Emily Doughty
Rachel Goldfeder
CS 229 Final Report

Characterizing and Diagnosing Hypertrophic Cardiomyopathy from ECG Data

Background

Hypertrophic cardiomyopathy (HCM) is a heart condition defined by a thickening of the heart muscle. This thickening makes it harder for the heart to pump blood throughout the body and also causes disturbances in the electrical functions of the heart that may lead to an arrhythmia. Approximately 1 in 500 people in the general population are affected by this condition [1]. Unfortunately, one of the potential symptoms of HCM is sudden cardiac death. In fact, HCM is a leading cause of sudden cardiac death in young athletes [2]. Risk factors for this disease include family history of sudden cardiac death, personal history of cardiac arrest, and tachycardia. People with these risk factors are screened for HCM; however, as the first symptom for HCM could be sudden cardiac death, early and accurate detection is critical. If HCM is diagnosed early, interventions such as alcohol septal ablation (destruction of the heart muscle) and septal myectomy (removal of part of the septum) can help reduce thickening of the heart; pacemakers can also be implanted to control and regulate electrical signals [3].

HCM is typically diagnosed using an echocardiogram, which allows the physician to measure heart wall thickness [4]. However, echocardiograms are expensive and require interpretation by a certified clinician. HCM diagnosis with a cheaper and more portable device that can be automatically interpreted, such as an electrocardiogram (ECG), would be a useful diagnostic tool in the clinic. ECGs noninvasively measure electrical activity of the heart over a period of time and produce data that can be mined for patterns. We hypothesize that the patterns of electrical activity recorded from a patient with HCM will be different from controls such that these patterns can be used to diagnose HCM.

Currently, the ECG features specifically associated with HCM are not well defined. In collaboration with Dr. Marco Perez and the Ashley lab at Stanford University, we obtained ECG records from patients diagnosed with HCM along with ECG records for controls. In this work, we compared five machine-learning algorithms on the task of classifying an ECG record as HCM or non-HCM using standard ECG measurements as input features.

Data & Methods

a. Data

<u>3-lead ECG data</u>

This dataset contained 279 patients with HCM that were diagnosed by a physician at Stanford Hospital and 1125 controls without HCM. The ECG analysis software, CardeaScreen [5], outputs 143 standard measurements from 3-lead ECGs, which were the attributes included in our dataset. These standard measurements include amplitudes, slopes, and lengths of various waveforms from the ECG.

<u>12-lead ECG data</u>

This dataset contained 260 patients with HCM (a subset of the patients with 3-lead data) that were diagnosed by a physician at Stanford Hospital and 1235 controls without HCM. The analysis software outputs 270 standard measurements from 12-lead ECGs, which were the attributes included in our dataset. Similar to the 3-lead measurements, these standard measurements include amplitudes, slopes, and lengths of various waveforms from the ECG. It must be noted that the 3-lead and 12-lead datasets did not contain exactly the same individuals: there were a total of 279 HCM patients and 1235 controls – most individuals had 3-lead and 12-lead ECGs, but some only had one type of ECG.

Preprocessing for both datasets included subject (HCM patients and controls) and feature removal. We removed features that did not provide any data (i.e., where all values of a given feature were missing or contained the same value) and subjects where the records were missing measurements. After this preprocessing, the 3-lead dataset contained 299 HCM records (there are some patients with multiple records), 1120 athlete records, and 136 standard measurements to be used as features and the 12-lead dataset contained 260 HCM records, 1235 control records, and 269 standard measurements to be used as features.

b. Methods

1. *Evaluation of five machine learning algorithms on full data.* We performed a 5-fold cross validation to evaluate the performance of random forest, support vector machines (SVM), AdaBoost, k-nearest neighbors (KNN), and naïve bayes in differentiating between ECG records from HCM patients and athletes. We performed our analysis in R. For random forest, we used the package **randomForest** [6] with 1000 trees and mtry = square root of the number of features. For SVM, we used the package **e1071** [7] with a linear kernel with a cost of constraint violation equal to 1. For KNN we used the package **e1071** with k=5. For naïve bayes we used package **e1071**. For AdaBoost we used package **ada** [8] with 50 boosting iterations. We averaged the following values across all folds: training error, test error, accuracy, sensitivity, precision, f-measure, and specificity.

2. *Feature selection and evaluation of machine learning algorithms.* We implemented feature selection using the mean decrease in Gini Index from random forest. Specifically, we trained a random forest classifier on all of the data and iteratively removed the least important features. For each feature subset, we repeated the above analysis for the other four algorithms.

Results

The results are summarized in Tables 1 and 2, with bolded values representing the algorithm that performed best for that metric. Random forest, SVM, and AdaBoost had the best performance across all metrics (PPV > 90%, accuracy > 95%, sensitivity > 80%). For this application, minimizing false negatives is more important than minimizing false positives, as improperly classifying a record from a person who truly has HCM could result in death. For this task, SVM had the best performance in terms of sensitivity. KNN performed the worst across all measures. In particular, KNN had the highest test error and lowest sensitivity, both of which are key for this application. Another important metric for this application is specificity. If we were to screen an entire population, we would want to minimize the number of people coming back to the clinic for additional testing. For example, 97% specificity with 100 healthy individuals would yield only three individuals returning for an unnecessary echocardiogram; however, if the total number of healthy individuals was 100 million, three million individuals may need to return to the clinic for an unnecessary and expensive echocardiogram. In our analyses, AdaBoost performed the best with a specificity of 0.988 for the 3-lead data (KNN had a specificity of 0.981 for the 12-lead data, but with a sensitivity of 0.434, KNN is not a useful method for this application area).

Table 1. 5-fold cross validation results for 3-lead data.

	Accuracy	Sensitivity	PPV	F-measure	Test Error	Training Error	Specificity
Random Forest	0.951	0.839	0.920	0.878	0.049	0.000	0.980
SVM	**0.961**	**0.900**	0.919	**0.908**	**0.039**	**0.005**	0.978
AdaBoost	0.952	0.819	**0.946**	0.878	0.048	0.007	**0.988**
Naïve Bayes	0.911	0.796	0.788	0.792	0.089	0.080	0.942
KNN	0.870	0.434	0.900	0.581	0.130	0.104	0.987

Table 2. 5-fold cross validation results for 12-lead data.

	Accuracy	Sensitivity	PPV	F-measure	Test Error	Training Error	Specificity
Random Forest	**0.934**	0.750	0.852	**0.797**	**0.066**	0.000	0.972
SVM	0.922	**0.785**	0.774	0.779	0.078	**0.004**	0.951
AdaBoost	0.930	0.712	**0.866**	0.780	0.070	0.007	0.977
Naïve Bayes	0.888	0.619	0.707	0.657	0.112	0.111	0.945
KNN	0.873	0.362	0.807	0.497	0.127	0.100	**0.981**

The top 15 features from random forest are depicted in Figure 1 A and B for the 3- and 12- lead data, respectively. For both datasets, the top 15 features were involved in the T wave. For example, in the 3-lead data T1 represents the largest T wave amplitude from V1 lead and TaVRM measures the T wave amplitude from the aVR lead. In the 12-lead data STI.1 measures the ST interval from lead V1 and ST1.aVR measures the ST interval from the aVR lead.

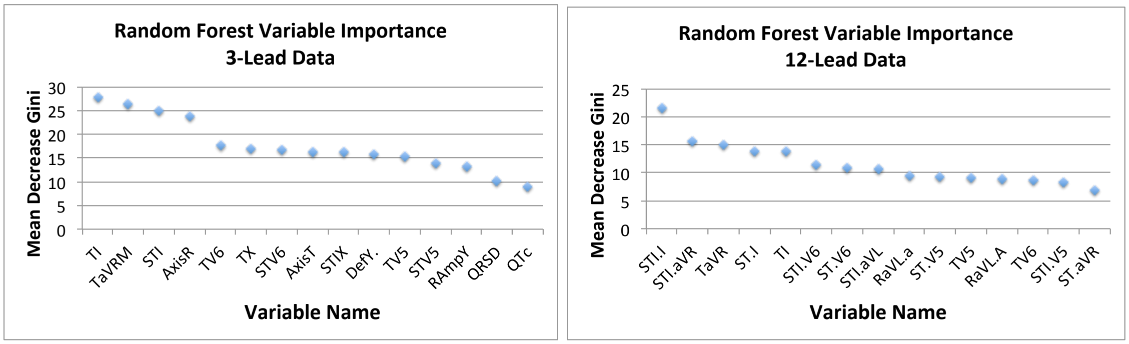

Figure 1. Important features from random forest for the 3- and 12-lead data.

Using the ordered top features based on Gini Index from random forest, we calculated all measures described above for the remaining four algorithms and for each subset of features for the 3- and 12-lead data. The impact of feature selection on sensitivity, specificity, and PPV are shown in Figure 2 for the 3-lead data. The 12-lead data showed similar trends but are not shown due to space restrictions. SVM reached maximum sensitivity at 65 top features. At greater than 20 features, KNN performed significantly worse than SVM, AdaBoost, and Naïve Bayes with regards to sensitivity. Thus, KNN is not a useful method for this task even with selected features. With regards to specificity, all four algorithms perform well (0.97 – 0.99 specificity) until 80 features are included, where Naïve Bayes began dropping in specificity. Feature selection showed greater variability on PPV than on sensitivity and specificity. AdaBoost stayed relatively constant for all amounts of features included, while the other three vary as the number of top features change. This shows that as the number of included top features increases, the false positives have a greater impact in relation to the true positives than the true negatives (as shown by specificity).

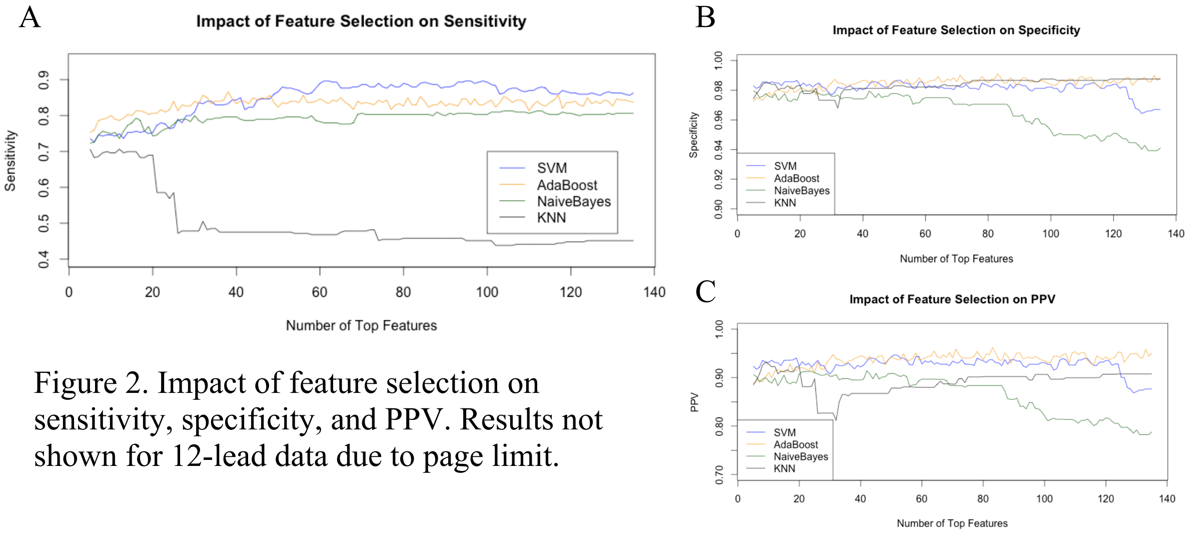

Figure 2. Impact of feature selection on sensitivity, specificity, and PPV. Results not shown for 12-lead data due to page limit.

Discussion

From this work, we 1) provide evidence that machine learning algorithms can accurately differentiate HCM from non-HCM ECG records and 2) provide clinicians with new features that are important for diagnosing HCM from ECGs. The algorithms performed better on 3-lead data than 12-lead data; this result was surprising since the 12-lead data is the known standard, while 3-lead is thought of as an approximation for the 12-lead data. Although the additional features provided by the 12-lead data may be useful for clinicians, it seems that the additional information is not informative for the machine learning algorithms.

With this idea in mind, we hypothesized that the algorithms may perform equally as well or better with fewer features. To test this, we reduced the dimensionality of the feature space by transforming the initial datasets (both 3-lead and 12-lead, separately) using Principal Component Analysis. We used four principal components to transform the data after reviewing the scree plot. We repeated our previous analysis and found that the algorithms performed significantly worse across all measures. All values decreased, with sensitivity having the most dramatic decrease between full data and reduced data for both 3- and 12-lead. Across all algorithms, the maximum sensitivity for the dimensionality-reduced data was 0.56, which is not acceptable for this task (results not shown due to space restrictions). Since PCA did not work well, instead, we decided to implement feature selection using the k most important features calculated by random forest.

This analysis showed that sensitivity, specificity, and PPV stay the same or increase for SVM, AdaBoost, and NaiveBayes even when using fewer features. This indicates that the rest of the features do not add useful information for the classification task.

We compared five algorithms in this analysis; however, in practice, a physician is likely to only use one algorithm when trying to diagnose a patient. To recommend one algorithm to use for this task, we can look at the metrics we used for evaluation (e.g., sensitivity, specificity, PPV); however, while a classifier that can predict HCM is useful in and of itself, availability of interpretable information about the features that lead to a given prediction is key for clinical utility. In particular, SVM and KNN lack interpretability, which may lead to limited clinical use. Thus, even though some of these classifiers may have high performance (e.g., SVM), they may not be the best choice for a clinical task. Using a classifier with interpretable information about how predictions were made can lead to a greater understanding of the disease. For instance, the most important features identified by random forest (i.e., the amplitude of the T wave and ST interval from lead V1 and lead aVR) give Dr. Perez new criteria to pay attention to when trying to use an ECG to determine if a patient has HCM.

Conclusion

In this work, we compared and contrasted the performance of five machine learning algorithms at the task of classifying ECG records as coming from a person with and without HCM. We found that SVM, random forest, and AdaBoost performed the best across nearly all metrics. Furthermore, all methods performed better on 3-lead data than on 12-lead data and performance is not decreased when only including the top 65 most important features. The algorithms have high enough sensitivity and specificity to suggest that ECGs can be used in place of echocardiograms for diagnosing HCM. This has a significant impact on clinicians' ability to screen for HCM and reduce the occurrence of sudden cardiac death. In conclusion, we have identified new electrical patterns that characterize HCM and shown that machine learning algorithms can accurately differentiate between patients with HCM and controls using ECG data.

Contributions

Emily, Rachel, and Dr. Perez conceived the project. Emily and Rachel performed all analyses, wrote the paper, and conveyed results to Dr. Perez.

References

[1] Maron, B.J., et al., Prevalence of hypertrophic cardiomyopathy in a general population of young adults. Echocardiographic analysis of 4111 subjects in the CARDIA Study. Coronary Artery Risk Development in (Young) Adults. Circulation, 1995. 92(4): p. 785-9.

[2] Maron, B.J., *Hypertrophic cardiomyopathy and other causes of sudden cardiac death in young competitive athletes, with considerations for preparticipation screening and criteria for disqualification.* Cardiol Clin, 2007. **25**(3): p. 399-414, vi.

[3] Maron, B.J., Hypertrophic cardiomyopathy: a systematic review. JAMA, 2002. 287(10): p. 1308-20.

[4] Wigle, E.D., et al., Hypertrophic cardiomyopathy. The importance of the site and the extent of hypertrophy. A review. Prog Cardiovasc Dis, 1985. 28(1): p. 1-83.

[5] https://www.cardeascreen.com/40/product.html

[6] Liaw, A and Wiener, M, Classification and Regression by randomForest. R News, 2002. 2(3): p. 18-22.

[7] Dimitriadou, E, et al.,. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. 2010. Available: http://cran.r-project.org/package=e1071.

[8] Michailides G, Johnson K, Culp M. ada: An r package for stochastic boosting. J Stat Softw. 2006;17:1–27.