th paper presents a sampling method of protein, and demonstrated to illustrate a free energy landscape of a small peptide (capped alanine) with transition states. The results look fine, but I am not sure this method is the best in the world, because there are many sampling method of protein conformation and energy. As the authors described, applications for larger system should be much interesting. Methodology section is not so easy to read, and should be improved.

Su *et al.*'s paper focus on a classification method for predicting a subject's emotional state, related to mental disorders. The method is well motivated and clearly described.

Bai *et al.* proposes an approach to detect exon skipping events from RNA-seq using a random forest classifier, ESclassifier, which aims to accurately detect the exon skipping (ES) events with RNA-seq data. The advantageous aspects of the method is that it can incorporate up to 10 different features (20 for two conditions) to decide an ES event. The experimental results on

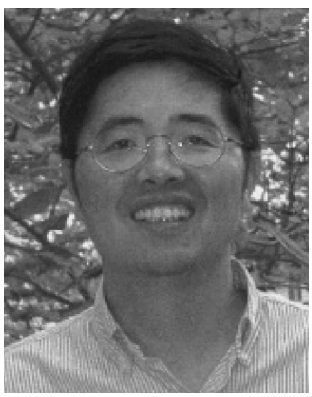
Digital Object Identifier 10.1109/TNB.2015.2446091

1536-1241 © 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

one published RNA-seq dataset verify the performance of the method.

XIAOHUA HU, *Guest Editor*
Drexel University
Philadelphia, PA 19104 USA

YADONG WANG, *Guest Editor*
School of Computer Science and Technology
Harbin Institute of Technology
Harbin, 150001 China



Xiaohua Hu is a full Professor and the Founding Director of the data mining and bioinformatics lab at the College of Computing and Informatics, Drexel University, Philadelphia, PA, USA. He is also the Founding Co-Director of the NSF Center (I/U CRC) on Visual and Decision Informatics (NSF CVDI), IEEE Computer Society Bioinformatics and Biomedicine Steering Committee Chair, and IEEE Computer Society Big Data Steering Committee Chair. He is a scientist, teacher, and entrepreneur. He joined Drexel University in 2002. He founded the *International Journal of Data Mining and Bioinformatics* (SCI indexed) in 2006, and the *International Journal of Granular Computing, Rough Sets and Intelligent Systems* in 2008. Earlier, he worked as a research scientist in world-leading R&D centers such as Nortel Research Center, and Verizon Lab (the former GTE labs). In 2001, he founded the DMW Software in Silicon Valley, CA, USA. He has a lot of experience and expertise to convert original ideas into research prototypes, and eventually into commercial products; many of his research ideas have been integrated into commercial products and applications in data mining fraud detection, and database marketing.

His current research interests are in data/text/web mining, big data, bioinformatics, information retrieval and information extraction, social network analysis, healthcare informatics, and rough set theory and application. He has published more than 270 peer-reviewed research papers in various journals, conferences and books such as various IEEE/ACM Transactions, and coedited 20 books/proceedings. He has received a few prestigious awards including the 2005 National Science Foundation (NSF) Career award, the best paper award at the 2007 International Conference on Artificial Intelligence, the best paper award at the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology. He is the Founding Editor-in-Chief of the *International Journal of Data Mining and Bioinformatics* (SCI indexed) and the *International Journal of Granular Computing, Rough Sets and Intelligent Systems*, and an Associate Editor/Editorial Board Member of four international journals (KAIS, IJDWM, IJSOI, and JCIB). His research projects are funded by the National Science Foundation (NSF), U.S. Department of Education, the Pennsylvania Department of Health, and the Natural Science Foundation of China (NSFC). He has obtained more than US$8.0 million research grants in the past 8 years as PI or Co-PI (PIs of 7 NSF grants, PI of 1 IMLS grant in the last 8 years). He has graduated 15 Ph.D. students from 2006 to 2015 and is currently supervising 12 Ph.D. students.

Yadong Wang received his B.S. degree in computer science from Heilongjiang University, Harbin, China, in 1986 and his M.Sc. degree in computer science and technology from Harbin Institute of Technology, Harbin, China, in 1989. Then he began a faculty position at Harbin Institute of Technology. He is now a full Professor and the Dean of the School of Computer Science and Technology. He is also the Director of the Center for Biomedical Information Technology and Systems Engineering Research at Heilongjiang province, the director of Heilongjiang Provinces key lab for Bioinformatics, a Member of the Chinese Association for Artificial Intelligence, and a Member of the expert committee for Biology and Medicine field, National High-Tech Research and Development (National 863 program). He has served on National 863 program and National Natural Science Foundation of China grant review panels. His research interests include bioinformatics, machine learning, and knowledge engineering. In the bioinformatics field, his research focuses on analysis of high-throughput biomedical data (microarray, next generation sequencing data, clinical data), personal genomics, and translational bioinformatics. He has published more than 150 technical papers in refereed journals and conference proceedings.

Prof. Wang is serving as the Editorial Board Member or the Guest Editor of a number of refereed journals and as the Program Committee Chair or Member of several international conferences. He has also reviewed papers for many refereed journals.

Utilizing ECG-Based Heartbeat Classification for Hypertrophic Cardiomyopathy Identification

Quazi Abidur Rahman*, Larisa G. Tereshchenko, Matthew Kongkatong, Theodore Abraham, M. Roselle Abraham, and Hagit Shatkay

***Abstract*—Hypertrophic cardiomyopathy (HCM) is a cardiovascular disease where the heart muscle is partially thickened and blood flow is (potentially fatally) obstructed. A test based on electrocardiograms (ECG) that record the heart electrical activity can help in early detection of HCM patients. This paper presents a cardiovascular-patient classifier we developed to identify HCM patients using standard 10-second, 12-lead ECG signals. Patients are classified as having HCM if the majority of their recorded heartbeats are recognized as characteristic of HCM. Thus, the classifier's underlying task is to recognize individual heartbeats segmented from 12-lead ECG signals as HCM beats, where heartbeats from non-HCM cardiovascular patients are used as controls. We extracted 504 morphological and temporal features—both commonly used and newly-developed ones—from ECG signals for heartbeat classification. To assess classification performance, we trained and tested a random forest classifier and a support vector machine classifier using 5-fold cross validation. We also compared the performance of these two classifiers to that obtained by a logistic regression classifier, and the first two methods performed better than logistic regression. The patient-classification precision of random forests and of support vector machine classifiers is close to 0.85. Recall (sensitivity) and specificity are approximately 0.90. We also conducted feature selection experiments by gradually removing the least informative features; the results show that a relatively small subset of 264 highly informative features can achieve performance measures comparable to those achieved by using the complete set of features.**

***Index Terms*—Electrocardiogram, feature selection, hypertrophic cardiomyopathy, machine learning, patient classification.**

I. Introduction

HYPERTROPHIC cardiomyopathy (HCM) is a genetic cardiovascular disease that may cause sudden cardiac death in young people [1]. The most consistent characteristic of HCM is the thickening (hypertrophy) of the muscle (myocardium) at the lower left chamber of the heart (left ventricle). Fig. 1 provides the illustration of two hearts, where the left one is normal while the one on the right shows the thickened muscle typical of hypertrophic cardiomyopathy. An imaging method, two-dimensional echocardiography, is often used to identify left ventricular hypertrophy (LVH). However, this method cannot reliably identify HCM patients when the thickening of the left ventricular muscle is not clearly detectable. Moreover, early prediction of the disease in patients not yet showing muscle thickening is not possible through echocardiography [2]. Therefore, the analysis of electrocardiogram (ECG) signals in patients with a family history of HCM and no clear muscle thickening has high diagnostic value for early detection and prediction. In a recent study we have also shown that the standard procedure of conducting ECG tests should be considered in mass pre-participation screening of young athletes [3].

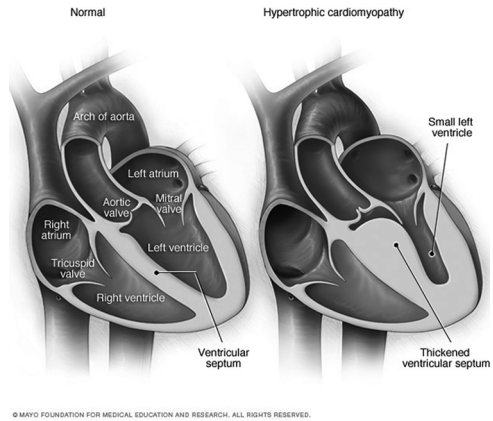


Fig. 1. Illustrations of a normal heart (left) and a heart with hypertrophic cardiomyopathy (HCM). The heart walls (muscle) are much thicker (hypertrophied) in the HCM heart [35].

Classifiers that automatically identify cardiovascular disease in patients may help reduce both cost and time of the pre-screening process. Historically, the main focus of ECG-classification research has been on identifying *arrhythmia* in cardiovascular patients. Arrhythmia is a condition where the heart beats too quickly, too slowly or in an irregular pattern. Early research has been concerned with using heartbeat classification to detect life threatening types of arrhythmia such as ventricular tachycardia (fast heart rhythm that originates in one of the ventricles of the heart) and ventricular fibrillation (uncontrolled quivering of the ventricular muscle) [4]–[6]. More recent research has expanded this idea to categorizing heartbeats along all categories of arrhythmia [7]–[9]. Traditional

Manuscript received March 27, 2015; accepted March 31, 2015. Date of publication April 24, 2015; date of current version August 05, 2015. This work was partially supported by HS's NSERC Discovery Award #298292-2009, NSERC DAS #380478-2009, CFI New Opportunities Award 10437, NIH Grant #U54GM104941, and Ontario's Early Researcher Award #ER07-04-085, and by TA's grant HL 098046 from the National Institutes of Health, and a grant from John Taylor Babbit Foundation. *Asterisk indicates corresponding author.*

*Q. A. Rahman is with the Computational Biology and Machine Learning Lab, School of Computing, Queen's University, Kingston, ON K7L 2N8, Canada.

L. G. Tereshchenko is with the Knight Cardiovascular Institute, Oregon Health and Science University, Portland, OR, 97239 USA.

M. Kongkatong, T. Abraham. and M. R. Abraham are with the Heart and Vascular Institute, Johns Hopkins University, Baltimore, MD 21218 USA.

H. Shatkay is with the Department of Computer and Information Sciences & Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE 19716 USA.

H. Shatkay is with the Computational Biology and Machine Learning Lab, School of Computing, Queen's University, Kingston, ON K7L 2N8, Canada.

Digital Object Identifier 10.1109/TNB.2015.2426213

1536-1241 © 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

machine learning methods such as artificial neural networks [9], support vector machines [8], random forests [10], and linear discriminants [11] have been used to detect arrhythmia. Random forests and support vector machines have been shown to perform well with accuracy greater than 95%.

As mentioned earlier, left ventricular hypertrophy is the most common indicator of the presence of HCM in cardiovascular patients. Several criteria, derived from, amplitude values of ECG waveforms have been proposed to detect cardiovascular patients with left ventricular hypertrophy (LVH) based on ECG signals. Many studies have been conducted to validate these LVH-detection criteria, which have generally achieved high specificity (approximately 100%) [12]–[14]. However sensitivity has been reported to be low (approximately 50%) across different studies [15]. Multiple linear regression and rule-based methods have also been used to detect cardiovascular patients with LVH [16], [17]. Corrado and McKenna have proposed a set of amplitude-thresholds for specifically detecting HCM patients [18]. Potter *et al*. have tested these thresholds on a small group of 56 HCM patients and 56 healthy control subjects [19]. The reported sensitivity and specificity from this study was approximately 90%. However, we are not aware of any previous work that employs machine learning methods for identifying HCM patients from ECG signals. Moreover, the number of HCM patients used in our classification experiment is 221, which is much higher than other previous work on HCM detection.

In this study, we aim to develop a classifier that can distinguish between ECG signals from HCM patients and those from non-HCM controls. Such a classifier will facilitate automated detection of HCM from ECG signals. However, we note that the classifier is not expected to replace extensive cardiovascular diagnosis. Rather, it is intended as an initial screening method that will hopefully detect patients that may have HCM. The automatically detected patients will be referred for further cardiovascular tests and be examined by expert cardiologists.

In order to develop a classifier for automated detection of patients with HCM, we have segmented ECG signals into individual heartbeats, extracted features from each heartbeat and then classified these heartbeats by applying machine learning methods. We assigned a patient to the HCM class if the number of heartbeats classified as HCM is equal to or greater than the number of heartbeats classified as control. For our classification experiments, we have extracted features that have been previously used, as well as some new morphological features (amplitude values of ECG waves) from ECG signals. We have applied random forests and support vector machines classifiers to distinguish between heartbeats from HCM and those from non-HCM patients. Using 5-fold and 10-fold cross validation for training and testing, we achieve high performance levels as measured in terms of *precision, recall (sensitivity), specificity* and *F-measure*. For comparison, we also applied logistic regression as a baseline classifier. We use feature selection to reduce the number of features required to achieve the same performance level as that obtained by using the complete set of features.

The rest of the paper is organized as follows: Section II describes the ECG dataset obtained from HCM patients and from control subjects, which is used in our classification experiments. In Section III, we discuss feature extraction, feature selection, and classification methods, as well as related tools. All classification results are presented in Section IV. We discuss and analyze the results and present directions for future work in Section V.

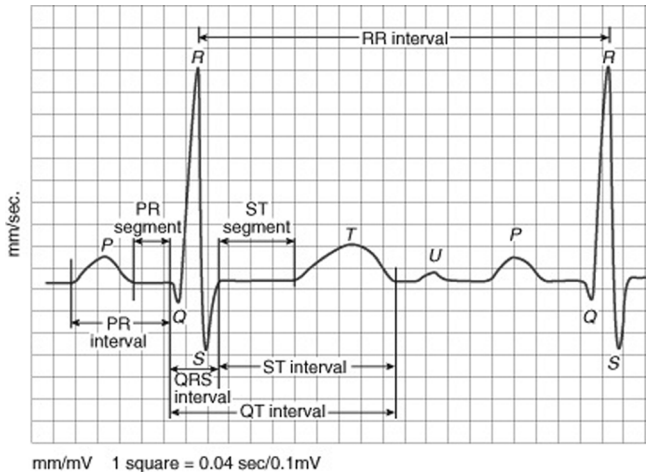


Fig. 2. A typical heartbeat comprising P, Q, R, S, T, U waveforms and inter-wave segments and intervals [36].

TABLE I
SUMMARY OF THE ECG DATASET USED IN THIS STUDY. EACH HCM PATIENT HAS ONE OR MORE ECG SIGNALS, WHEREAS EACH OF THE CONTROLS HAS ONLY ONE SIGNAL IN THE DATASET

Type of patient	Number of patients	Total number of ECG recordings	Total number of Heartbeats
HCM	221	754	6488
ICD (Control)	541	541	4442

II. DATA

The ECG dataset used in this study comprises standard 10-second, 12-lead ECG signals from two groups of cardiovascular patients. The first group consists of 221 hypertrophic cardiomyopathy (HCM) patients. Each HCM patient has one or more ECG recordings in the dataset. The total number of ECG signals in the HCM patients' dataset is 754. In the second group there are 541 subjects, all of which were diagnosed with ischemic or non-ischemic cardiomyopathy, and had implantable cardioverter defibrillator (ICD) installed for primary prevention of sudden cardiac death. As none of the ICD patients was diagnosed with HCM, their ECG data is used as the control in the experiments described here. While there may be cases in which a set of healthy controls would be preferable (e.g., pre-screening for HCM among young athletes), we have chosen the ICD patients' ECG dataset as the control because most of the patients referred for ECG tests in a hospital do not usually have a normal cardiac diagnosis; accordingly distinguishing HCM patients from other cardiovascular patients is a realistic, essential task. That said, we expect the methods used in this study to be applicable in other scenarios of distinguishing HCM patients from another group. Each patient in our control dataset has exactly one ECG recording, resulting in a total of 541 ECG signals the control set.

We segmented each ECG signal into individual heartbeats using the freely available ECGPUWAVE tool [20]. A heartbeat is a single cycle in which the heart's chambers relax and contract

TABLE II
COMPLETE LIST OF THE 42 FEATURES EXTRACTED FROM EACH OF THE 12-LEAD ECG SIGNALS FOR CLASSIFYING HEARTBEATS. (THE TOTAL NUMBER OF FEATURES IS $42 \times 12 = 504$)

Group	Feature	Definition	Number of features
Temporal (based on length of intervals)	Pre-RR interval	The interval between the current heartbeat and the previous heartbeat	6
	Post-RR interval	The interval between the current heartbeat and the following heartbeat	
	Average RR-interval	The mean of the RR intervals of a recording and the it is used as the same for all the heartbeats in a recording	
	P-wave duration	The interval between the P-wave onset and offset	
	QRS interval	The interval between the QRS onset and offset	
	T-wave duration	The interval between QRS-offset and T-wave offset	
Morphological (based on amplitude values)	QRS morphology	10 uniformly sampled amplitude values between the QRS onset and the QRS offset.	36
		Maximum and minimum of original sampled amplitude values in the QRS complex.	
	P and T wave morphology	10 uniformly sampled amplitude values between the wave onset and the wave offset.	
		The maximum and the minimum of the original sampled amplitude values in the P and T wave.	

to pump blood, where each heartbeat comprises multiple waveforms. The ECG waves are created by the electrical signal that passes through the heart chambers (atria and ventricles). Fig. 2 shows a typical heartbeat and its waves: P, Q, R, S, T, and U. It also shows inter-wave segments and intervals. While identifying each heartbeat, ECGPUWAVE detects the onset and offset points of the P-wave and the QRS-complex. It also identifies the offset point of the T-wave and the peak of the QRS-complex.

The segmentation of ECG signals was conducted on signals from each of the 12 leads. We then identified the heartbeats that are simultaneously detected on all 12-leads. Each of these heartbeats was classified using machine learning methods as described in Section III-B. The summary of the dataset is presented in Table I.

III. METHODS AND TOOLS

After segmenting the 12-lead ECG signals into individual heartbeats, we extracted features from each heartbeat and represented it as a feature vector for classification. We also applied feature selection to identify highly informative features, and repeated the classification experiments using the selected features. We compared the results obtained from the different classification experiments and assessed the statistical significance of the observed differences. Finally, we identified HCM patients, by classifying each subject based on his/her respective number of heartbeats classified as HCM. The methods and tools used are discussed next.

A. Feature Extraction

As described in Section II, we utilized the ECGPUWAVE tool to detect individual waveforms from heartbeats of HCM and ICD patients. We utilized the onset and offset points of various waveforms detected by the tool for extracting temporal and morphological features from each heartbeat. The peak of the QRS-complex was used to measure the length of intervals between the R-waves of consecutive heartbeats. The temporal features and the morphological features extracted from the QRS complex and the T-wave have been used in the literature for heartbeat classification in a different context, namely, automatic detection of arrhythmia in cardiovascular patients [11], [21] . In the current study, we add morphological features of the P-wave that have not been used before. The complete list of features is shown in Table II. To represent each heartbeat, we extract all 42 features from each of the 12 leads, resulting in a total of 504 features.

B. Heartbeat Classification and HCM Patient Detection

As a first step to automatically detect HCM patients from 12-lead ECG signals, we developed a classifier whose task was to assign each instance (heartbeat) into one of two possible classes: HCM or control. As noted before, in this study heartbeats from ICD patients serve as controls. We applied two standard classification methods: random forests [22] and support vector machine (SVM) [23]. We have chosen these two methods because they have been previously used and were reported to perform well when classifying heartbeats for arrhythmia detection [8], [10]. For comparison, we also conducted experiment using a logistic regression classifier [24], which is often employed in biomedicine for classification tasks [25]–[27].

Random forests form an ensemble classifier based on a collection of decision trees, learned from multiple random samples taken from the training set. Decision tree classifiers are constructed using the information content of each attribute; thus the decision-tree learning algorithms first select the most informative attributes for classification. Random samples from the training dataset are selected uniformly, with replacement, such that the total size of each random sample is the same as the size of the whole training set. To classify a new instance, each decision tree is applied to the instance, and the final classification decision is made by taking a majority vote over all the decision trees. We applied the standard random forests classification package in WEKA [28], using 500 trees in the random forests implementation. The number of features selected at random at each tree-node was set to $2\sqrt{n}$, where n is the total number of features. We chose this number because in our classification experiments we found it to perform well compared to several alternatives proposed in the literature (e.g., $\log_2 n$, $\sqrt{n}$, $2\sqrt{n}$, $\sqrt{n/2}$ [22] [29] [30]).

The second classification method, support vector machines (SVM), is primarily a binary linear classifier. A hyperplane is

learnt from the training dataset in the feature-space to separate the training instances for classification. The hyperplane is constructed such that the margin, i.e., the distance between the hyperplane and the data points nearest to it is maximized. If the training instances are not linearly separable, these can be mapped into high dimensional space to find a suitable separating hyperplane. In our experiments, we used the WEKA libsvm [31], employing the Gaussian radial basis function kernel.

Another classification method we used for comparison is *logistic regression*. Given a training dataset D consisting of instances $\vec{\mathbf{X}}^{\mathbf{1}}, \vec{\mathbf{X}}^{\mathbf{2}}, \ldots, \vec{\mathbf{X}}^{\boldsymbol{m}}$ where each is represented as a feature vector $< \vec{\mathbf{X}}_{\mathbf{1}}^{\boldsymbol{j}}, \vec{\mathbf{X}}_{\mathbf{2}}^{\boldsymbol{j}}, \ldots, \vec{\mathbf{X}}_{\boldsymbol{n}}^{\boldsymbol{j}} >$, a linear combination of the input features for $\vec{\mathbf{X}}^{\boldsymbol{j}}$ is defined as: $s^j = w_0 + \sum_{i=1}^{n} w_i \vec{\mathbf{X}}_{\boldsymbol{i}}^{\boldsymbol{j}}$. The conditional probabilities of the binary class variable, C over the values $\{HCM, Control\}$ given the instance $\vec{\mathbf{X}}^{\boldsymbol{j}}$ are calculated as:

$$\Pr(C = HCM \mid \vec{\mathbf{X}}^{\boldsymbol{j}}) = \frac{1}{1 + e^{-s^j}} \text{ and}$$
$$\Pr(C = Control \mid \vec{\mathbf{X}}^{\boldsymbol{j}}) = 1 - \Pr(C = HCM \mid \vec{\mathbf{X}}^{\boldsymbol{j}}),$$

where $g(s) = 1/1 + e^{-s}$ is known as the *logistic function*. The training dataset D is used to estimate the values of the parameter vector $\vec{\mathbf{W}} =< w_0, w_1, \ldots, w_n >$ such that the conditional data likelihood is maximized. The conditional data likelihood is the conditional probability of the observed heartbeat-classes in the training dataset *given* their corresponding feature vector. Thus, $\vec{\mathbf{W}}$ is estimated such that the following condition is satisfied:

$$\vec{\mathbf{W}} = \arg\max_{\vec{\mathbf{w}}} \prod_{\vec{\mathbf{x}}^{\boldsymbol{j}} \in \boldsymbol{D}} \Pr(C^j \mid \vec{\mathbf{X}}^{\boldsymbol{j}}, \vec{\mathbf{W}}).$$

An instance, $\vec{\mathbf{X}}^{\boldsymbol{j}}$ is assigned the class label $C = Control$ if $w_0 + \sum_{i=1}^{n} w_i \vec{\mathbf{X}}_{\boldsymbol{i}}^{\boldsymbol{j}} < 0$, and $C = HCM$ otherwise.

We used the *logistic* package in WEKA for implementing the logistic regression classifier that estimates the parameter vector, $\vec{\mathbf{W}}$ following the estimation method proposed by Cessie and Houwelingen [24].

In our classification experiment, we represented each heartbeat as a 504-dimensional vector of features where 42 features were extracted from each of 12-leads as described in Section III-A. We used the stratified 5-fold cross-validation procedure for training and testing.

Although we are classifying here individual heartbeats, recall that the goal of this study is to classify patients into two groups: HCM vs. control. Hence, we partitioned both HCM patients and control patients into 5 equal sized groups. Heartbeats from one group of HCM patients and from one group of control patients were included in the test set and the other four groups were used for training. We repeated the process 5 times such that each heartbeat from a HCM patient or from a control subject is tested exactly once. We also applied 10-fold cross validation in the same manner to verify the stability of the classification performance.

After classifying all heartbeats from a subject, we classified that subject as a HCM patient based on the number of heartbeats classified as HCM. If the number of heartbeats classified as HCM is equal to or higher than that of heartbeats that have been classified as control, the subject is classified as a HCM patient.

To evaluate the performance of both the heartbeat and the patient classification, we have used several standard measures, namely, *precision, recall (sensitivity),* and *specificity*. These measures are defined below, where true positives (*TP*) and true negatives (*TN*) are correctly classified HCM and control heartbeats (or patients), respectively; False positives (*FP*) denote control heartbeats (or patients) that are misclassified as HCM; HCM heartbeats (or patients) incorrectly classified as control are false negatives (*FN*);

$$Precision = \frac{TP}{TP + FP},$$
$$Recall(Sensitivity) = \frac{TP}{TP + FN},$$
$$Specificity = \frac{TN}{TN + FP}.$$

In addition to these three measures, we also calculate the *F-measure*, which is the harmonic mean of precision and recall, defined as:

$$F - measure = 2.\frac{Precision.Recall}{Precision + Recall}.$$

We compared the performance measures obtained by random forests, SVM and logistic regression, where the paired t-test was used to assess the statistical significance of the differences along each performance measure [32].

C. Feature Selection

We initially used all 504 features to classify heartbeats as HCM or control beats. Building classifiers from a large feature set can possibly lead to overfitting; moreover, including features that carry only negligible information about the heartbeat-class may incur unnecessary extra training time. To address these issues, we performed feature selection to reduce the number of features.

To select features that have high predictive value, we utilized the well-known *Information Gain* criterion [33]. For each feature, the information gain measures how much information is gained about the heartbeat-class when the value of the feature is obtained. It is calculated as the difference between the unconditional entropy associated with the heartbeat-class and the conditional entropy of the heartbeat-class given the value of a feature. These measures are formally defined as follows: Let $C = \{HCM, Control\}$ be the set of heartbeat-classes and V_F be the value of the feature F. The maximum likelihood estimate for the probability of a heartbeat to be recorded from a HCM patient, $Pr(C = HCM)$, is calculated as:

$$Pr(C = HCM) \approx \frac{\#\ of\ heartbeats\ from\ HCM\ patients.}{Total\ \#\ of\ heartbeats},$$

while the same estimate for a Control heartbeat is calculated as:

$$Pr(C = Control) = 1 - Pr(C = HCM).$$

Similarly, we define the conditional probability of the heartbeat-class to be *HCM* (or *Control*), given the value of feature F, as: $Pr(C = U \mid V_F = x_i, 1 \leq i \leq k)$ where U is either *HCM* or *Control* and x_i is one of k possible values of F. The conditional probabilities are estimated from the observed proportions; e.g., the probability of the heartbeat-class to be *HCM* given that

the value of feature F is x_i, $Pr(C = HCM \mid V_F = x_i)$, is estimated as:

$$Pr(C = HCM \mid V_F = x_i) \approx \frac{\begin{array}{c}\#\ of\ heartbeats\ from\ HCM\ patients \\ that\ have\ x_i\ as\ the\ value\ of\ F\end{array}}{\begin{array}{c}Total\ \#\ of\ heartbeats \\ that\ have\ x_i\ as\ the\ value\ of\ F\end{array}}.$$

For a heartbeat-class variable, C, the entropy $H(C)$ is defined as:

$$H(C) = -[Pr(C = HCM)\log_2 Pr(C = HCM) + Pr(C = Control)\log_2 Pr(C = Control)].$$

The conditional entropy associated with C given that the value of the feature F is x_i is defined as:

$$H(C \mid V_F = x_i) = -[Pr(C = HCM \mid V_F = x_i) \times \log_2 Pr(C = HCM \mid V_F = x_i) + (C = Control \mid V_F = x_i) \log_2 \Pr(C = Control \mid V_F = x_i)].$$

Based on this definition, the conditional entropy associated with C given a feature F is calculated as:

$$H(C \mid V_F) = \sum_{i=1}^{k} Pr(V_F = x_i) H(C \mid V_F = x_i).$$

The information gain associated with a feature F, $IG(C, V_F)$, is thus formally defined as:

$$IG(C, V_F) = H(C) - H(C \mid V_F).$$

The above formal definition of information gain is based on the assumption that the features are discrete. As all features in our study are continuous, they first must be discretized. We calculated the information gain using the feature selection package in WEKA, which first discretizes continuous features following Fayyad and Irani's algorithm [34].

After calculating the information gain for each feature, we removed the 20 least-informative features and repeated the 5-fold cross validation experiment. We continued conducting this procedure by gradually removing 20 features at a time until we observed decline in performance. Notably, only the training dataset is used for information gain calculation and feature selection. Once the reduced feature set has been determined, the test set is represented based on the selected features.

IV. Results and Discussion

As explained in Section III-B, the first step in our experiment toward identifying HCM patients was to classify individual heartbeats such that each heartbeat is assigned to one of the two classes: HCM or control. We applied random forests and support vector machine using the complete set of 504 features for heartbeat classification. As noted earlier, we also used logistic regression for comparison. Table III shows the results from the 5-fold cross validation experiments using all three classifiers. Both random forests and SVM performed better than logistic regression. Differences in precision and specificity between logistic regression and the other two classifiers are statistically significant ($p < 0.05$). Therefore we do not use logistic regression further in the rest of our experiments, namely, patient classification and feature selection. For both random forests and SVM classifiers, precision (0.94) and F-measure (0.91) are the same. The small differences in recall and specificity for these two classifiers are not statistically significant ($p > 0.35$). We also conducted 10-fold cross validation experiments using the complete feature set and the results are shown in Table IV. All four performance measures are exactly the same for both 5-fold and 10-fold cross-validation. Hence we apply 5-fold cross validation for training and testing random forests and SVM classifiers using the reduced set of features as described below.

TABLE III
Heartbeat Classification Results Using all 504 Features (5-fold Cross Validation). Standard Deviation Is Shown in Parentheses

Classifier	Precision	Recall (Sensitivity)	Specificity	F-measure
RF (all features)	0.94 (0.02)	0.87 (0.03)	0.92 (0.02)	0.91 (0.03)
SVM (all features)	0.94 (0.03)	0.88 (0.03)	0.91 (0.03)	0.91 (0.02)
Logistic Regression (all features)	0.90 (0.02)	0.85 (0.02)	0.86 (0.02)	0.87 (0.02)

TABLE IV
Heartbeat Classification Results Using all 504 Features (10-fold Cross Validation). Standard Deviation Is Shown in Parentheses

Classifier	Precision	Recall (Sensitivity)	Specificity	F-measure
RF (all features)	0.94 (0.02)	0.87 (0.02)	0.92 (0.03)	0.91 (0.02)
SVM (all features)	0.94 (0.03)	0.88 (0.02)	0.91 (0.03)	0.91 (0.02)

To investigate how the four performance measures change when the number of features is reduced, we first calculated information gain for each feature. The highest information gain was 0.67 and the lowest was 0.001. Fig. 3 shows a histogram of the information gain distribution across features, where the x-axis shows the information gain values and the y-axis shows the number of features associated with each information gain. As values on the x-axis are rounded to 2 decimal points, an information gain of less than 0.01 is shown as zero (the leftmost column on the graph). We observe that more than 300 features (four columns from the left) are associated with a negligible information gain (less than 0.04). We expect that removing some of these features will not lead to significant reduction in the classification performance. Therefore, as described in Section III-C, we gradually removed the least-informative features, 20 at a time, and repeated the heartbeat classification experiment using both random forests and SVM. The change in performance in terms of all four measures using random forests for classification is shown in Fig. 4. All four performance measures fluctuate slightly as we continue removing features until the number of features reaches 264. After that, the performance steadily declines as additional features are

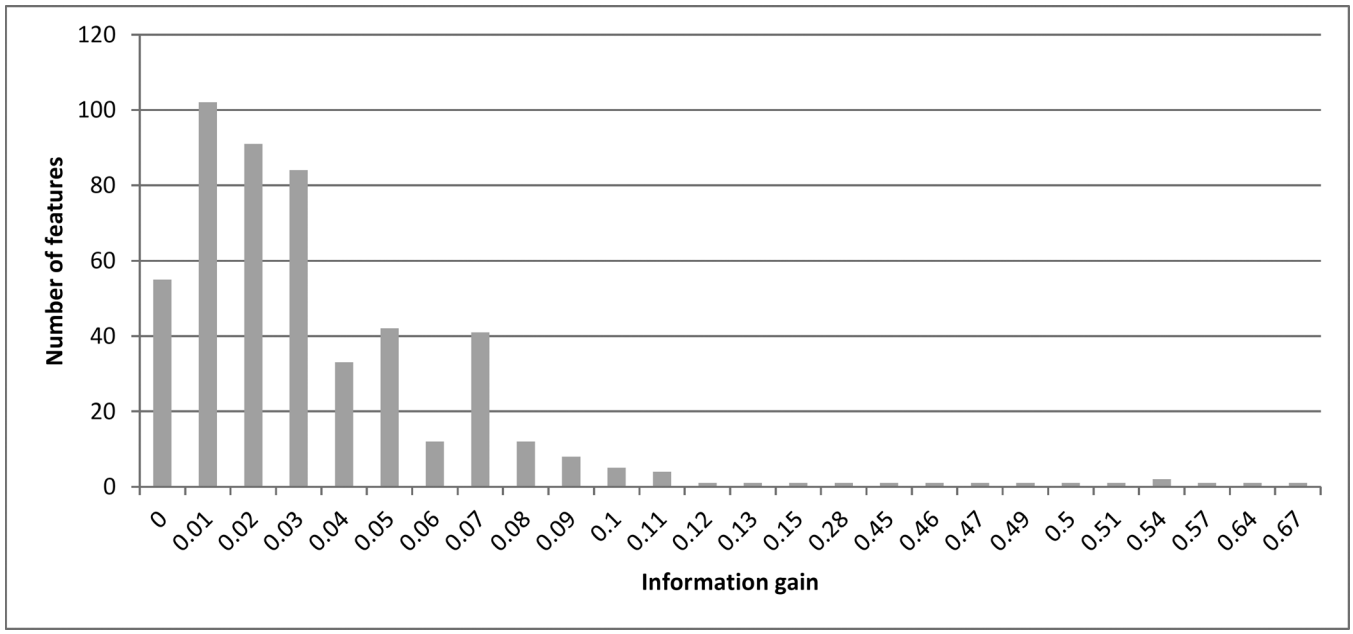


Fig. 3. Histogram of the information gain distribution across 504 features.

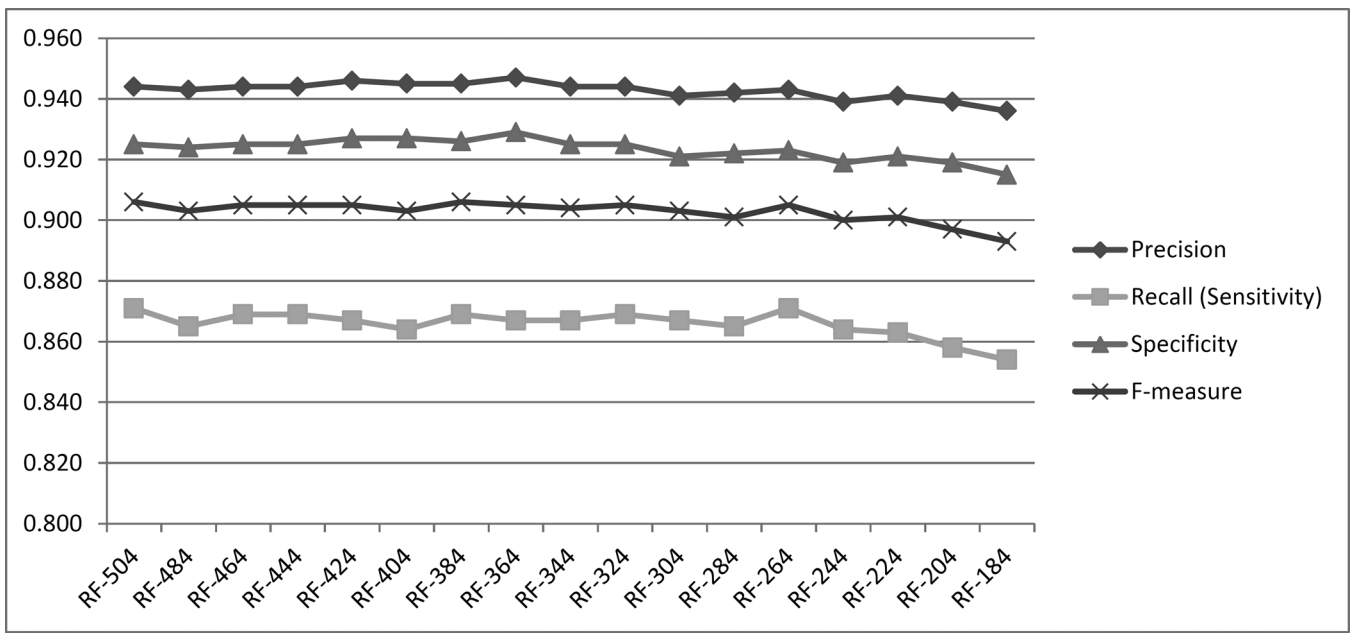


Fig. 4. Performance measures from heartbeat classification using random forests while gradually removing 20 features at a time.

removed. All four measures, obtained when using 264 features in our representation, are exactly the same as those obtained when using the complete set of 504 features. We have also plotted the performance measures for SVM while removing 20 features at a time, as shown in Fig. 5. The performance remains almost the same when gradually reducing the number of features from 504 to 404. Beyond that, the performance declines steadily as we remove additional features.

The next step in identifying HCM patients was to classify each subject as belonging to one of two classes: HCM or non-HCM. If the percentage of heartbeats classified as HCM was 50% or more, the subject was classified as an HCM patient. Table V shows results of patient classification, where the heartbeats used in the classification were represented based on all 504 features. Random forests and SVM perform almost the same and the marginal difference in performance measures is not statistically significant ($p > 0.85$)

As 264 features for the random forests classifier and 404 features for the SVM classifier performed the same as the complete feature set when classifying individual heartbeats, we used the respective reduced feature sets to identify HCM patients based on the number of heartbeats categorized as HCM. Patient classification results are presented in Table VI, where heartbeats were represented using 264 features for random forests and 404 features for SVM. The paired t-tests show no statistically-significant performance-difference between SVM and random forests for classifying patients, when the reduced feature-sets are used for heartbeat classification ($p > 0.58$).

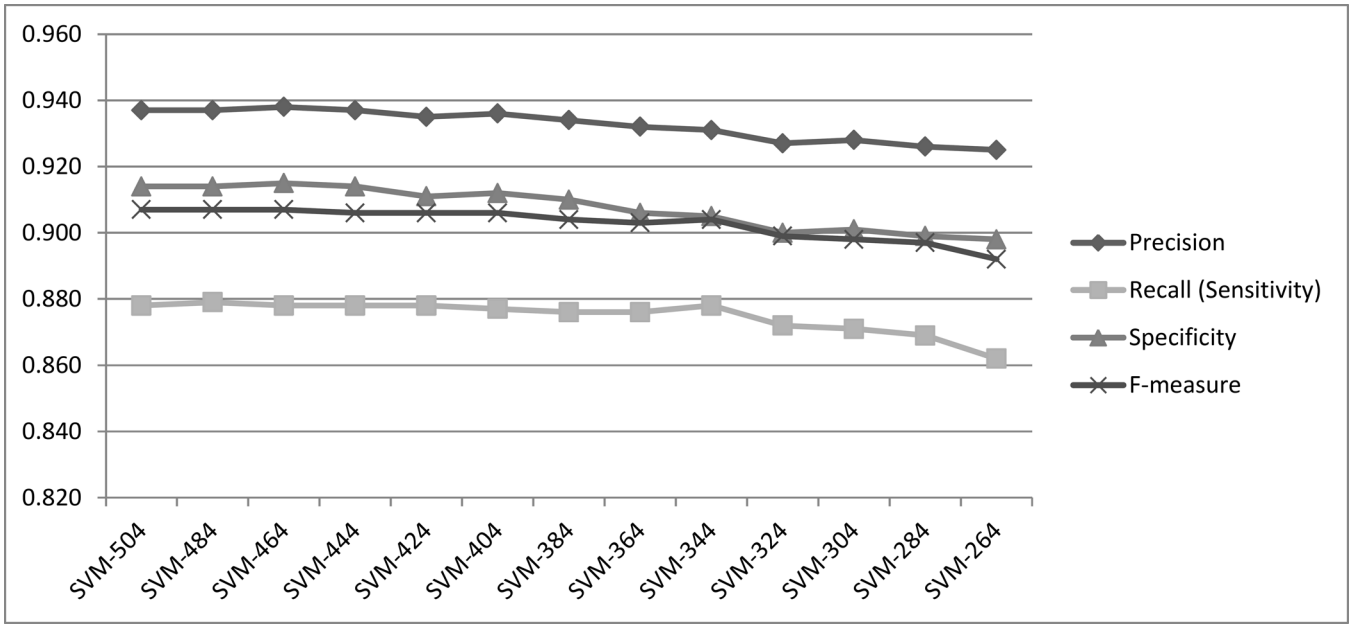


Fig. 5. Performance measures from heartbeat classification using SVM while gradually removing 20 features at a time.

TABLE V
RESULTS FROM THE PATIENT CLASSIFICATION EXPERIMENT, WHERE HEARTBEATS WERE CLASSIFIED USING THE COMPLETE SET OF 504 FEATURES. STANDARD DEVIATION IS SHOWN IN PARENTHESES

Classifier	Precision	Recall (Sensitivity)	Specificity	F-measure
RF (all features)	0.84 (0.05)	0.89 (0.04)	0.93 (0.02)	0.86 (0.04)
SVM (all features)	0.83 (0.05)	0.90 (0.03)	0.92 (0.03)	0.87 (0.03)

TABLE VI
RESULTS FROM THE PATIENT CLASSIFICATION EXPERIMENT WHERE HEARTBEATS WERE CLASSIFIED USING REDUCED SETS OF 264 (RF) AND 404 (SVM) FEATURES. STANDARD DEVIATION IS SHOWN IN PARENTHESES

Classifier	Precision	Recall (Sensitivity)	Specificity	F-measure
RF (264 features)	0.84 (0.05)	0.89 (0.04)	0.93 (0.03)	0.86 (0.04)
SVM (404 features)	0.82 (0.05)	0.89 (0.03)	0.92 (0.03)	0.85 (0.03)

The classification results described above show that we were able to achieve high performance level while identifying HCM patients from 12-lead ECG data by classifying individual heartbeats using a set of 504 features. We also demonstrate that reduced feature-sets, obtained by gradually removing the least informative features, performs equally well. The statistical tests applied show that the difference in performance obtained by random forests and by support vector machines is not statistically significant.

V. CONCLUSION

We have classified individual heartbeats from standard 10-second, 12-lead ECG signals to identify hypertrophic cardiomyopathy (HCM) patients. We have used ECG signals from HCM patients and from non-HCM controls to train and test heartbeat classifiers by applying random forests and support vector machines. A comprehensive set of 504 features extracted from ECG signals was used for heartbeat representation and classification. A subject was identified as a HCM patient if the majority of heartbeats for the patient were classified as HCM. The four performance measures from the patient classification experiment using random forests are: *precision 0.84, recall 0.89, specificity 0.93* and *F-measure 0.86*; similar performance measures were obtained by using SVM, as confirmed by the paired t-test. For comparison, we have also applied the logistic regression method to classify heartbeats, which showed a diminished level of performance compared to both random forests and SVM. We have used the information-gain criterion for selecting highly informative features to represent the heartbeats in the training and in the test set. For random forests, performance measures using 264 selected features were similar to the measures obtained using the complete set of 504 features. For SVM, this was true for a set of 404 informative features.

This work is the first study of its kind, setting out to automatically identify HCM patients from 12-lead ECG signals by classifying heartbeats using machine-learning methods. We have shown that it is possible to attain high performance using random forests or SVMs. We also showed that the information-gain criterion can be effectively used to choose a reduced set of temporal and morphological features that retain a similar level of performance. While in this study we have classified patients simply based on the percentage of individual heartbeats classified as HCM, in future research we shall focus on analyzing and modeling the sequence of heartbeats using advanced machine learning methods.

REFERENCES

[1] B. J. Maron and L. Salberg, *Hypertrophic Cardiomyopathy: For Patients, Their Families and Interested Physicians*. Chichester, U.K.: Wiley-Blackwell, 2008.

[2] B. J. Maron, "The electrocargiogram as a diagnostic tool for hypertrophic cardiomyopathy: Revisited," *Ann. Noninvasive Electrocardiol.*, vol. 6, no. 4, pp. 277–279, 2001.

[3] Q. A. Rahman, S. Kanagalingam, A. Pinheiro, T. Abraham, and H. Shatkay, "What we found on our way to building a classifier: A critical analysis of the AHA screening questionnaire," in *Proc. BHI 2013 LNAI*, ser. Lecture Notes in Artificial Intelligence, K. Imamura, S. Usui, T. Shirao, T. Kasamatsu, L. Schwabe, and N. Zhong, Eds., 2013th ed. Heidelberg, Germany: Springer, 2013, vol. 8211, pp. 225–236.

[4] S. Kuo and R. Dillman, "Computer detection of ventricular fibrillation," *Proc. Comput. Cardiol.*, pp. 347–349, 1978.

[5] R. H. Clayton, A. Murray, and R. W. F. Campbell, "Comparison of four techniques for recognition of ventricular fibrillation from the surface ECG," *Med. Biol. Eng. Comput.*, no. 31, pp. 111–117, 1993.

[6] M. Nygards and J. Hulting, "Recognition of ventricular fibrillation utilizing the power spectrum of the ECG," *Proc. Comput. Cardiol.*, pp. 393–397, 1977.

[7] M. Llamedo and J. Marte, "Heartbeat classification using feature selection driven by database generalization criteria," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 3, pp. 616–625, Mar. 2011.

[8] F. Melgani and Y. Bazi, "Classification of electrocardiogram signals with support vector machines and particle swarm optimization," *IEEE Trans. Inf. Technol. Biomed.*, vol. 12, no. 5, pp. 667–677, Sep. 2008.

[9] S. Yu and K. Chou, "Integration of independent component analysis and neural networks for ECG beat classification," *Expert Syst. Appl.*, vol. 34, no. 4, pp. 2841–2846, May 2008.

[10] N. Emanet, "ECG beat classification by using discrete wavelet transform and random forest algorithm," in *Proc. 5th Int. Conf. Soft Comput., Comput. With Words Perceptions Syst. Anal., Decision, Control*, 2009, pp. 1–4.

[11] P. De Chazal and R. B. Reilly, "Patient-adapting heartbeat classifier using ECG morphology and heartbeat interval features," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 12, pp. 2535–2543, 2006.

[12] D. D. Savage, J. I. Drayer, W. L. Henry, E. C. Mathews, J. H. Ware, J. M. Gardin, E. R. Cohen, S. E. Epstein, and J. H. Laragh, "Echocardiographic assessment of cardiac anatomy and function in hypertensive subjects," *Circulation*, vol. 59, no. 4, pp. 623–632, Apr. 1979.

[13] A. Cohen, A. D. Hagan, J. Watkins, J. Mitas, M. Schvartzman, A. Mazzoleni, I. M. Cohen, S. E. Warren, and W. V. R. Vieweg, "Clinical correlates in hypertensive patients with left ventricular hypertrophy diagnosed with echocardiography," *Amer. J. Cardiol.*, vol. 47, no. 2, pp. 335–341, Feb. 1981.

[14] A. A. Carr, L. M. Prisant, and L. O. Watkins, "Detection of hypertensive left ventricular hypertrophy," *Hypertension*, vol. 7, no. 6_Pt._1, pp. 948–954, Nov. 1985.

[15] G. Schillaci, F. Battista, and G. Pucci, "A review of the role of electrocardiography in the diagnosis of left ventricular hypertrophy in hypertension," *J. Electrocardiol.*, vol. 45, no. 6, pp. 617–623, 2012.

[16] W. Kaiser, T. S. Faber, and M. Findeis, "Automatic learning of rules. A practical example of using artificial intelligence to improve computer-based detection of myocardial infarction and left ventricular hypertrophy in the 12-lead ECG," *J. Electrocardiol.*, vol. 29 Suppl., pp. 17–20, Jan. 1996.

[17] R. A. Warner, Y. Ariel, M. D. Gasperina, and P. M. Okin, "Improved electrocardiographic detection of left ventricular hypertrophy," *J. Electrocardiol.*, vol. 35 Suppl., pp. 111–115, Jan. 2002.

[18] D. Corrado and W. J. McKenna, "Appropriate interpretation of the athlete's electrocardiogram saves lives as well as money," *Eur. Heart J.*, vol. 28, no. 16, pp. 1920–2, Aug. 2007.

[19] S. L. P. Potter, F. Holmqvist, P. G. Platonov, K. Steding, H. Arheden, O. Pahlm, V. Starc, W. J. McKenna, and T. T. Schlegel, "Detection of hypertrophic cardiomyopathy is improved when using advanced rather than strictly conventional 12-lead electrocardiogram," *J. Electrocardiol.*, vol. 43, no. 6, pp. 713–8, Jan. 2010.

[20] P. Laguna, R. Jane, E. Bogatell, and D. Anglada, ECGPUWAVE [Online]. Available: http://www.physionet.org/physiotools/ecgpuwave/

[21] P. De Chazal, M. O'Dwyer, and R. B. Reilly, "Automatic classification of heartbeats using ECG morphology and heartbeat interval features," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 7, pp. 1196–1206, 2004.

[22] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[23] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 297, pp. 273–297, 1995.

[24] S. Le Cessie and J. Van Houwelingen, "Ridge estimators in logistic regression," *Appl. Stat.*, vol. 41, no. 1, pp. 191–201, 1992.

[25] A. Subasi and E. Erçelebi, "Classification of EEG signals using neural network and logistic regression," *Comput. Methods Programs Biomed.*, vol. 78, no. 2, pp. 87–99, May 2005.

[26] P. H. C. Eilers, J. M. Boer, G.-J. van Ommen, and H. C. van Houwelingen, "Classification of microarray data with penalized logistic regression," in *Proc. BiOS 2001 Int. Symp. Biomed. Opt.*, 2001, pp. 187–198.

[27] J. G. Liao and K.-V. Chin, "Logistic regression for disease classification using microarray data: model selection in a large P and small n case," *Bioinformatics*, vol. 23, no. 15, pp. 1945–1951, May 2007.

[28] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software," *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, p. 10, Nov. 2009.

[29] R. Genuer, J.-M. Poggi, and C. Tuleau, "Random forests: some methodological insights," *Arxiv:0811.3619v1, stat.ML*, Nov. 2008.

[30] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognit. Lett.*, vol. 31, no. 14, pp. 2225–2236, Oct. 2010.

[31] Y. El-Manzalawy, WLSVM: Integrating LibSVM into Weka environment [Online]. Available: http://weka.wikispaces.com/libSVM

[32] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998.

[33] T. M. Mitchell, *Machine Learning.* New York: McGraw-Hill, 1997, p. 432.

[34] U. Fayyad and K. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, 1993, pp. 1022–1027.

[35] Hypertrophic cardiomyopathy Mayo Clinic [Online]. Available: http://www.mayoclinic.com/health/medical/IM00586

[36] ECG Wave [Online]. Available: http://lifeinthefastlane.com/ecg-library/basics/t-wave/