**Classification of Alcoholics and Non-Alcoholics using Heart Rate Variability Analysis on ECG Signals**

**Akarsh N Kolekar, Apoorv Vatsal** and **Rakshith Vishwanatha**

**Guided by – Dr. B. Niranjana Krupa**

**Abstract:** Several medical studies reveal alcohol consumption has pronounced effects on the heart rate variability (HRV) of the consumer. In this article, electrocardiogram (ECG) samples of chronically alcoholic subjects and normative subjects are collected for HRV analysis and feature extraction. The features extracted are fed to machine learning algorithms to enable the algorithms to classify new subjects into alcoholic or normative classes ~~on collecting the subject’s ECG~~. For this classification problem, Support Vector Machines and Extreme Learning Machines have been trained, and their performance has been compared. While time domain, frequency domain and non-linear features are generally extracted from ECG signals for HRV analysis, in this study a new set of features obtained from Autoregressive Modelling (using Exogenous Inputs) have also been used to improve the accuracy of the algorithms being trained. ASSUMPTION THAT HRV IS NOT DUE TO OTHER FACTORS like heart problems?

**Keywords:** Electrocardiogram, Heart Rate Variability, Classification, Support Vector Machine, Extreme Learning Machine, Autoregressive Modelling with Exogenous Input

1. **INTRODUCTION**

REASON FOR BELIEF IN VALIDITY:

The effects of consumption of alcohol in sudden bursts or gradually for a prolonged time in humans have been studied and documented extensively in [add ref]. ADD a few changes seen in ECG

One of the many physiological factors that alcohol consumption affects is the heart rate variability (HRV) of the consumer.

Consumption of alcohol has been known to act as a depressant on the brain and nervous tissue. Several studies [add ref] have linked chronic consumption of alcohol to changes in the HRV of the subject. HRV is the variation or change in the inter-beat interval of the heart, and the studies [add same ref] have also detailed the correlation between the amount of alcohol consumed and the extent of change in the HRV, while accounting for factors like the subject’s gender, weight, BMI, etc. From the results of these studies it is possible to draw conclusions about the differences in alcoholic and normative subjects based purely on HRV.

OVERALL FLOW OF PROJECT:

In this article, HRV analysis has been performed on the ECG signals to obtain three different types of features. They include the time domain, frequency domain (also depends on HRV since it PSD of signal varies do to number of R peaks that occur in the same interval of time?) and non-linear features. Seven time domain features, three non-linear and thirteen frequency domain features have been extracted. These twenty-three features were sufficient to get accuracies in the range of 80% and higher for both classifiers. To further the accuracy of the system, a new set of features were extracted from the ECG signals using autoregressive modelling with exogenous inputs (ARX). ARX features proved to increase training accuracy

Two different types of classifiers have been used. The Support Vector Machine (SVM) uses the concept of hyperplanes and separating surfaces, and the other classifier, the Extreme Learning Machine (ELM) uses ideas from neural networks. Both these distinct types of classifiers have been used on the same dataset and have been compared.

1. **METHODOLOGY**

**2.1 Dataset Description**

In order to classify test subjects as alcoholics or normative with a reasonable accuracy, the classifier was trained with ECG data recorded at the Autonomic Lab, Department of Neurophysiology, NIMHANS, Bengaluru. The dataset consisted of 67 ECG samples, of which 38 were samples of alcoholic patients and 29 samples were that of normative people. The ECG samples had a sampling frequency of 1kHz and each sample was recorded for approximately 5 minutes (300 seconds). Of the samples recorded, only 28 alcoholic and 28 normative samples were used to train the classifiers to prevent biasing of the algorithms.

**2.2 Pre-processing**

The ECG dataset contained baseline wandering and power-line noise, which is removed in the pre-processing stage. Eight level wavelet decomposition is performed on the signal, and the 8th component when removed from the original signal filters the signal of baseline wandering. The power-line noise has a much smaller amplitude than the total swing in the ECG signal, and thus is left unfiltered.

**More** Wavelet **explanation** and **graphs**

**2.3 Feature Extraction**

Features for the classifiers are extracted from the filtered signal. Four types of features have been obtained. They are obtained as mentioned in the sections 2.3.1, 2.3.2, 2.3.3, and 2.3.4.

**2.3.1 Time Domain Features**

In this type of feature extraction, the time instants of occurrences of R peaks is stored in an array, and the RR intervals (the time passed between two R peaks) is calculated. The series of RR intervals is used to calculate the time domain features. The following features are extracted:

1. Mean of RR intervals (RR\_mean)

This feature is the arithmetic mean calculated on the RR interval series and was calculated as follows

EQN

1. Standard Deviation of RR intervals (RR\_std)

This feature is the standard deviation of the RR interval series and was calculated using

EQN

1. Mean Heart Rate (HR\_mean)

From the RR interval series, the average frequency of occurrence of the RR intervals per minute was calculated using [ADD EQN NO]. The inverse was then taken which resulted in the mean heart rate.

EQNS

1. Standard Deviation of Heart Rate (HR\_std)

The standard deviation of the heart rate was calculated in a similar manner as the mean heart rate. First the standard deviation of the of the frequency of occurrence of the RR intervals per minute was calculated. Then, the inverse was taken to obtain the standard deviation of the heart rate.

EQNS

1. Root Mean Square of RR intervals (RR\_rms)

The square root of the mean of the sum of the squares of all the entries in the RR interval series results in the RMS of the RR interval series. The formula for the same is as follows:

EQN

1. Number of Intervals Varying by Larger than a Threshold (RR\_50)

This feature is slightly different from the rest, in the sense that it involves an additional step of taking differences. While obtaining the RR interval series required taking successive differences of the time instants at which R peaks occurred, here, successive differences are taken for the values in the RR interval series itself. On this new series of difference, the number of time differences that are larger than 50ms are counted to yield the RR\_50 feature.

1. Relative Number of Intervals Varying Larger than a Threshold (RR\_r50)

The previous feature obtained divided by the total length of the the RR interval series gives rise to the final time domain feature. This can be represented by the following equation:

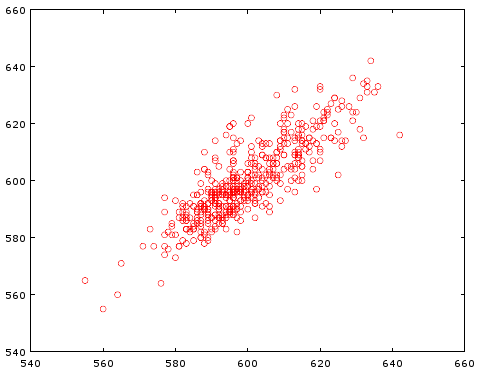
EQN

**2.3.2 Non-Linear Features**

Three non-linear features have been extracted. There were two types of analysis done, one of which yielded two features while the other gave rise to a single feature. These two methods are as given below:

1. Poincare Plot

This feature is extracted ‘graphically’. A graph of Poincare points is plotted, where the horizontal axis value of all the points is some i-th value in the RR interval series, while the vertical axis value of the point is the (i+1)-th value of the RR interval series. Then, the standard deviations of the points along two different axes are calculated to yield two of three non-linear features denoted by SD1 and SD2 respectively. The axes along which the standard deviations are calculated are the x = y line and the x = -y line. A sample of the Poincare plot obtained for the first alcoholic sample in the dataset has been shown below.



1. Approximate Entropy

This feature is essentially a measure of how much irregularity exists within the RR interval series.

Explain the rest with EQNS

**2.3.3 Frequency Domain**

The frequency domain features calculated are obtained from the power spectral density (PSD) of the ECG signal. The frequency domain features that have been extracted are:

1. Peak Frequency in VLF, LF, HF (pk\_freq\_vlf, pk\_freq\_lf, pk\_freq\_hf):

EQN

1. Absolute Power in VLF, LF, HF (ab\_pow\_vlf, ab\_pow\_lf, ab\_pow\_hf):

EQN

1. Total Power of the Signal (pw\_ttl):

EQN

1. Relative Power in VLF, LF, HF (rp\_vlf, rp\_lf, rp\_hf):

EQN

1. Normalized Power in LF, HF (norm\_lf, norm\_hf):

EQN

1. Ratio of Absolute Power of LF and absolute power of HF (ratio):

EQN

**2.3.4 Coefficients of Autoregressive Model with Exogenous Input**

What to add?

What is the actual model? … A predictive model that tries to relate initial half of signal to second half of the signal … a method to measure gradual, grouped change

**2.4 Feature Reduction ??**

**TBD**

**2.5 Classifiers**

Two classifier algorithms, each having a different ideology behind it has been trained on the dataset. One of the classifiers is the Support Vector Machine (SVM) which uses the idea of hyperplanes and decision boundaries, while the other is based off neural network concepts and is the Extreme Learning Machine (ELM).

**2.5.1 Support Vector Machine**

A support vector machine with regularization was implemented on the dataset of 28 alcoholic and 28 normative samples. An RBF kernel was used [ADD ref] in order to better separate the datapoints in a higher dimension. The SVM used the Simplified SMO algorithm [Add ref] to solve the Lagrangian problem and obtain the weights for the hyperplane.

RBF KERNEL EQN

The dataset was divided randomly into k-folds. One fold was used as the validation set, while the rest were used to train the system. The averaged result of k-such validation accuracies resulted in the k-fold validation accuracy. Such a process of obtaining the k-fold validation accuracy was conducted a large number of times (two hundred times). In each of the 200 iterations, the dataset was relabelled randomly into k new folds, and the optimal C and sig for such a labelled dataset was obtained by looking at which C and sig pair yielded the highest k fold cross validation accuracy. After all the 200 loops had been iterated, the pair of C and sig that was selected the most (ie. The pair that yielded the highest k-fold accuracy when the dataset was labelled randomly into seven folds 200 times ~~was chosen the most number of times~~) was used to train the final SVM classifier model.

k-folds of training and validation, the optimal C and sig were chosen?...other way around right? accuracy obtained, the dataset was split randomly into training and testing folds. For each such round of validation a pair of regularizaiton and variance values were selected. The most frequently used pair was then used to train the final model and obtain the weights of the hyperplane.

The regularization parameter and RBF variance parameter were obtained by checking which pair of values gave the best randomized k-fold accuracy when tested multiple (two hundred times)

**2.5.2 Extreme Learning Machine**

Classifier, regularization/hidden neurons, kernel,

**2.5.3 Validation**

While the training the classifier systems, leave-one-out validation and k-fold cross validation was used to verify/validate the accuracy of the model that was trained. For both classifiers was used while performing k-fold cross validation, allowing each fold to contain eight samples with four samples from each of the two classes.

Along with the accuracy obtained through leave-one-out and k-fold cross validation, the sensitivity and specificity (eqs no1 and no2) were obtained from the confusion matrix.

ADD EQNS

Calculation of sensitivity and specificity ensured that classification accuracy remained high for both the positive and negative classes, and that the system was not biased toward a particular class.

verNoise removed. Features Extracted.

Classifiers description.

System trained. Validation method.

To increase, ARX added

System retrained and validated.

1. **3 RESULTS**
2. **3.1 Results of SVM**

The process of selecting the and pair, training the SVM, obtaining the k-fold validation accuracy have been mentioned in Section 2. Here, the results obtained for specific groups of feature sets and feature optimization performed by combnk has been provided.

The accuracies and optimal and pair obtained for the different types of feature sets are as given the table below.

|  |  |  |
| --- | --- | --- |
| Features Used | Optimal (, ) Pair | 8-Fold Accuracy |
| Time, non-linear, frequency | (0.1, 0.3) | 80% |
| Time, non-linear, frequency, ARX order 3 | (0.3, 1) | 82% |
| Time, non-linear, frequency, ARX order 5 | (0.3, 1) | 86% |

From the table above, it is seen clearly that the accuracy of the system improves when ARX coefficients are used along with the time domain, non-linear and frequency domain features.

~~As mentioned earlier, it is important to understand how many samples of each class were classified correctly. For this,~~ The confusion matrix was obtained and the sensitivity and specificity (given by eqns ADD and ADD) were calculated for each of the cases given above in (ADD table no).

|  |  |  |
| --- | --- | --- |
| Features Used | Confusion Matrix | (Sensitivity, Specificity) |
| Time, non-linear, frequency | -- | -- |
| Time, non-linear, frequency, ARX order 3 | -- | -- |
| Time, non-linear, frequency, ARX order 5 | |  |  | | --- | --- | | 25 | 3 | | 5 | 23 | | (89%, 82%) |

The table above shows that the system is very slightly biased toward the positive class since a few more samples in the negative class are being misclassified. However, this is not of much concern as both sensitivity and specificity are well above 80% accuracy.

**3.2 Results of ELM**

The process of selecting the number of neurons in the hidden layer and , training the ELM, obtaining the k-fold validation accuracy were as given in Section 2. Here, the results obtained for specific groups of feature sets and feature optimization performed by combnk has been provided.

The accuracies and optimal number of hidden layer neurons and obtained for the different types of feature sets are as given the table below.

|  |  |  |
| --- | --- | --- |
| Features Used | Optimal (No. of Hidden Layer  Neurons, ) | 7-Fold Accuracy |
| Time, non-linear, frequency | (6, 1.4) | 89.29% |
| Time, non-linear, frequency, ARX order 3 | (7, 2) | 92.86% |
| Time, non-linear, frequency, ARX order 5 | (15, 1.8) | 94.64% |

|  |  |  |
| --- | --- | --- |
| Features Used | Optimal (Number  of Hidden Layer  Neurons, ) | Leave-One-Out Validation Accuracy |
| Time, non-linear, frequency, ARX order 5 | (6, 2) | 92.86% |

From the table above, it is seen clearly that the accuracy of the system improves when ARX coefficients are used along with the time domain, non-linear and frequency domain features.

Again, the confusion matrix was obtained and the sensitivity and specificity were calculated for each of the cases given above in (ADD table no).

|  |  |  |
| --- | --- | --- |
| Features Used | Confusion Matrix | (Sensitivity, Specificity) |
| Time, non-linear, frequency | -- | -- |
| Time, non-linear, frequency, ARX order 3 | -- | -- |
| Time, non-linear, frequency, ARX order 5 | |  |  | | --- | --- | | 26 | 2 | | 4 | 24 | | (92.86%, 85.71%) |

The confusion matrix above shows how the system is classifying samples from both classes well and thus, is not biased.

**3.3 Comparative Results and Points of Discussion**

Accuracies obtained for k-fold cross validation for SVM and ELM are compared in table ( )

|  |  |  |
| --- | --- | --- |
|  | SVM | ELM |
| Without ARX Coefficients | 80% | 89% |
| With ARX Coefficients of order 5 | 86% | 94% |

Sensitivity and Specificity obtained for SVM and ELM are given in table ( ).

|  |  |  |
| --- | --- | --- |
|  | SVM | ELM |
| Sensitivity | 89% | 92.86% |
| Specificity | 82% | 85.71% |

**Points of Discussion**

* Classifiers can be applied to features extracted via HRV analysis and good accuracies are obtained.
* ELM classifier was found to have a better accuracy than SVM for the same feature set.
* ARX features improve the accuracy of both ELM and SVM.

A**dd: Real time data was captured on AD8232 sensor and classified successfully?**

**REFERENCES**

**APPENDIX**

**To Do:**

**Add equations where necessary**

**Cite sources**