Prediction of Arrhythmias using Support Vector Machines and Reduced Feature Subset of ECG Variability Metrics

Christopher Hoang

Thomas Jefferson High School for Science and Technology

Computer Systems Research Lab 2015-2016

Abstract

Sudden cardiac arrest, a life-threatening condition during which the heart suddenly stops beating, is caused by arrhythmias or irregular changes in heartbeat rhythm. This paper investigated the idea of using a support vector machine (SVM) approach with variability metrics derived from electrocardiogram (ECG) signals to predict short-term risk of arrhythmia episodes. The goal of this study was to present a classification model that could differentiate between healthy and at-risk patients given samples of the patients' ECG signal. The SVM classifiers used a feature vector composed of heart rate variability metrics and recently proposed morphological variability metrics; the feature vector was also subjected to sequential forward feature selection. The SVM classifiers were trained and tested on patient records in various medical databases in the PhysioNet repository. The strongest performing classification model had an overall accuracy of 92.12% for 165 patients. The proposed classification approach could be utilized by health care professionals as an early predictor of arrhythmia episodes and as an assistive tool in implantable cardioverter defibrillator recommendation.

I. Introduction

Sudden cardiac arrest is the number one killer in the United States, with an annual death toll of 325,000 adults [1]. This condition is caused by arrhythmia or an irregular change in heartbeat rhythm, which can be harmless or potentially severe, as it may result in sudden cardiac arrest. In cases of arrhythmia, the heart's electrical system does not function properly, causing the signals to travel in a manner which result in an irregular heartbeat rate [2]. Doctors will frequently utilize an implantable cardioverter defibrillator (ICD) to administer electrical pulses to the heart to restore a regular heartbeat rhythm, with larger electrical shocks for life-threatening situations which may result in sudden cardiac arrest [3].

Evaluation of patient risk for arrhythmia is based on a variety of risk factors such as family history of heart problems, electrocardiogram results, or presence of long QT syndrome [4]. Those at high risk are recommended for ICD treatments, but the surgical procedure for ICD implementation and the defibrillators themselves are potentially harmful to patients. The procedure to implant an ICD can be life-threatening and costly, with costs of upwards of $50,000. In addition, the ICDs in many patients are never activated, indicating the disjunction of ICD recommendation to patient ICD necessity. In a recent study [5], 22.5% of patients who

received an ICD did not match the guidelines for defibrillator use recommendation, yet 37% of the patients who received an ICD without guideline recommendation for defibrillator use suffered a cardiac incident in the 40 days following ICD implantation. A death rate of 0.57% was reported for the patients who did not meet medical guidelines to receive ICD treatment. Nevertheless, the study projects that more than 100,000 people who require ICD treatment fail to receive it every year [6].

Reaction time to sudden cardiac arrest incidents is also crucial to patient survival. The American Heart Association projects survival rates of those in ventricular fibrillation (VF) sudden cardiac arrest to decrease by 7 – 10% every minute after collapse of the victim [7]. If patients could be classified by their risk of suffering future arrhythmia episodes correctly, then there lies potential to improve ICD recommendation and to decrease reaction time for treatment of cardiac events caused by arrhythmias in at-risk patients.

This study proposed a support vector machine (SVM) approach to this classification problem. A feature vector of variability metrics was derived from samples of the electrocardiogram (ECG) signal records of patients in a physiological signals database. The feature vector was subjected to feature selection in order to obtain a smaller feature subset representative of the variability of the original feature vector. Using the feature subset as the input data, the classification model was trained and tested to predict the condition of a variety of patients as healthy or at risk of suffering a future arrhythmia episode. The final proposed SVM classifier had a classification accuracy of 92.12%. The variability metrics in the feature vector were also evaluated through the results of the feature selection algorithm used in this study.

II. Related Work

Heart rate variability (HRV), measures computed from the beat-to-beat (RR) intervals of an ECG signal record, were suggested as risk measures for adverse cardiac events in [8]. This extensive study of 2,501 subjects is the first to investigate the association of HRV to risk of cardiac events for a large community-based population. The reduced HRV metrics of standard deviation of all normal RR intervals within a 2-hour window, very-low-frequency power, low-frequency power, and total power were significantly associated with increased risk of cardiac events. In the multivariate proportional hazards model, mean heart rate and all HRV metrics except the ratio between low-frequency and high-frequency power had statistically significant

association to subsequent cardiac events. HRV metrics were included in this investigation's feature vector as components $x_1 - x_{12}$.

Morphological variability (MV) was first introduced in [9] as a new measure of small differences in heartbeat morphology. This study suggested that MV could point to irregularities in the myocardium which could potentially lead to arrhythmia episodes. The MV metrics were compared to more traditional HRV metrics: standard deviation of the averages of beat-to-beat (RR) intervals of five minute windows (SDANN) and the ratio of power in the frequency spectrum between 0.04-0.15 Hz and 0.15-0.4 Hz (LF/HF). These metrics were extended to the morphological distance time-series for two MV metrics, MV-SDANN and MV-LF/HF. This investigation concluded that while MV-SDANN did not have an association with death, MV-LF/HF had a statistically significant correlation with death greater than that of the other HRV metrics. MV metrics are included in this investigation's feature vector as components $x_{13} - x_{14}$.

In [10], MV was evaluated on its ability to identify post-non-ST-elevation acute coronary syndrome (NSTEACS) patients with short-term risk of cardiac arrhythmias. This study found MV to have a strong association to arrhythmia episodes for patients following NSTEACS and established its use for short-term risk stratification in such patients. As another application, MV was utilized as a metric in a feature vector with a combination of ECG features including wavelet coefficients from a Daubechies 2 wavelet, normalized energy of beat segments, and RR interval metrics for heartbeat classification using a SVM approach in [11].

SVMs were first introduced in [12] as a statistical learning algorithm that optimizes the margin boundary of a separating linear-like hyperplane. Specifically, the margin between the decision boundary hyperplane and the support vectors, data points closest to the hyperplane, is maximized. The algorithm implements decision functions in a dual kernel representation, resulting in a significantly smaller number of dimensions than the original feature vector space. The resulting classification function dependent on a smaller training data set of the support vectors is able to minimize the maximum loss and represents a unique solution.

A SVM approach coupled with feature selection towards cardiac arrhythmia detection was proposed in [13]. The 13 morphological, spectral, and complexity features were computed from ECG signal records found in the PhysioNet repository. These features were then ranked using a combination of filter-type feature selection methods: correlation criterion, Fisher criterion, and mRMR criterion. The features were then evaluated on their classification

relevancy using bootstrap resampling to estimate SVM performance. The SVM classifier was used to classify the test set of ECG signal records for two detection decisions: ventricular fibrillation rhythm and shockable arrhythmia identification.

The use of SVMs for prediction of cardiac arrhythmias was proposed in [14]. The features in the study were obtained using singular value decomposition analysis of the spectral energy distribution of the ECG record signal. In this study, a binary SVM classifier was used with a Gaussian kernel function to classify the patient as either healthy or in predicted critical condition of an arrhythmia episode. The study also used continuous Gaussian white noise in the ECG record signals to emulate noisy data.

III. Methods

The ECG signal records were retrieved from PhysioBank of the PhysioNet repository, a collection of multi-parameter physiologic signals from a variety of types of patients from healthy subjects to those suffering from various health conditions, including a variety of arrhythmia states. The data has been de-identified [15]. These are the databases that were used in this study.

Database Name	Patients	Record Details
MIMIC II Version 2 Database (MIMIC) [16]	Patients in an intensive care unit (adults and neonates)	225 records used of a set with expert-reviewed severe arrhythmia alert annotations Long-term (days to weeks)
Creighton University Ventricular Tachyarrhythmia Database (CUDB) [17]	Patients who experienced sustained ventricular tachycardia, ventricular flutter, and ventricular fibrillation	31 records used of a set with reference annotations of notable events and episodes 8 minutes long
Fantasia Database [18]	Patients monitored while watching the movie, *Fantasia* 20 young (21 - 34 years old) 20 elderly (69 - 85 years old)	40 records with annotated heartbeats verified by inspection 120 minutes long
MIT-BIH Malignant Ventricular Arrhythmia Database (VFDB) [19]	Patients who experienced sustained ventricular tachycardia, ventricular flutter, and ventricular fibrillation	22 records with rhythm reference annotations 30 minutes long
MIT-BIH Supraventricular Arrhythmia Database (SVDB) [20]	Patients who experienced supraventricular arrhythmias	78 records with annotations 30 minutes long
MIT-BIH Normal Sinus Rhythm Database (NSRDB)	Patients in Arrhythmia Laboratory at Beth Israel Deaconess Medical Center with no significant arrhythmias	18 records with annotations Different parts of record used (effectively 36 records) Long-term (days)

TABLE I: Database, Patient, and Record Descriptions

The Fantasia, VFDB, and SVDB ECG records were used in their entirety while excerpts of the MIMIC, CUDB, and NSRDB ECG records were selected. For the MIMIC, the first, expert-validated severe arrhythmia event with at least five minutes of preceding ECG signal was noted, and the five-minute sample of the signal preceding the arrhythmia episode was used. For the CUDB, the first three minutes of each ECG signal record were utilized because most of the records did not contain episodes of ventricular flutter/fibrillation during this time; four records from the CUDB were later removed because they contained episodes of these arrhythmias within the three-minute sample. Finally, two 30-minute segments from records in the NSRDB were extracted from each ECG signal record and were treated as separate data points.

To extract the variability features, annotations of the beats in each record were required. A QRS detector from the PhysioToolkit was used to produce beat annotations for each record. The detector, optimized for sensitivity, was implemented in the MATLAB environment as the function, *gqrs* [21] [22]. Records in the Fantasia database were accompanied by beat annotations which were utilized in place of those produced by the detector [15] [18].

HRV metrics were calculated using PhysioNet's HRV toolkit, an open-source software for HRV analysis. The software input the beat annotations mentioned above, calculated the RR intervals, and extracted the normal sinus to normal sinus (NN) intervals. The frequency domain was calculated using the Lomb-Scargle periodogram, which was also included in the PhysioToolkit [15] [23]. The package calculated the HRV metrics shown in Table II.

Metric Name	Description
NN/RR	Ratio of NN intervals and total RR intervals
Time-Domain Measures	
AVNN (x_1)	Average of all NN intervals
SDNN (x_2)	Standard deviation of all NN intervals
SDANN (x_3)	Standard deviation of averages of NN intervals in all 5-minute segments
SDNNIDX (x_4)	Mean of standard deviations of NN intervals in all 5-minute segments
rMSSD (x_5)	Square root of mean of squares of differences between consecutive NN intervals
PNN50 (x_6)	Percentage of differences between consecutive NN intervals that greater than 50 milliseconds
Frequency-Domain Measures	
TOTPWR (x_7)	Total spectral power of NN intervals up to 0.04 Hz
ULF (x_8)	Total spectral power of NN intervals up to 0.003 Hz
VLF (x_9)	Total spectral power of NN intervals between 0.003 and 0.04 Hz
LF (x_{10})	Total spectral power of NN intervals between 0.04 and 0.15 Hz
HF (x_{11})	Total spectral power of all NN intervals between 0.15 and 0.4 Hz
LF/HF (x_{12})	Ratio of LF to HF

TABLE II: HRV Metrics

Morphological variability (MV) metrics were calculated by implementing the process described in [9] in the MATLAB environment. The ECG signal records were first pre-processed and filtered by the method described in [24]. Figure 1 depicts an example of a processed and filtered ECG signal record from the Fantasia database. Using the beat annotations, the records were divided into segments separated by each detected QRS complex.

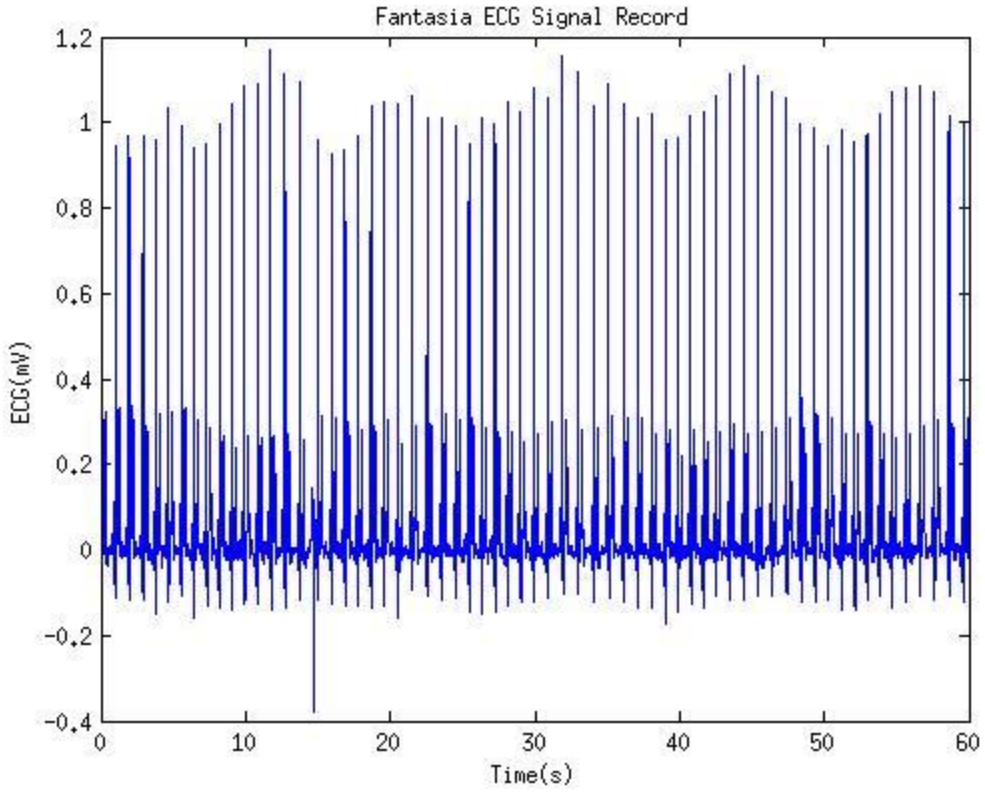

Figure 1. A 60-second sample of a processed and filtered Fantasia ECG signal record "f2y10" is shown [18]. The ECG signal record was retrieved from the PhysioBank physiological signal collection and processed using median filters of 200 ms and 600 ms as proposed in [24]. The graph was produced in the MATLAB environment [21].

```
%Find MD using dynamic time warping function (dtw)
marker = 1;
for i=1:N,
    sample = samp(i);
    sample2 = samp(i+1);
    if (sample > 525000)
        MD(N) = [];
        t(N) = [];
        break
    end
    if (sample2 > 525000)
        sample2 = 525000;
    end
    sig1 = horzcat(time(marker:sample), rsignal(marker:sample));
    s = sample + 1;
    sig2 = horzcat(time(s:sample2), rsignal(s:sample2));
    marker = sample;
    window = sample2 - sample;
    d = dtw(sig1,sig2,window);
    t(i) = time(sample);
    MD(i) = d;
end
```

Figure 2. A sample of the MATLAB code is shown. This section is responsible for calculating and storing the energy differences between consecutive heartbeats using the imported *dtw* function [25].

Using a dynamic time-warping implementation in the MATLAB environment, the cost of the optimized alignment path between the ECG segments was calculated for each consecutive pair of ECG segments, resulting in a new time-series of energy differences [25]. The *sig1*, *sig2*, and *window* variables contains the consecutive heartbeat signals and length of the difference the signals respectively. The *dtw* function used these variables in order to align similar parts of the two heartbeat signals and compute the resulting energy difference between the two aligned signals. The graph below illustrates this process.

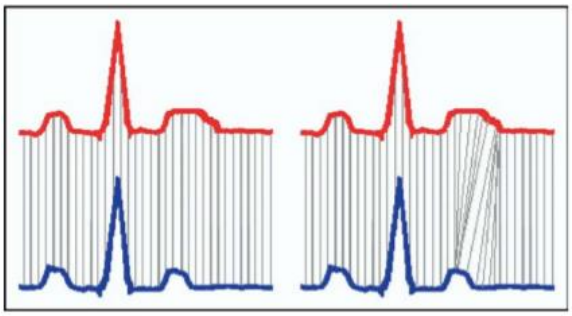

Figure 3. Here is an example of the results of dynamic time warping on two heartbeat signals. The second pair of signals is aligned so that similar sections of the heartbeat may be compared. If the differences between these aligned signals are still large, then this could indicate a higher risk of suffering from a future arrhythmia.

The morphological distance time-series was then smoothed using a median filter of length 8 as proposed in [9], and an example of this smoothed energy differences signal corresponding to the same ECG signal record used in Figure 1 is shown in Figure 2. The power spectral density of the signal was then estimated with the Lomb-Scargle periodogram using the MATLAB function, *plomb* [21] [26]. The MV metrics were computed from the power spectral density of the signal. Table III shows the MV metrics calculated.

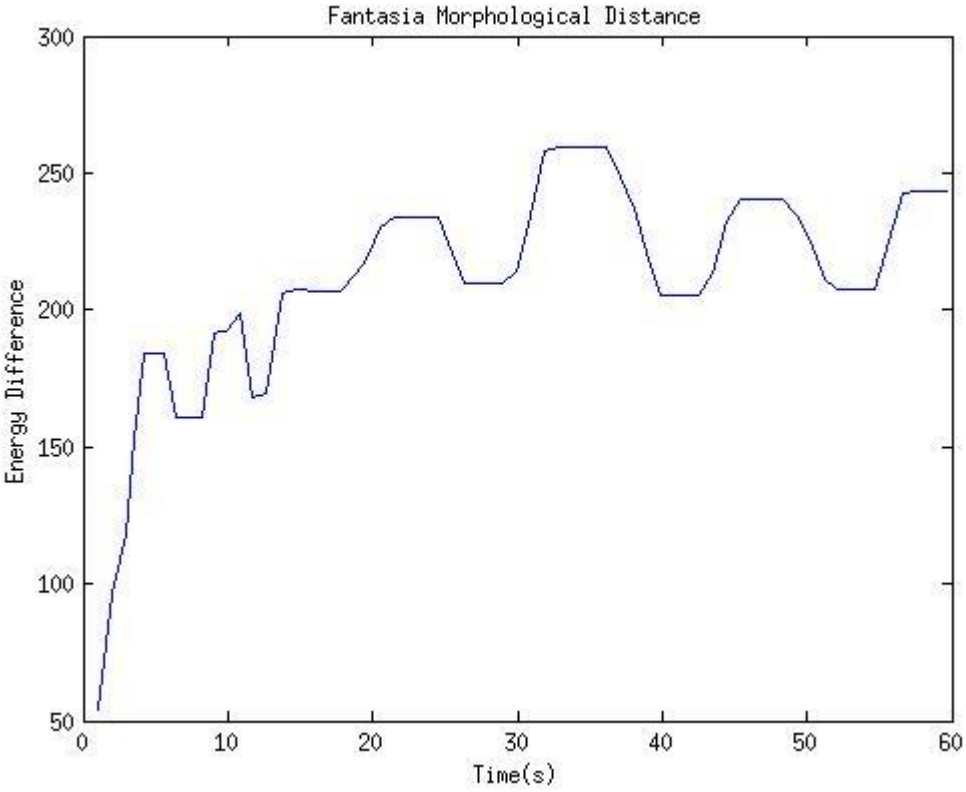

Figure 4. A 60-second sample of the smoothed energy difference time-series for Fantasia record "f2y10" is shown [18]. The energy differences, defined as the optimal alignment path between each pair of consecutive ECG segments, were calculated using the dynamic time-warping implementation [25]. The resulting time-series was smoothed using the median filter proposed in [9]. The graph was produced in the MATLAB environment [21].

Metric Name	**Description**
MV (x_{13})	Energy in MD time-series power spectral density between 0.30-0.55 Hz
MV-LF/HF (x_{14})	Ratio of LF/HF* in MD time-series *LF: 0.04-0.15 Hz, HF: 0.15-0.4 Hz

TABLE III: MV Metrics

The metrics NN/RR, SDANN, and SDNNIDX were not included in the final feature vector. NN/RR is used primarily as an indicator of data reliability and does not represent an ECG variability measure. SDANN and SDNNIDX are normally used for long-term records and would be unreliable for the records used in this study. The feature vector was also subjected to forward sequential feature selection with 10-fold cross-validation using the MATLAB function, *sequentialfs*. The function selected features until the local minimum of the misclassification rate was reached as defined by the criterion value, resulting in a feature subset with higher classification performance than the original feature set of all metrics [21].

The finalized feature vector was partitioned into training and test datasets. The class label vectors were defined as follows. The Fantasia data set and both data sets of the NSRDB were labeled "healthy" as they contained no significant arrhythmic episodes. The VFDB, SVDB, CUDB, and MIMIC data sets were labeled "unhealthy" as they contained episodes of various types of arrhythmia episodes.

With the finalized feature vector, the SVM classifiers were trained using the MATLAB function, *fitcsvm*, with a linear kernel function for two-class learning. The MATLAB function, *predict*, was used to predict class labels for the test datasets, which returned both the predicted labels and class likelihood measures. Receiver operating characteristic (ROC) curves were created using the MATLAB function, *perfcurve*, and the area under the curve (AUC) was computed in order to evaluate the performance of the SVM classifiers. A better performing classifier would have an AUC value closer to 1 while a lesser performing classifier would have an AUC value closer to 0.5 [21].

Confusion matrices were also created to evaluate each SVM classifier using the MATLAB function, *confusionmat* [21]. Extracting the true and false class prediction values from the confusion matrices, sensitivity (SE), specificity (SP), positive predictive value (PPV), F-Score, and Youden's index were all computed for each SVM classifier as defined by the methods in [27]. PPV and F-Score were used in this study due to their nature of emphasis on correct classification of the positive class ("unhealthy"). Youden's index, as an evaluation of a classifier's avoidance of failure, was chosen in conjunction with the goals of this study: to avoid failure in prediction of arrhythmia episodes. All classifier evaluation metrics ultimately improved in conjunction with the overall classification accuracy.

IV. Results

A total of 9 SVM classifiers were constructed and trained in this study. Table IV describes each classifier and includes their corresponding classification performance evaluation metrics. The first 2 classifiers used all 436 records collected, with the training and test datasets drawing equal data points from each database and the odd ECG signal records going towards the training data set. For the remainder of the classifiers, the 4 aforementioned CUDB records containing arrhythmia episodes were removed. For SVM Classifiers 6 and 7, the SVDB and VFDB records were not utilized because they contained arrhythmia episodes within the signal recordings. In SVM Classifiers 8 and 9, these SVDB and VFDB records were instead used exclusively in the training dataset, thereby maintaining the nature of the classifiers as predictive models rather than detection models.

Sequential forward feature selection was used to select the feature subsets used in SVM Classifiers 2, 4, 7, and 9. In SVM Classifier 5, the same feature subset as SVM Classifier 2 was used for the training and test datasets, resulting in slightly better performance across all evaluation measures. Overall, each classifier with a reduced feature subset improved in performance across all evaluation metrics. In the case of SVM Classifiers 6 and 7 of which the SVDB and VFDB records were omitted, the performance increase due to the feature selection was not as substantial.

SVM Classifier	Details	Features	AUC	Accuracy	SE	SP	PPV	F-Score	Youden's Index
1	Original dataset	All	0.410	66.82%	80.45%	2.63%	79.56%	80%	0.821
2	Feature selection	AVNN, SDNN, rMSSD, MV-LF/HF	0.952	89.86%	97.77%	52.63%	90.67%	94.09%	1.494
3	4 CUDB records removed	All	0.413	67.61%	81.71%	2.63%	79.44%	80.56%	0.833
4	4 CUDB records removed with feature selection	AVNN, rMSSD, pNN50, MV-LF/HF	0.950	89.20%	97.71%	50%	90%	93.70%	1.467
5	4 CUDB records removed with features used in SVM Model 2	AVNN, SDNN, rMSSD, MV-LF/HF	0.951	89.67%	97.71%	52.63%	90.48%	93.96%	1.493
6	4 CUDB records, SVDB, and VFDB removed	All	0.928	91.52%	96.06%	76.32%	93.13%	94.57%	1.714
7	4 CUDB records, SVDB, and VFDB removed with feature selection	AVNN, SDNN, rMSSD, LF/HF, MV, MV-LF/HF	0.929	92.12%	96.06%	78.95%	93.85%	94.94%	1.740
8	4 CUDB records removed, SVDB & VFDB moved to training dataset	All	0.213	58.79%	74.02%	7.89%	72.87%	73.44%	0.809
9	4 CUDB records removed, SVDB & VFDB moved to training dataset with feature selection	AVNN, rMSSD, MV-LF/HF	0.945	86.06%	97.64%	47.37%	86.11%	91.51%	1.440

TABLE IV: SVM Classifiers and Classification Evaluation Metrics

The first model created, SVM Classifier 1, with the complete, original dataset and no feature selection, had an AUC value of 0.410 and accuracy of 66.82%. In comparison, SVM Classifier 7 had the most promising performance out of all the classifiers, reporting the highest classification accuracy, SP, PPV, F-score, and Youden's index of 92.12%, 78.95%, 93.13%, 94.57%, and 1.740 respectively. It also reported high values for sensitivity and AUC as well as the lowest number of false negatives and the highest number of true negatives as shown in Table V. This could be attributed to the reduced training and test dataset, resulting in a smaller percentage of true positives to total positives. Consequently, this would have also influenced the AUC metric, which is dependent on the true positive and false positive rates [27]. Figure 3 illustrates the large difference in the ROC curves of SVM Classifiers 1 & 7.

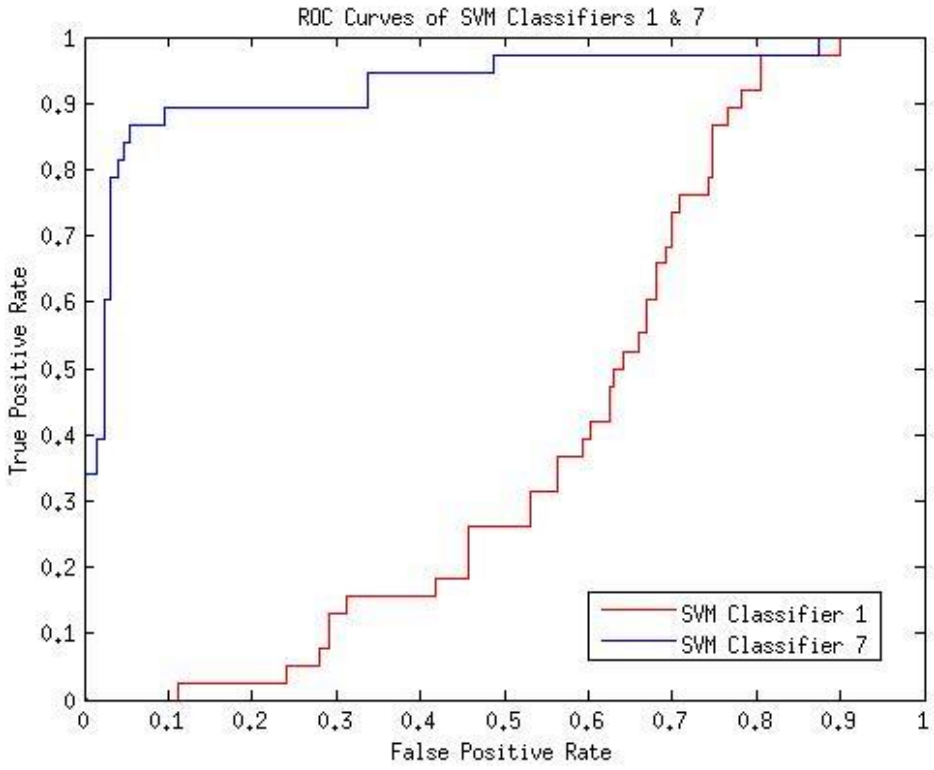

Figure 5. The graph shows the comparison of the ROC curves and AUC values between SVM Classifiers 1 & 7. The stronger classification model, SVM Classifier7, underwent sequential forward feature selection and had a final feature vector containing AVNN, SDNN, rMSSD, LF/HF, MV, and MV-LF/HF. It had an overall classification accuracy of 92.12% for 165 ECG signal records and an AUC value of 0.929. The graph was produced in the MATLAB environment using the function, *perfcurve* [21].

As shown in Table V for the class label predictions of SVM Classifier 9, with feature selection and the transfer of all the SVDB and VFDB records to the training data set, 2 more positive cases were classified correctly in comparison to SVM Classifier 7 at the expense of 12 additional misclassified cases as false positives. SVM Classifier 9 had higher scores for the evaluation metrics, AUC and SE, but lower scores for the remaining evaluation metrics: classification accuracy, SP, PPV, F-Score, and Youden's index.

SVM Classifier	True Positives*	False Positives	False Negatives	True Negatives
1	144	37	35	1
2	175	18	4	20
3	143	37	32	1
4	171	19	4	19
5	171	18	4	20
6	122	9	5	29
7	122	8	5	30
8	94	35	33	3
9	124	20	3	18

*The positive class was defined as a patient experiencing a future arrhythmia episode; the negative class was defined as a healthy patient.

TABLE V: Class Predictions of SVM Classifiers

V. Discussion

The majority of misclassified patients had the class prediction of false positive, indicating that the classifiers were prone to identify the patients as positive or at increased risk for future arrhythmia episodes. Misclassified healthy patients could be detrimental to the ICD recommendation process, but beneficial to short-term arrhythmia episode prediction. On balance, a false alarm would outweigh a missed alarm in terms of maximizing patient survival, accounting for the costs of responding to a false alarm. However, the positive and negative classes were balanced for the class prediction thresholds of each classifier.

In this study, MV-LF/HF was found to have the greatest impact on the reduction of the misclassification rate of the SVM model during the sequential forward feature selection. Thus, we may conclude that for this particular data set of a variety of different patient conditions, the MV-LF/HF metric served as the strongest indicator of increased risk of a future arrhythmia episode. The results of the feature selection also revealed the nature of this classification

problem. In particular, the features chosen by the selection algorithm differed as the training data sets changed. Although the features, AVNN, rMSSD, and MV-LF/HF, were chosen repeatedly for each classifier with feature selection, the others, SDNN, LF/HF, and pNN50, were only characteristic of some of the classifiers proposed. Thus, the different data sets used in each classifier benefited from specific combinations of the metrics that lowered the misclassification rate of the respective classifier.

VI. Contributions

Previously, MV-LF/HF was introduced as a risk variable in [9], and was found to have the strongest association with death in comparison to other ECG variability metrics such as SDANN or LF/HF. MV-LF/HF has not been used beforehand as a metric for a classifier in arrhythmia risk prediction. In addition, the specific feature subsets with a combination of the metrics, AVNN, SDNN, rMSSD, pNN50, LF/HF, MV, and MV-LF/HF, have not been previously proposed and used for cardiac arrhythmia prediction.

This study also validated and extended the conclusions presented in [10], which concluded that high MV values were strongly associated to arrhythmias, defined as more than eight beats of ventricular tachycardia or cardiac pause of over three seconds. MV was used as a feature in SVM Classifier 7, the strongest performing classifier in terms of overall accuracy. This result of the feature selection supports the conclusions in [10] and suggests that MV may contribute to prediction of patients for short-term risk of arrhythmia episodes using a SVM classification approach. However, the risk stratification model in the aforementioned study used patients who had recently suffered from non-ST-elevation acute coronary syndrome while this study aimed to investigate both healthy patients and those who later suffered an arrhythmia episode.

With the highest classification accuracy of 92.12% of 165 ECG records, SVM Classifier 7 outperformed the SVM classification model proposed in [14], which had a classification accuracy of 89% from a clean data set of 35 patient records. The feature vector with AVNN, SDNN, rMSSD, LF/HF, MV, and MV-LF/HF had a stronger performance than the feature vector used in the study mentioned. In addition, the larger data set used in this investigation serves as an extended evaluation of the viability of using a SVM classifier for cardiac arrhythmia prediction as a step from the smaller data set used by the classifier in the aforementioned study.

VII. Conclusion

This study proposed a number of SVM classifiers and evaluated their ability to predict patient short-term risk of a future arrhythmia episode using HRV and MV metrics. The classifiers were trained and tested on various databases in the PhysioBank archive which contained ECG signal recordings from both healthy patients and those who suffered a variety of arrhythmia episodes. Using sequential forward feature selection, features were chosen from the original 15 computed metrics based on their ability to reduce predictive error of the classifiers. The SVM classifiers with the reduced feature subsets improved in classification performance across all computed classification performance statistics [27].

The performance of SVM Classifiers 7 & 9 suggests that the proposed SVM approach with sequential forward feature selection is a viable, assistive tool for short-term prediction of future arrhythmia episodes. This predictive classification model could be applied in ICD recommendation and decrease response time for patients at risk of a future arrhythmia episode in a short-term time window. This study also validated the viability of MV-LF/HF as a feature in SVM classification models for short-term arrhythmia episode risk analysis.

The training and test data sets utilized were unbalanced towards the class of patients at risk of arrhythmic episodes. In order to counteract this imbalance, more data points from healthy patients would have to be included. This could also reduce the number of false positives, as the classifier would have more support feature vectors to maximize the margin boundary from, optimizing the decision hyperplane distance from the nearest healthy data point. The classifier could also be modified to favor classification of the positive class, in order to maximize the reduction of the number of missed alarms of arrhythmia episodes.

Symmetry could be further advanced across the databases used to construct and test the proposed classifiers. The ECG signal records in each database differ in various characteristics such as the sampling rate, length of window before a cardiac arrhythmia episode, and analog-to-digital converter resolution. Resampling and signal scaling techniques could be applied in order to achieve a uniform sampling rate across all ECG signal records, at the cost of reduced amplitude resolution. With a uniform sampling rate, the length of the time window would contain an equal number of samples per ECG signal record and the time window before an arrhythmia event could be kept constant.

Potential avenues for future work on this classification problem exist. Further evaluation

of the classification approach and feature vector proposed in this paper on other datasets could be completed. This study could be extended to include longer-term prediction of risk of arrhythmia episodes. With larger time windows, ICD recommendation and response time to arrhythmia episodes would be improved even further. The classification models could also account for a greater number of classes to differentiate between different types of arrhythmias and adverse cardiac events.

A user-interactive website is also currently under development. It will educate people on the problem using example patients and explain the proposed SVM approach. It will also allow users to upload their ECG signal and associated heartbeat intervals as text files. After calculating the user's HRV and MV metrics, the previously-constructed SVM classifier 7 will be used to predict the user's short-term risk of suffering an arrhythmia episode. The user can also choose to contribute this data to this project in order to improve the current classification models.

VIII. Acknowledgements

I would like to thank George Washington University for allowing me to use its facilities at the Science and Engineering Hall. I would also like to thank Professor Howie Huang for mentoring me as a research intern, Professor Matthew Kay for guiding me in the project methodology, and graduate student Hang Liu for his help with the MATLAB code. Finally, I am grateful for Dr. Peter Gabor and Dr. Shane Torbert for their support throughout the research process as my research lab directors.

References

[1] Cleveland Clinic, 'Sudden Cardiac Death (Sudden Cardiac Arrest)', 2015. [Online]. Available: http://my.clevelandclinic.org/services/heart/disorders/arrhythmia/sudden-cardiac-death. [Accessed: 30- Jun- 2015].

[2] NIH - National Heart, Lung, and Blood Institute, 'What Is an Arrhythmia?', 2011. [Online]. Available: http://www.nhlbi.nih.gov/health/health-topics/topics/arr. [Accessed: 2- Jul- 2015].

[3] NIH - National Heart, Lung, and Blood Institute, 'What Is an Implantable Cardioverter Defibrillator?', 2011. [Online]. Available: http://www.nhlbi.nih.gov/health/health-topics/topics/icd. [Accessed: 2- Jul- 2015].

[4] NIH - National Heart, Lung, and Blood Institute, 'Who Needs an Implantable Cardioverter Defibrillator?', 2011. [Online]. Available: http://www.nhlbi.nih.gov/health/health-topics/topics/icd/whoneeds. [Accessed: 2- Jul- 2015].

[5] A. Kadish and J. Goldberger, 'Selecting Patients for ICD Implantation', *JAMA*, vol. 305, no. 1, pp. 91-92, 2011.

[6] D. Grady, 'Many Defibrillators Implanted Unnecessarily, Study Says', *The New York Times*, 2011. [Online]. Available: http://www.nytimes.com/2011/01/05/health/05device.html. [Accessed: 8- Jul- 2015].

[7] Resuscitationcentral.com, 'Early Defibrillation Programs', 2015. [Online]. Available: http://www.resuscitationcentral.com/defibrillation/early-defibrillation-sca-chain-of-survival/. [Accessed: 8- Jul- 2015].

[8] H. Tsuji, M. Larson, F. Venditti, E. Manders, J. Evans, C. Feldman and D. Levy, 'Impact of reduced heart rate variability on risk for cardiac events: the Framingham heart study', *Circulation*, vol. 94, no. 11, pp. 2850-2855, 1996.

[9] Z. Syed, B. Scirica, C. Stultz and J. Guttag, 'Risk-stratification following acute coronary syndromes using a novel electrocardiographic technique to measure variability in morphology', *Computers in Cardiology*, pp. 13 - 16, 2008.

[10] Z. Syed, B. Scirica, C. Stultz and J. Guttag, 'Electrocardiographic prediction of arrhythmias',*Computers in Cardiology*, pp. 565 - 567, 2009.

[11] J. Wiens and J. Guttag, 'Active learning applied to patient-adaptive heartbeat classification',*Advances in Neural Information Processing Systems*, vol. 23, pp. 2442 - 2450, 2010.

[12] B. Boser, I. Guyon and V. Vapnik, 'A training algorithm for optimal margin classifiers',*Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*, pp. 144 - 152, 1992.

[13] F. Alonso-Atienza, E. Morgado, L. Fernandez-Martinez, A. Garcia-Alberola and J. Rojo-Alvarez, 'Detection of Life-Threatening Arrhythmias Using Feature Selection and Support Vector Machines', *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 3, pp. 832 - 840, 2014.

[14] H. Namarvar and A. Shahidi, 'Cardiac arrhythmias predictive detection methods with wavelet-SVD analysis and support vector machines', *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 365 - 368, 2004.

[15] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. Ivanov, R. Mark, J. Mietus, G. Moody, C. Peng and H. Stanley, 'PhysioBank, PhysioToolkit, and PhysioNet : components of a new research resource for complex physiologic signals', *Circulation*, vol. 101, no. 23, pp. e215-e220, 2000.

[16] M. Saeed, M. Villarroel, A. Reisner, G. Clifford, L. Lehman, G. Moody, T. Heldt, T. Kyaw, B. Moody and R. Mark, 'Multiparameter Intelligent Monitoring in Intensive Care II: A public-access intensive care unit database*', *Critical Care Medicine*, vol. 39, no. 5, pp. 952-960, 2011.

[17] F. Noelle, F. Badura, J. Catlett, R. Bowser and M. Sketch, 'CREI-GARD, a new concept in computerized arrhythmia monitoring systems', *Computers in Cardiology*, vol. 13, pp. 515 - 518, 1987.

[18] N. Iyengar, C. Peng, R. Morin, A. Goldberger and L. Lipsitz, 'Age-related alterations in the fractal scaling of cardiac interbeat interval dynamics', *The American Physiological Society*, vol. 271, pp. 1078 - 1084, 1996.

[19] S. Greenwald, 'Development and analysis of a ventricular fibrillation detector', M.S., MIT Dept. of Electrical Engineering and Computer Science, 1986.

[20] S. Greenwald, 'Improved detection and classification of arrhythmias in noise-corrupted electrocardiograms using contextual information', Ph.D, Harvard-MIT Division of Health Science and Technology, 1990.

[21] *MATLAB*. Natick, Massachusetts, United States: The MathWorks Inc., 2012.

[22] I. Silva and G. Moody, 'An open-source toolbox for analysing and processing PhysioNet

databases in MATLAB and Octave', *Journal of Open Research Software*, vol. 2, no. 1, 2014.

[23] G. Moody, 'Spectral analysis of heart rate without resampling', *Proceedings of Computers in Cardiology Conference*, pp. 715 - 718, 1993.

[24] P. deChazal, M. O'Dwyer and R. Reilly, 'Automatic classification of heartbeats using ECG morphology and heartbeat interval features', *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 7, pp. 1196 - 1206, 2004.

[25] Q. Wang, *Dynamic Time Warping (DTW)*. The MathWorks Inc., 2014.

[26] N. Lomb, 'Least-squares frequency analysis of unequally spaced data', *Astrophys Space Sci*, vol. 39, no. 2, pp. 447-462, 1976.

[27] M. Sokolova, N. Japkowicz and S. Szpakowicz, 'Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation', *Advances in Artificial Intelligence*, vol. 4304, pp. 1015 - 1021, 2006.