# Determining Best HRV Indices for PAF Screening using Genetic Algorithm

Nermin ÖZCAN[1], Mehmet KUNTALP[2]

[1] Dokuz Eylul University, Graduate School of Natural and Applied Sciences, Turkey
nermin.ozcan@ceng.deu.edu.tr
[2] Dokuz Eylul University, Graduate School of Natural and Applied Sciences, Turkey
mehmet.kuntalp@deu.edu.tr

## Abstract

**Heart Rate Variability (HRV) analysis is used for diagnosis of various cardiac diseases. In this study, the genetic algorithm (GA) is used for the selection of optimal subset of HRV features to detect paroxysmal atrial fibrillation (PAF) patients from their ectopic-free ECG records. The K-nearest neighbors (K-NN) algorithm is used as the classifier. This GA-based method reduced the number of HRV features from 33 to 7 and increased the classifier's performance from 90.2% to 92.2%.**

## 1. Introduction

There are effective software systems designed to help physicians in the process of diagnosis and decision making based on patient data [1]. However, the biological datas in diagnostic systems usually are high size. High-dimensional data clusters cause classification complexity and these reduce classification performance. Therefore, it is necessary to reduce the data size while preserving the important information of the original data in the process. There are two basic methods for dimension reduction. These are feature selection and feature extraction methods [2].

The purpose of feature extraction is to reach from unhandled data to the information that both increases the variance between classes and reduces the variance within the classes. Algorithms like Principal Component Analysis (PCA), Multidimensional scaling (MDS) and Independent Component Analysis (ICA) are widely used for feature extraction. Feature selection is an operation type that selects the most appropriate property subset that preserves sufficient information from the original data. Feature selection algorithms such as Genetic Algorithms (GA), Support Vector Machine (SVM) and Particle Swarm Optimization (PSO) have been developed and used. Among these, GA is widely used and the GA method is chosen for this study.

GA is an heuristic algorithm and is used as an optimization method in feature selection [3]. GA imitates natural selection, genetic crossing over and the mutation that causes genetic alterations in living cells. A fitness function is calculated for each individual in the randomly generated population. The fitness function is determined as the minimization or maximization function according to the work to be performed and then the natural selection process begins. After the selection process, crossing over and mutation events are used to transfer better gene to new population. Individuals with poor fitness values are removed from the generation while individuals with good values form a new population [4]. This process is repeated until the best population is obtained.

Atrial fibrillation is an arrhythmia type where the vibrations of the atria occur outside of the required state [5]. In this rhythm disturbance, a large number of electrical activities within the atrium are caused by completely irregular and very fast excitation, moving in different directions. As a result, very rapid and irregular contractions occur at each of the atria. During these rapid contractions, blood cannot be adequately filled into the ventricles and there may be a reduction in the amount of blood pumped into the body. This condition may lead to stroke and even death. Paroxysmal atrial fibrillation (PAF) is a type of atrial fibrillation which can lead to more dangerous forms of AF. The best way to detect PAF condition is the ECG record taken during an ectopic attack. However, PAF ectopic attacks occur randomly and may disappear in a very short time. Therefore, it is hard to take an ECG record during an ectopic attack. It would be of great help if it is possible to detect PAF condition from normal, i.e. ectopic-free, ECG records of a subject.

The aim of this study is to determine whether a subject is a PAF patient or not from his/her ectopic-free ECG recordings (i.e. PAF screening) by using the best HRV features chosen by the GA method.

## 2. Method

In the first part of this section, the ECG data set used and the HRV indices obtained from this set are mentioned. In the second part, K-NN classifier is described. In the third part, Genetic Algorithms are discussed.

### 2.1. Dataset Description

The ECG data used in this study is obtained from the Physio-Net database, a web-based resource that is freely supplied data to researchers with a wide range of physiological signals. The dataset contains two channel ECG records. The ECG signals were sampled at 128 Hz and digitized with 12-bit resolution. There are 100 ECG recording sets in the database. Each record set contains two 30 minute records from the same subject. Of these 100 recording groups, 53 are from persons who had previously been labeled as PAF patient; the remaining 47 sets of records belong to persons who do not have any PAF history [6].

A system for diagnosing PAF patients has been previously proposed based on ECG records taken at normal sinus rhythm for the same data set (İrem Hilavin, 2016). A combination of time domain, frequency domain and non-linear HRV features has been proposed to separate PAF patients and non-PAF

persons. All 33 HRV features in the mentioned study are presented in Table 1. We used these HRV features as the raw dataset in this study.

**Table 1.** HRV indices obtained from ECG data

| Feature | Description |
|---|---|
| Mean RR | Mean of RR intervals |
| Std RR | Standard deviation of RR intervals |
| Std HR | Standard deviation of HR |
| RMSSD | Square root of the mean squared |
| NN50 | Number of successive RR interval |
| PNN50 | NN50 divided |
| VLF peak | VLF band peak frequency |
| LF peak | LF band peak frequency |
| HF peak | HF band peak frequency |
| VLF power | Absolute power of VLF band |
| VLF power prc | Relative power of VLF band |
| LF power | Absolute power of LF band |
| LF power prc | Relative power of LF band |
| LF power nu | Power LF band in normalized units |
| HF power | Absolute power of HF band |
| HF power prc | Relative power of HF band |
| HF power nu | Power of HF band normalized units |
| LF/HF power | Ratio of LF and HF power bands |
| ApEn | Approximate entropy |
| SampEn | Sample entropy |
| DFA Alpha1 | Short term fluctuation analysis |
| DFA Alpha2 | Long term fluctuation analysis |
| CorDimD2 | Correlation dimension |
| RPA Lmax | Max of Recurrence plot analysis |
| RPA Lmean | Mean of Recurrence plot analysis |
| RPA Rec | Recurrence plot recurrence rate |
| RPA Det | Recurrence plot determinism |
| RPA ShanEn | Recurrence shannon entropy |
| SD1 | Dispersion of points |
| SD2 | Dispersion of points |
| CCM | Complex correlation measure |
| AEN | Atrial ectopic number |
| VEN | Ventricular ectopic number |

## 2.2. K-NN Classifier

K-nearest neighbors (K-NN) is a sample-based learning algorithm and one of the simplest but most successful of machine learning algorithms [7]. It is an algorithm that computes closeness of a selected data or a group of data and decides their classes. In general, Euclid distance is used for the proximity account in this study. The number K indicates how many neighbors in the circumference are to be computed. The use of odd numbers for K is preferred as dual numbers can cause equality.

The algorithm is designed to run the K-NN classifier for each individual in the population. Thus, the classification accuracy is calculated for each individual. This accuracy indicates the contribution of that individual and forms the fitness value of the individual. Different K values were tried used in this and K=3 was chosen since it produced the best results.

## 2.3. Genetic Algorithm

Genetic algorithm (GA) is a search method that mimics the natural development process. This heuristic method is used to produce useful solutions to optimization and search problems using techniques inspired by natural development such as inheritance, mutation, selection and crossover [7].

Dimension reduction problem is very suitable for formulation as an optimization problem. Given a set of d-dimensional input patterns, GA's task is to find a transformed set of m-dimensional space (m<d) patterns that maximizes a set of optimization criteria [6]. The basic elements of a GA mechanism used during feature selection are given below.

### 2.3.1. Initialization Population

In the population, each row specifies one individual while each column specifies the characteristics of the data set. A binary system has been used to show the properties of individuals. It is meant that if the value of individual in i. column is 1 it will participate in the classification. If the value of individual in i. column is 0 it will not participate classification.

### 2.3.2. Fitness Value and Selection

Fitness is the value calculated for each individual and can vary depending on the purpose of the study. For feature selection, this value is accepted as the classification accuracy.

The selection process is aimed to creat groups that have two individuals from the population for the crossingover. The roulette wheel method was used in the study [8]. In this method each individual has a slice according to the fitness value of the wheel. The wheel is turned and the individual in the slice on which the spin is stopped is selected. The roulette wheel method gives each individual a chance to be selected. The individual who has the better fitness value has the bigger slice at the same time, thus its characteristic features are transferred to the next generation.

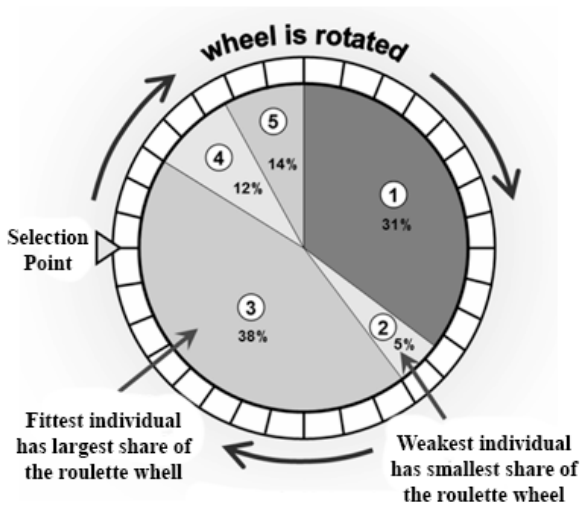


**Fig. 1.** Method of Roulette Wheel.

### 2.3.1. Crossing Over and Mutation

Crossingover (see Fig. 2) is also known as track change. It is the fact that some of the genes are taken from the mother while the other part is taken from the father.

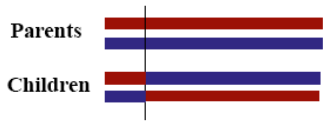


**Fig. 2.** Crossing Over.

Mutation (see Fig. 3) is a randomly changing of a gene in chromosomes. It is important for diversity. When the bit value is 0, it changes to 1 and if the bit value is 1, it will change to 0 [9].

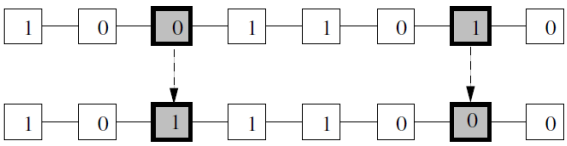


**Fig. 3.** Mutation in GA.

### 2.4. Feature Selection

The main bearing of the implementation of attribute selection techniques is to remove unnecessary and irrelevant features in the data set without causing any loss of information [3]. The algorithm is applied during feature selection is as shown in Figure 4.

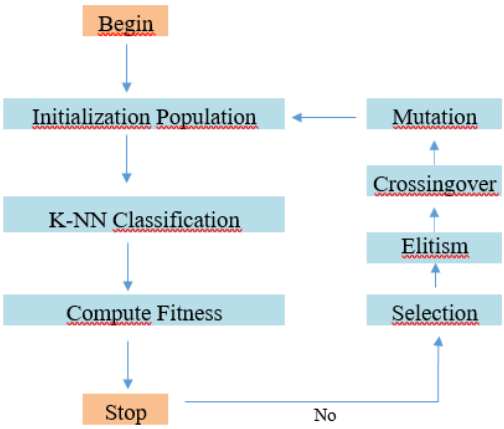


**Fig. 4.** Algorithm of GA.

The initial population is randomly generated, consisting of 0 and 1 's. Each row is regarded as an individual. New data is obtained by choosing random population's columns that is 1 from the columns of original data. Each selected data is inserted into the K-NN classifier and the classification success is calculated for each data set. This indicates how much the selected bits affect the classification, that is, how important it is to the data set. At the end of the classification, each individual is given a classification success value. At the same time it is the fitness value of genetic algorithm. Individuals are ranked according to their fitness values. Their slices in the roulette wheel are determined and dual groups are selected randomly. In some cases, individuals with high fitness values are prevented from being lost in this random selection. This is called elitism. Individuals that have higher fitness value are transferred to the next step directly. After the parents have been identified, crossingover occurs and mutation steps take place. At a result of these processes, the children chromosomes with high fitness values, form the new population and the low ones are removed from the population. The transactions are repeated until the stopping criterion is reached.

In genetic algorithm, there are some parameters of operations such as crossingover, mutation, elitism. The parameters used in the study are shown in Table 2.

**Table 2.** Parameters of Genetic Algorithm

| Description | Feature | Value |
|---|---|---|
| Stopping Criteria | i | 200 |
| Crossover Probability | Pc | 0.85 |
| Elitism Probability | Pe | 0.01 |
| Mutation Probability | Pm | 0.05 |
| Population Type | Ptype | Bit String |

## 3. Results

Given the parameters and all 33 HRV features, the GA identified individuals with the best fitness values after 200 iteration. Only 7 of the 33 features were selected for the best individual. The selected HRV features are shown in Table 3.

**Table 3.** Selected Features

| Column | Feature |
|---|---|
| 1 | Mean RR |
| 2 | Std RR |
| 9 | HF peak |
| 13 | LF power prc |
| 16 | HF power prc |
| 23 | DFA Alpha1 |
| 31 | CCM |

The classification made with these indices has an error value of 7.8%. This value indicates that the success of the K-NN classifier is 92.2%. When K-NN is run without considering the feature selection, that is using all 33 features, 90.2% success rate is obtained. The study is performed using 2015 version of MATLAB software package.

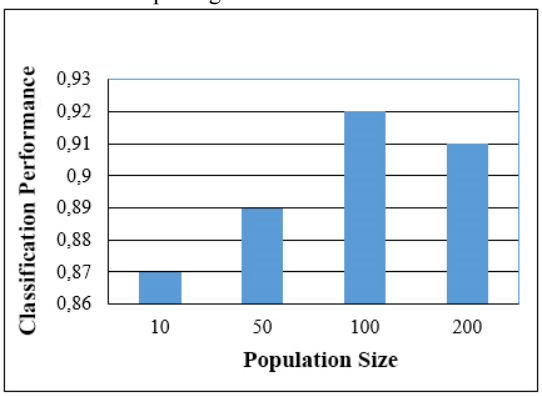


**Fig. 5.** Effect of population size on performance.

Genetic algorithms start the population with randomly generated vectors. This sometimes prevents achieving the best result. In order to avoid such a problem, different population sizes were chosen. The applied population values and their effects on the overall performance are given in Figure 5. It can be seen from this figure that there is an optimal value for the population size.

## 4. Conclusions

In this study, the best HRV feature subset was determined for PAF screening purposes. The proposed system accommodates the genetic algorithm (GA) to search and form the feature subsets and the K-NN classifier as the evaluation tool to select the best feature subset. The performance of the proposed method was analyzed in terms of number of features reduced and the classification accuracy produced by the K-NN classifier. From the conducted experiments, it is evident that the subset consisting of only 7 HRV features selected by the GA outperforms the case where all 33 HRV features are used. The obtained accuracy of 92.2% shows that the HRV features selected by the GA in this study can be used for PAF screening purposes.

## 5. References


[1] Ömer Deperlioğlu, Gür Emre Güraksın ve Utku Köse, "Web-Based Clinical Decision Support System: Features and Development,"Academic Informatıcs Conference, s.2-3, February 2016.

[2] Min Zhu, Jing Xia, Molei Yan, Guolong Cai, Jing Yan ve Gangmin Ning, "Dimensionality Reduction in Complex Medical Data: Improved Self-Adaptive Niche Genetic Algorithm," *Computational and Mathematical Methods in Medicine*, s.1-2, Jully 2015.

[3] Muammer ALBAYRAK ve Ahmet ALBAYRAK, "Feature Selection with Genetic Algorithm in Classification of Mesothelioma Disease Data," *Tıptekno16*, s.2-3, October 2016.

[4] D. Asir Antony Gnana Singh, E. Jebamalar Leavline, R. Priyanka ve P. Padma Priya, "Dimensionality Reduction using Genetic Algorithm for Improving Accuracy in Medical Diagnosis," *Modern Education and Computer Science*, s.3, January 2016.

[5] Prof. Dr. Ali Oto, "Turkish Society of Cardiology Atrial Fibrillation Diagnosis and Treatment Guide," *Turkish Society of Cardiology,* December 2003.

[6] İrem HİLAVİN, "Development of a System to Diagnose Atrial Fibrillation Patients From Arrhythmia Free ECG Records," *A Thesis Submitted to the Graduate School of Natural and Applied Sciences of Dokuz Eylül University, s.74-82, April 2016.*

[7] Pat Langley ve Wayne Iba, "Average-Case Analysis of a Nearest Neighbor Algorithm," *Conference on Artificial Intelligence*, s.1, 1993.

[8] Fernando Gómez ve Alberto Quesada, "Genetic algorithms for feature selection in Data Analytics," Artelnics, August 2017.

[9] Carlos Alberto Nasillo González, "Polymorphic Virus Signature Recognition via Hybrid Genetic Algorithm," *github.com/carlosnasillo/Hybrid-Genetic-Algorithm,* August 2017.

[10] D Beasley, DR Bull ve RR Martin, An Overview of Genetic Algorithms. UK 1993.