# Cardiac Arrhythmia Classiﬁcation.

**Rajan Sanghavi**

**The University of Texas at Dallas.**

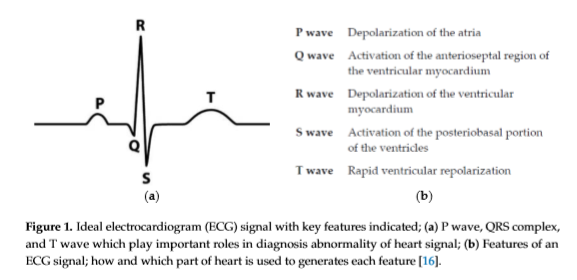
## **Abstract**

Heart arrhythmia (also known as arrhythmia, dysrhythmia, or irregular heartbeat) is a group of conditions in which the [heartbeat](https://en.wikipedia.org/wiki/Cardiac_cycle) is irregular, too fast, or too slow. A [heart rate](https://en.wikipedia.org/wiki/Heart_rate) that is too fast – above 100 beats per minute in adults – is called [tachycardia](https://en.wikipedia.org/wiki/Tachycardia) and a heart rate that is too slow – below 60 beats per minute – is called bradycardia. [Wikipedia link](https://en.wikipedia.org/wiki/Heart_arrhythmia)

**Data Set Information:**

This database contains 279 attributes, 206 of which are linear valued, and the rest are nominal.   
  
Concerning the study of H. Altay Guvenir: "The aim is to distinguish between the presence and absence of cardiac arrhythmia and to classify it in one of the 16 groups. Class 01 refers to 'normal' ECG classes 02 to 15 refers to different classes of arrhythmia and class 16 refers to the rest of unclassified ones. For the time being, there exists a computer program that makes such a classification. However, there are differences between the cardiology’s and the programs classification. Taking the cardiology’s as a gold standard we aim to minimize this difference by means of machine learning tools."   
  
Further Dataset information at [UCI Machine Learning Repository.](https://archive.ics.uci.edu/ml/datasets/Arrhythmia)

Explanation of Features:



Pic Credit: Wikipedia

**DataPreprocessing**

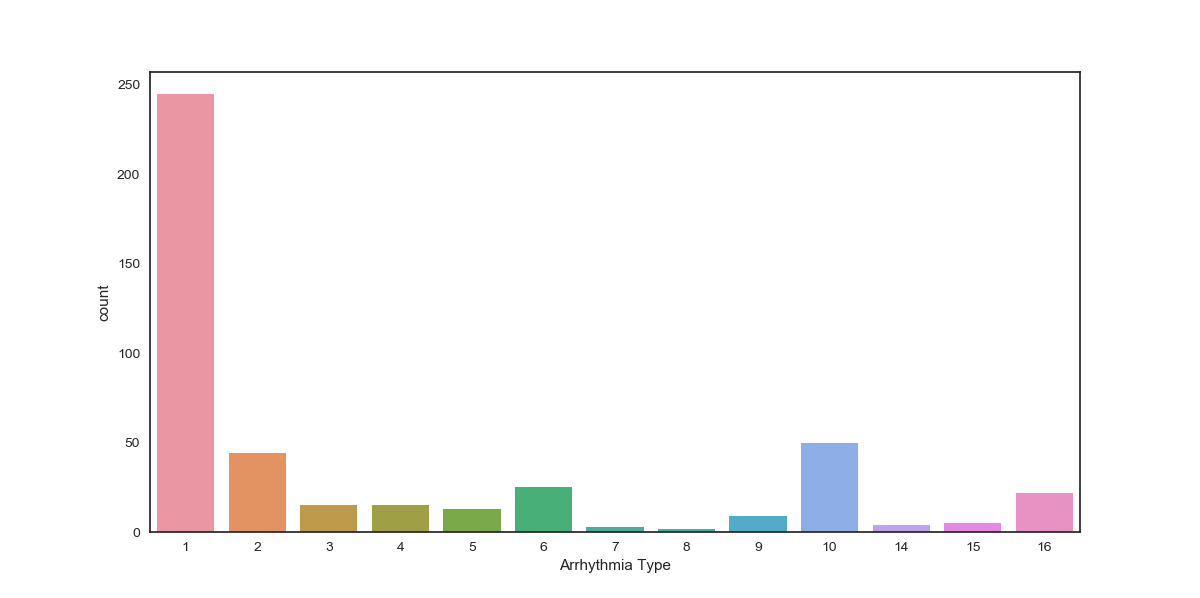
The original data contains columns with both missing values and single valued columns having the same value for all the patient records. The one column with more than 80% missing values was deleted from the data set. Single value column does not provide any relevant information for machine learning algorithms and hence they were dropped from our analysis. There were such sixteen columns which were dropped. The resulting data set contained 452 instances and 261 features and 1 target variable.

**FeatureSelection**

One of the reasons for using fewer features is the limited number of data records (452) compared to features (261). This helps in avoiding overﬁtting and gives insight into the important features which have maximum correlation with the output labels. We chose to take features which fall in top 20% in descending order of feature importance to our target variable.

**Distribution and Train Test Splits**

The dataset is highly bias with majority of records skewed towards no Cardiac Arrhythmia.



We will use stratified Train Test split method, with 30% data for testing our models.