

# Cardiac Arrhythmia Risk Stratification using Multiple Features

Lichu Zhao, Matt Wiggins, George Vachtsevanos and Brian Litt\*

School of Electrical and Computer Engineering  
Georgia Institute of Technology  
777 Atlantic Drive, Atlanta, GA 30332-0250  
USA

\*Department of Neurology  
University of Pennsylvania  
3400 Spruce St., Philadelphia, PA 19104  
USA

## ABSTRACT

*Sudden cardiac death (SCD) accounts for approximately 300,000 deaths in the U.S. annually. Most sudden deaths are due to ventricular tachycardia (VT) or ventricular fibrillation (VF), but unfortunately, most patients with sudden death die before reaching the hospital. Because the number of patients surviving the hospitalization with sustained ventricular arrhythmias represents only a fraction of the patients who may be at risk for an arrhythmic event, research has focused in recent years on developing risk stratification methods to identify these patients that may be at the highest risk.*

*In this paper, a risk stratification algorithm based on multiple features is proposed. First, the available data set is clipped at 30 second segments. Second, the data set is preprocessed to eliminate the baseline drift and powerline noise. Next, the best feature set is selected for risk stratification following the following steps: 1) Potential feature set selection, where eight potential features are selected based on the knowledge of ECG signals with cardiac arrhythmia present; 2) Feature normalization, where the best normalization method is selected for each feature among the proposed six normalization methods respectively; 3) A similarity measurement between any two different features is defined; 4) Best two-feature set selection, where the two features are least correlated. Finally, a K Nearest Neighbor (KNN) classifier is explored for risk stratification, where half the data (randomly selected) is used for training and the whole data set for testing. The proposed algorithm achieves 89.9% accuracy for risk stratification of the whole data set. The proposed risk stratification method doesn't require beat classification or QRS detection while there is no requirement for prior knowledge about patient data.*

## 1. Introduction

Sudden cardiac death (SCD) accounts for approximately 300,000 deaths in the U.S. annually. Most sudden deaths are due to ventricular tachycardia (VT) or ventricular fibrillation (VF), but unfortunately, most patients with sudden death die before reaching the hospital. Because the number of patients surviving to hospitalization with sustained ventricular arrhythmias represents only a

fraction of the patients who may be at risk for an arrhythmic event, a tremendous amount of research has focused on developing risk stratification methods to identify the patient at highest risk.

Many articles published describe risk factors for sudden death in patients with previous MIs. The risk of sudden death after a MI is approximately 5% per year for at least three to five years after an infarction. The risk is proportional to the degree of left ventricular dysfunction, with a left ventricular ejection fraction below 40% being the best predictor of high risk in the long-term. Asymptomatic ventricular arrhythmias including premature ventricular contractions (PVCs) and non-sustained VT (more than three PVC beats lasting less than 30 s) are also valuable for predicting outcome in patients with coronary artery disease. Isolated ventricular ectopy is common after MI, and up to 80% of patients will have some ectopy present on a 24-h ambulatory (Holter) monitor.

Usually, the risk stratification methods in literature are either based on patients' medical information, such as left ventricular ejection fraction or beat classification (PVC beats/hour, non-sustained VT), or based on Heart Rate Variability (HRV) analysis. In our risk stratification method, no beat classification or QRS detection is needed. Also, there is no requirement for prior knowledge about patients.

In this paper, a risk stratification algorithm based on multiple features is proposed. First, generate the data set, which was clipped 30 seconds segment from the MIT Malignant Ventricular Arrhythmia Database and AHA Database. Second, preprocess the data set, which eliminate the baseline drift and powerline noise. Next, select the best feature set for risk stratification. The best feature selection scheme is as follows: 1) Potential feature set selection, where eight potential features are selected based on the knowledge of ECG signals with cardiac arrhythmia; 2) Feature normalization, where the best normalization method is selected respectively for each feature among the proposed six normalization methods; 3) Similarity measurement between any two different features; 4) Best two-feature set selection, where the two features are least correlated. Finally, KNN classifier is explored for risk stratification, where half data (randomly selected) were used for training and the whole data set was used for testing. The proposed algorithm achieved

89.9% accuracy for risk stratification of the whole data set. The proposed risk stratification method also has no need for beat classification or QRS detection meanwhile there is no requirement for prior knowledge about patients.

## 2. Methodology

### 2.1. Data Set Generation

67 high-risk records are clipped several seconds before ONSET of VF/VT from the MIT Malignant Ventricular Arrhythmia Database and the AHA Database, and 60 low-risk records are clipped randomly from 10 non-Ventricular Arrhythmia persons from the AHA Database. The data segments are 30 seconds in length with a sampling frequency of 250Hz. If the data segment has a lot of noise, it will be discarded and regenerated. High pass filtering is used to remove baseline drift and a 60 Hz notch filter to remove power-line noise.

### 2.2. Processing

#### 2.2.1. Feature Extraction

Eight potential features are selected based on our knowledge about cardiac arrhythmia and ECG signal. They are Energy, Curve Length, Nonlinear Energy, Spectral Entropy, Power in band [3 6]Hz, Fractal Dimension, AR Error and Wavelet Accumulate Energy.

#### 2.2.2. Potential Feature Set Selection

The feature normalization method based on Fisher’s Discriminant Ratio is used to eliminate the ill-conditioned problem caused by different magnitude of features. [7] Two features could be thought of as two individual variables. The correlation coefficient was selected as the similarity measure between two features. The higher the correlation coefficient, the more correlated between two features. Table 1 shows the similarity measure between different features.

Table 1 Feature Similarity between Features

	E	CL	FD	AR	P36	SE	NE	WAE
E	1	0.5444	0.5481	0.0615	0.1273	0.9929	0.4753	0.2248
CL	0.5444	1	0.9992	0.8033	<b>0.0691</b>	0.4849	0.8806	0.8249
FD	0.5481	0.9992	1	0.7979	0.0693	0.4882	0.875	0.8211
AR	<b>0.0615</b>	0.8033	0.7979	1	0.1727	<b>0.008</b>	0.6869	0.8447
P36	0.1273	<b>0.0691</b>	<b>0.0693</b>	0.1727	1	0.136	<b>0.2969</b>	0.3196
SE	0.9929	0.4849	0.4882	<b>0.008</b>	0.136	1	0.4079	<b>0.1602</b>
NE	0.4753	0.8806	0.875	0.6869	0.2969	0.4079	1	0.861
WAE	0.2248	0.8249	0.8211	0.8447	0.3196	0.1602	0.861	1

Bolded values indicate the least correlated features. From the table, curve length and fractal dimension capture the same information of the signal. Under this condition, only Curve length is kept because of the easily computation. Also, energy and spectral entropy have high correlation

(correlation coefficient equal to 0.9929). The selected potential best two-feature sets are:

Set1: Energy and AR Error

Set2: Nonlinear Energy and Power in band [3 6]

Set3: Curve Length and Power in band [3 6]

Set4: Wavelet Accumulate Energy and Spectral Entropy

### 2.2.3. K-Nearest Neighbor Classifier

The K Nearest Neighbor (kNN) is a very intuitive method that classifies unlabeled examples based on their similarity to examples in the training set. For a given unlabeled example  $x_u$ , find the k “closest” labeled examples in the training data set and assign  $x_u$  to the class that appears most frequently within the k-subset. The kNN only requires

- An integer k
- A set of labeled examples (training data)
- A metric to measure “closeness”

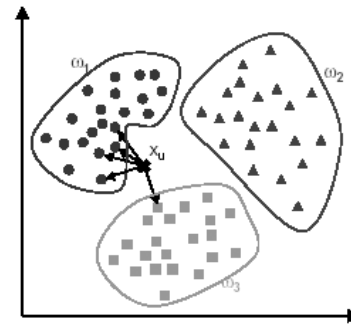


Figure 1 KNN Classifier

The example shown in Figure 1 has three classes, uses the Euclidean distance as closeness measure and selects  $k=5$ . The goal is to find a class label for the unknown example  $x_u$ . Of the 5 closest neighbors, 4 belong to  $\omega_1$ , and 1 belongs to  $\omega_3$ , so  $x_u$  is assigned to  $\omega_1$ , the class with the highest confidence.

### 2.3. Performance Analysis

The confusion matrix and performance metric (Sensitivity, Specificity, Positive Prediction Value, Negative Prediction Value and Classification Accuracy) is defined in Table 2 and Table 3 respectively.

Table 2 Confusion Matrix

	Arrhythmia	Non-Arrhythmia
Arrhythmia	NTP	NFN
Non-Arrhythmia	NFP	NTN
N=Total Number	NTP+NFP	NFN+NTN

Where,

*NTP=number of true positives*

*NFN=number of false negatives*

*NFP=number of false positives*

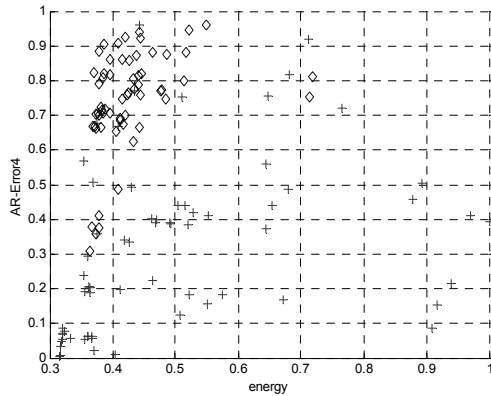
*NTN=number of true negatives*

*N=total number of records classified*

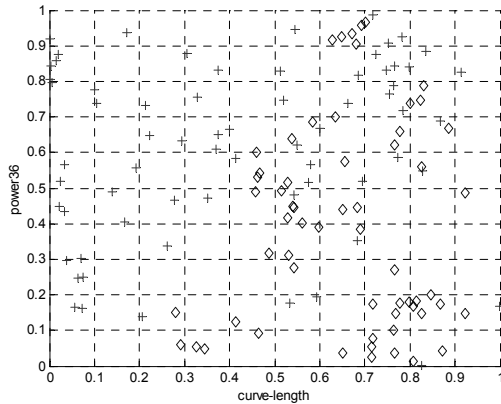
Table 3 Performance Metric Definition

Name	Definition
Sensitivity	$Sen = \frac{NTP}{NTP + NFN}$
Specificity	$Spec = \frac{NTN}{NTN + NFP}$
Positive Prediction Value (PPV)	$PPV = \frac{NTP}{NTP + NFP}$
Negative Prediction Value (NPV)	$NPV = \frac{NTN}{NTN + NFN}$
Classification Accuracy	$Acc = \frac{NTP + NTN}{N}$

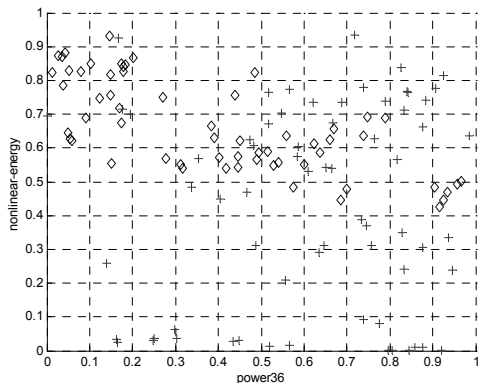
### 3. Experimental Results



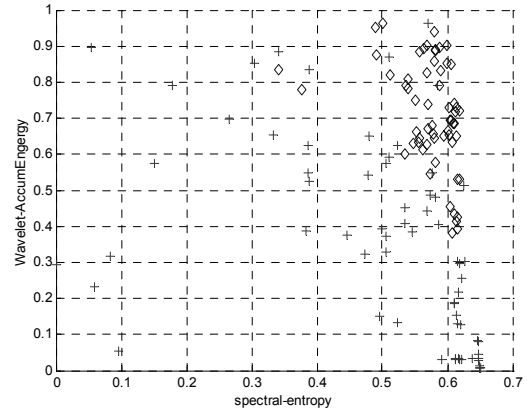
(a)



(b)



(c)



(d)

Figure 2 2D feature distribution: red cross stands for high risk patients and blue diamond stands for low risk people.

Figure 2 (a)~(d) shows 2D feature space distribution of the whole data set for feature Set1, Set2, Set3 and Set4. KNN Classifier is applied to verify the hypotheses. From the distribution of (a)~(d), it is easy to conclude that Set1 and Set4 are better than Set2 and Set3

Half of high-risk patients and half of the normal persons were randomly selected as the training data set and the whole data set as the testing data set. The feature vector was fed into the KNN classifier. In order to eliminate the influence of training set selection, the experimental results were averaged by applying KNN 50 times. Table 4 lists the performance measures for the four feature sets. Set1 and Set4 outperform the other two feature sets. But it is difficult to tell whether Set1 is better than Set4. For the two feature sets, the classification accuracy is almost 90% and positive prediction value is 85%. Note that the Number of False Negatives (NFN) is not good enough (around 10 out of 67). This issue will be addressed in future work.

Table 4 Performance Measures for the Four Feature Set

	Set1	Set2	Set3	Set4
TP	57.24	53.50	51.04	57.64
FP	4.02	8.06	7.46	3.38
TN	55.98	51.94	52.54	56.62
FN	9.76	13.50	15.96	9.36
Sensitivity	0.8543	0.7985	0.7618	0.8603
Specificity	0.9330	0.8657	0.8757	0.9437
PPV	0.9344	0.8691	0.8725	0.9446
NPV	0.8515	0.7937	0.7670	0.8581
Classification Accuracy	0.8915	0.8302	0.8156	0.8997

### 4. Discussion and Conclusion

An ECG data-driven risk stratification algorithm is proposed and studied. Also, a methodology for feature selection, feature normalization and best feature set

selection is proposed using quantitative measures. The methodology is not specific to this application and could be generalized to other classification problems. The experimental results show that feature set chosen gives good class separability under the performance metric defined. Using KNN classifier where a randomly selected half of data was used for training and the whole data for testing, the proposed algorithm could achieve 89.9% accuracy for risk stratification of the whole data set.

Only two-feature sets are explored here. Further research is required to study feature sets with more than two features. Also, soft output (risk probability for each patient) rather than hard output will be explored in the future work. It is also worthy combining HRV signal together for patients' risk stratification.

### 5. Acknowledgement

We are grateful for the help and co-operation of Dr. David Callans and Dr. Edward Gerstenfeld of University of Pennsylvania. This research is currently funded by the Whitaker Foundation, the DANA Foundation and the National Institute of Health (grant #5R01NS041811-03).

### References

- [1]. Patrick J. Welch, Richard L. Page and Mohamed H. Hamdan, "Management of Ventricular Arrhythmias", *Journal of the American College of Cardiology*, Vol 34, No 3, 1999, p621-630
- [2]. James J. Bailey, Alan S. Berson, Harry Handelsman and Morrison Hodges, "Utility of Current risk Stratification Tests for Predicting Major Arrhythmic Events After Myocardial Infarction", *Journal of the American College of Cardiology*, Vol.38, No.7, 2001, p1902-1911
- [3]. C.W. Therrien, "Decision, estimation and classification", John Wiley & Sons, 1989
- [4]. Carl J. Huberty, *Applied Discriminant Analysis*. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics Section. John Wiley & Sons, 1994
- [5]. Pabitra Mitra, C.A.Murthy and Sankar K. Pal, "Unsupervised Feature Selection using Feature Similarity", *IEEE Trans on Pattern Analysis and Machine Intelligence*, Vol 24, No.3, p301-312, March 2002
- [6]. Matt Wiggins, Lichu Zhao, George Vachtsevanos and Brian Litt, "Non-Invasive, Cardiac Risk Stratification Using Wavelet Coefficients", *WSEAS Transaction on Computers*, p720-722, Vol2(3), 2003
- [7]. Lichu Zhao, Matt Wiggins, George Vachtsevanos and Brian Litt, "Feature Normalization For Cardiac Arrhythmia Risk Stratification", The 6th IASTED International Conference on Signal and Image Processing (SIP 2004), Honolulu, Hawaii, USA, August 23-25, 2004