Identifying Hypertrophic Cardiomyopathy Patients by Classifying Individual Heartbeats from 12-lead ECG Signals

Quazi Abidur Rahman[1], Larisa G. Tereshchenko[2,3], Matthew Kongkatong[2], Theodore Abraham[2], M. Roselle Abraham[2] and Hagit Shatkay[1,4]

[1]Computational Biology and Machine Learning Lab, School of Computing, Queen's University, Kingston, ON, Canada
[2]Heart and Vascular Institute, Johns Hopkins University, Baltimore, MD, USA
[3]Knight Cardiovascular Institute, Oregon Health & Science University, OR, USA
[4]Dept. of Computer and Information Sciences & Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA
quazi@cs.queensu.ca, tereshch@ohsu.edu, matthew.kongkatong@gmail.com, tabraha3@jhmi.edu, mabraha3@jhmi.edu, shatkay@cis.udel.edu

***Abstract*— Test based on electrocardiograms (ECG) that record the heart electrical activity can help in early detection of patients with hypertrophic cardiomyopathy (HCM) where the heart muscle is partially thickened and blood flow is (potentially fatally) obstructed. This paper presents a cardiovascular-patient classifier we developed to identify HCM patients using standard 10-seconds, 12-lead ECG signals. Patients are classified as having HCM if the majority of the heartbeats are recognized as 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. The patient-classification precision of both classifiers are 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 304 highly informative features can achieve performance measures comparable to that achieved by using the complete set of features.**

Keywords-Hypertrophic Cardiomyopathy; Electrocardiogram; Patient classification; Machine learning methods; Feature selection

I. INTRODUCTION

Hypertrophic cardiomyopathy (HCM) is a genetic cardiovascular disease that may cause sudden cardiac death in young people [1]. In HCM patients, a portion of the heart muscle (myocardium) is thickened (hypertrophied). The most consistent trait of HCM is the thickening of the muscle at the lower left chamber of the heart (left ventricle). An imaging method, two-dimensional echocardiography, is often used to identify left ventricular hypertrophy (LVH). However, this method is not reliable when the thickening of the muscle of the left ventricle is not clearly detectable. Moreover, early prediction of the disease for patients without any thickening of the muscle 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 in the left ventricle has high diagnostic value for early detection and prediction of HCM. We have also shown in a recent study [3] that the standard procedure of conducting ECG tests should not be ignored in mass pre-participation screening of young athletes.

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 life-threatening arrhythmia in cardiovascular patients. Traditional machine learning methods such as artificial neural networks [4], support vector machines [5], random forests [6], and linear discriminants [7] have been used for reliable detection of life-threatening arrhythmia. Morphological (e.g., amplitude values) and temporal (e.g., length of various intervals) features have been primarily extracted for heartbeat-classification.

As mentioned earlier, left ventricular hypertrophy is the most consistent trait among HCM patients. Several criteria have been proposed to detect cardiovascular patients with left ventricular hypertrophy (LVH) from ECG signals. These criteria are usually derived from amplitude values of ECG waveforms. Many studies have been conducted to validate these LVH-detection criteria. While these studies have generally achieved high specificity (approximately 100%), sensitivity has been low (approximately 50%) across different studies [8]. Besides this criterion-based approach, multiple linear regression [11] and rule-based methods [12] have been used to detect cardiovascular patients with LVH, where morphological features were extracted from ECG signals. Potter et al. [9] have tested a set of criteria specifically proposed for detecting HCM [10] on a small group of 56 HCM patients and 56 healthy control subjects. 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.

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.

978-1-4799-5669-2/14/$31.00 ©2014 IEEE

However, we note that the classifier is not expected to replace extensive diagnosis by an expert cardiologist. 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 extensive cardiovascular tests and be examined by expert cardiologists.

In order to develop a patient classifier for automated detection of 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 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 cross validation for training and testing, we achieve high performance levels as measured in terms of *precision, recall (sensitivity), specificity* and *F-measure*. We also reduce, through feature selection, 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 used for classification experiments, obtained from HCM patients and from control subjects. In Section III, we discuss feature extraction, classification and feature selection methods and related tools. All classification results are presented in Section IV. Finally, we discuss and analyze the results and present directions for future work in Section V.

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 implantable cardioverter defibrillator (ICD) were inserted in their hearts 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.

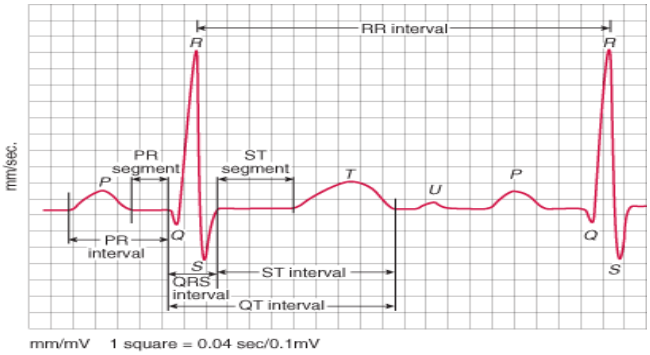

Figure 1. A typical heartbeat comprising P, Q, R, S, T, U waveforms and inter-wave segments and intervals [22].

We segmented each ECG signal into individual heartbeats using the freely available ECGPUWAVE tool [13]. A heartbeat is a single cycle in which the heart's chambers relax and contract 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). Figure 1 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.

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

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 RR-intervals between consecutive heartbeats. The temporal features and the morphological features extracted from the QRS complex and the T-wave have been used in the literature [7], [14] for heartbeat classification in a different context, namely, automatic detection of arrhythmia in cardiovascular patients. 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. Classification of heartbeats and detection of HCM patients

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. Heartbeats from ICD patients are the controls in this study. We applied two standard classification methods: random forests [15] and support vector machine (SVM) [16]. We used the standard classification package in WEKA [17] for random forests and the libsvm package [18] in WEKA for SVM. The random forests algorithm was implemented with 500 trees; the number of features selected at random at each node of a tree was set to $2\sqrt{n}$, where n is the total number of features. Our SVM experiments used Gaussian radial basis function as the kernel.

In the 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 individual heartbeats, the final goal of this study is to classify patients into two groups: HCM and ICD. Hence, we partitioned both HCM patients and ICD 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 an HCM patient or from a control subject is tested exactly once.

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 the 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:

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

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 42X12=504)

Group	Feature	Definition	Number of features
Temporal (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 (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.	

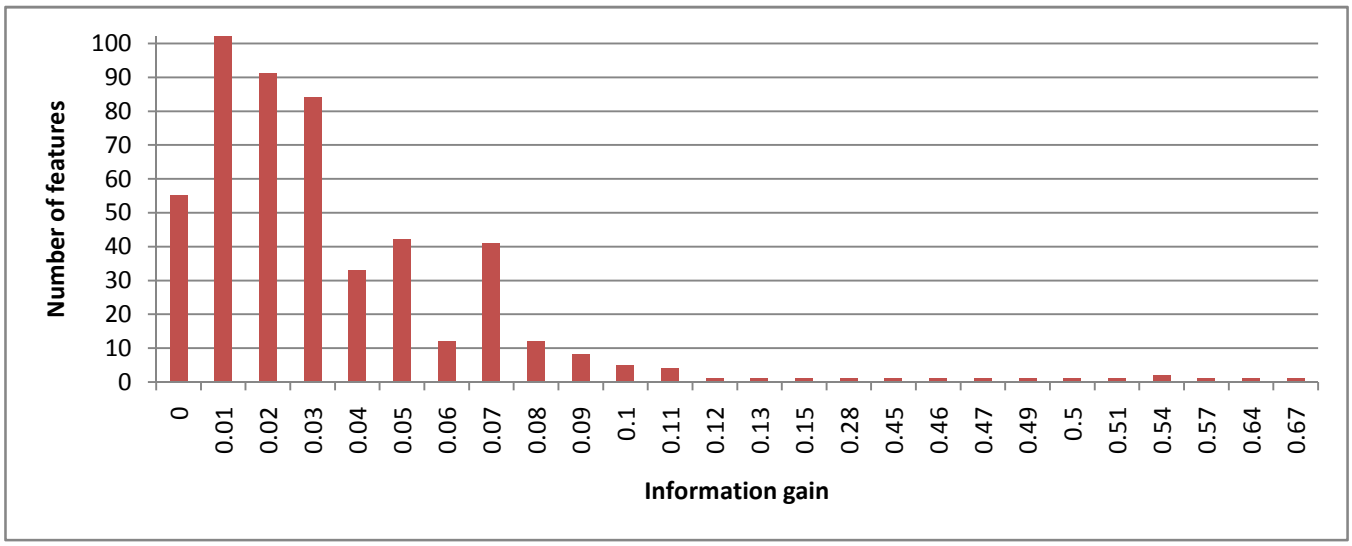

Figure 2. Histogram of the information gain distribution across 504 features.

We compared the performance measures obtained by random forests and SVM, where the paired t-test [19] was used to assess the statistical significance of the differences along each performance measure.

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 capability, we utilized the well-known *Information Gain* criterion [20] that measures how much information is gained about the heartbeat-class when the value of the feature is known. This criterion is primarily defined for nominal attributes. As all features in our study are numerical, they need to be discretized for calculating information gain. We calculated the information gain using the feature selection package in WEKA, which discretizes numerical attributes following Fayyad and Irani's algorithm [21].

After calculating information gains, we removed the 20 least-informative features and conducted the classification experiment again using both random forests and SVM. We repeated this procedure by gradually removing 20 features at a time until we observed decline in performance.

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 classifiers using the complete set of 504 features for heartbeat classification. Table III shows the results from the 5-fold cross validation experiments. Precision (0.94) and F-measure (0.91) are the same for both classifiers. The small differences in recall and specificity are not statistically significant ($p > 0.35$).

To investigate how these performance measures change when the number of features is reduced, we calculated information gain for selecting highly predictive features. The highest information gain was 0.67 and the lowest was 0.001. Figure 2 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).

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 Figure 3. All four performance measures marginally increase and decrease as we continue removing features until the number of features reaches 304. After that, the performance steadily declines as additional features are removed. All four measures for 304 features 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 Figure 4. The performance remains almost the same between 504 features and 384 features. Then it starts to decline steadily as we remove more features. The performance measures attained using 504 features and 384 features are exactly the same.

Table III. Heartbeat classification results using all 504 features.

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

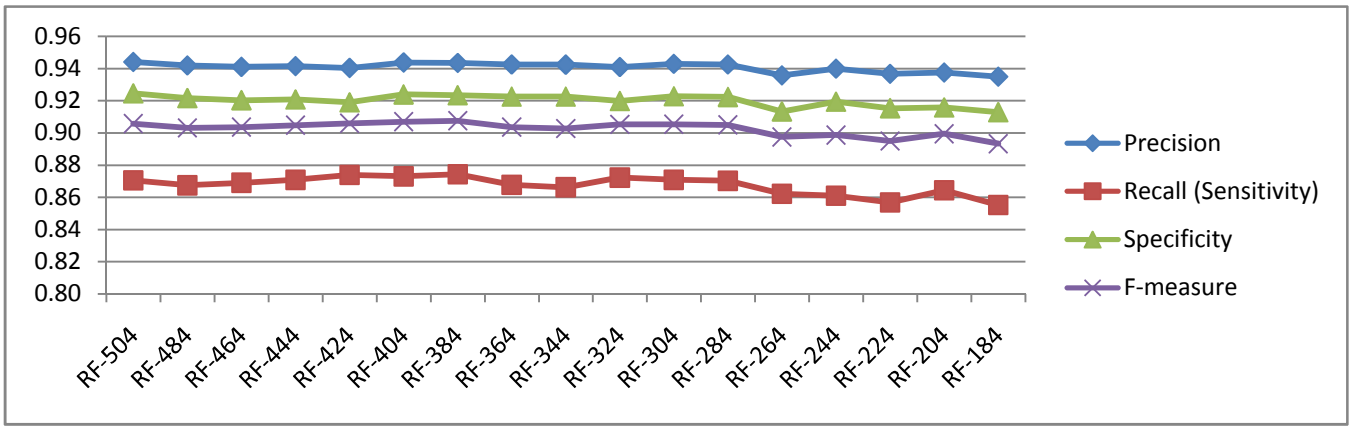

Figure 3. Performance measures from heartbeat classification using random forests while gradually removing 20 features at a time.

The next step in identifying HCM patients was to classify each subject as 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 IV shows patient classification results where heartbeats were classified using all 504 features. Random forests and SVM perform almost the same and the marginal difference between performance measures is not statistically significant *(p>0.85)*

As 304 features for the random forests classifier and 384 features for the SVM classifier performed exactly 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 classified as HCM. Patient classification results are presented in Table V, where heartbeats were represented using 304 features for random forests and 384 features for SVM. Paired t-tests show no performance-difference between SVM and random forests for classifying patients, when the reduced feature-sets are used for heartbeat classification *(p>0.38)*.

The classification results described above show that we were able to achieve high performance level identifying HCM patients on results described above show that we were able to achieve high performance level identifying HCM patients from 12-lead ECG data by classifying individual heartbeats using a set of 504 features. We also demonstrate that the reduced feature-sets, obtained by gradually removing the least informative features, performed 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 was 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.* SVM performed similarly, as confirmed by the paired t-test. We have also used the information-gain criterion for

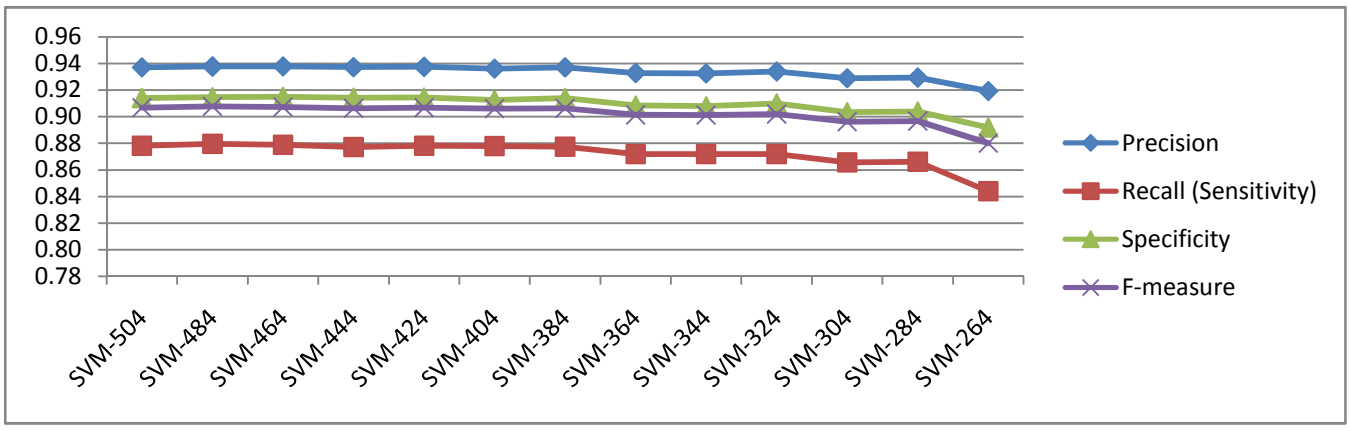

Figure 4. Performance measures from heartbeat classification using SVM while gradually removing 20 features at a time.

TABLE IV. PATIENT CLASSIFICATION RESULTS WHERE ALL 504 FEATURES ARE USED FOR HEARTBEAT CLASSIFICATION.

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

feature selection. For random forests, performance measures using 304 selected features were similar to the measures obtained using the complete set of 504 features. For SVM, this was true for a set of 384 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.

ACKNOWLEDGMENT

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.

REFERENCES

[1] B. J. Maron and L. Salberg, *Hypertrophic Cardiomyopathy: For Patients, Their Families and Interested Physicians*. Wiley-Blackwell, 2006.

[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 *BHI 2013, LNAI*, vol. 8211, K. Imamura, S. Usui, T. Shirao, T. Kasamatsu, L. Schwabe and N. Zhong, Eds. Springer, 2013, pp. 225–236.

[4] 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.

[5] 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. 667–77, Sep. 2008.

[6] N. Emanet, "ECG beat classification by using discrete wavelet transform and Random Forest algorithm," in *Fifth International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control*, 2009.

TABLE V. PATIENT CLASSIFICATION RESULTS WHERE REDUCED SETS OF 304 (RF) AND 384 (SVM) FEATURES ARE USED FOR HEARTBEAT CLASSIFICATION.

Classifier	Precision	Recall (Sensitivity)	Specificity	F-measure
RF (304 features)	0.85	0.89	0.93	0.87
SVM (384 features)	0.82	0.89	0.92	0.85

[7] 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.

[8] 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. 617–23, 2012.

[9] 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.

[10] 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.

[11] 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. 111–5, Jan. 2002.

[12] 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.

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

[14] 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.

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

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

[17] 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.

[18] Y. EL-Manzalawy and V. Honavar, (2005), *WLSVM: Integrating LibSVM into Weka Environment* [Online]. Available: http://www.cs.iastate.edu/~yasser/wlsvm.html

[19] T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998.

[20] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

[21] U. Fayyad and K. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *Thirteenth International Joint Conference on Articial Intelligence*, pp. 1022–1027, 1993.

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