# Cardiac arrythmia prediction based on machine learning

Today, arrhythmia is one of the diseases that can be easily diagnosed, but also successfully treated with the necessary medication. This paper will present some of the main algorithms in the field of machine learning (ML) that can be used to draw various conclusions, as well as diagnose diseases.

**Author**
Mario Negovetić
Faculty of Engineering
University of Rijeka, Croatia

Keywords: Accuracy, Algorithms, Arrhythmia, Machine Learning, Python.

## 1 Introduction

Although AI is slowly beginning to be used in the field of cardiology, expert cardiologists are still necessary to diagnose the disease. The use of AI would speed up the process of diagnosing disease and treating it, but in some cases, it might also allow for a more precise definition of the patient's condition.
**The main aim is to distinguish between the normal condition and types of cardiac arrhythmia, and then to classify it in one of the 16 groups.**

## 2 Methodology

The dataset consists of 452 instances and 279 attributes, and needed appropriate filtering and values scaling. Then, dataset correlation was carried out as to filter out low-correlation columns into a new dataset (*Correlation dataset*), while still preserving the old one (*All dataset*).

Using 5 hyperparemeter-tuned algorithms, prediction was done on the datasets, where 75% of the datasets was put to train data, and 25% to test data.

**Algorithms**

- Decision Trees
- Random Forest
- Gradient Boosting
- Support Vector Machine (C-Support Vector)
- K-Nearest Neighbour

## 3 Results

### *All dataset*

- best ROC-AUC: Gradient Boosting Classifier (0,80)
- worst ROC-AUC: KNN (0,67)
- best F1: Gradient Boosting Classifier (0,37)
- worst F1: KNN (0,27)

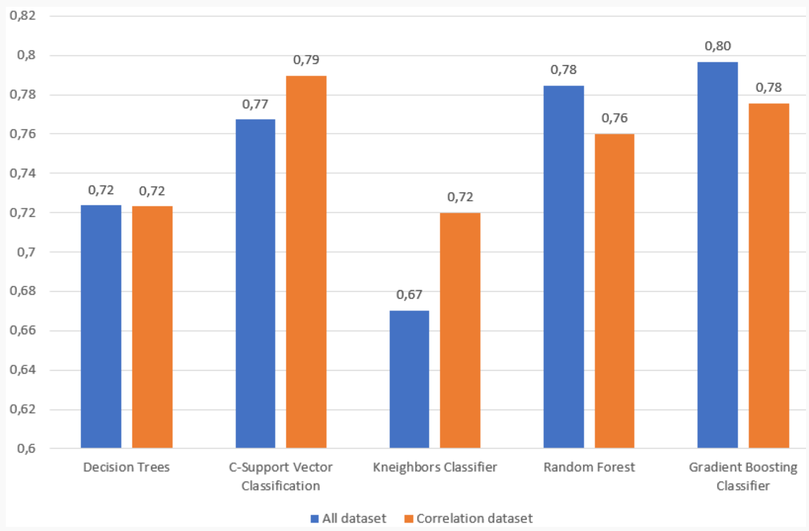


**Figure 1.** ROC-AUC scores for all algorithms

### *Correlation dataset*

- best ROC-AUC: C-Support Vector (0,79)
- worst ROC-AUC: KNN (0,72)
- best F1: Random Forest (0,29)
- worst F1: KNN (0,23)

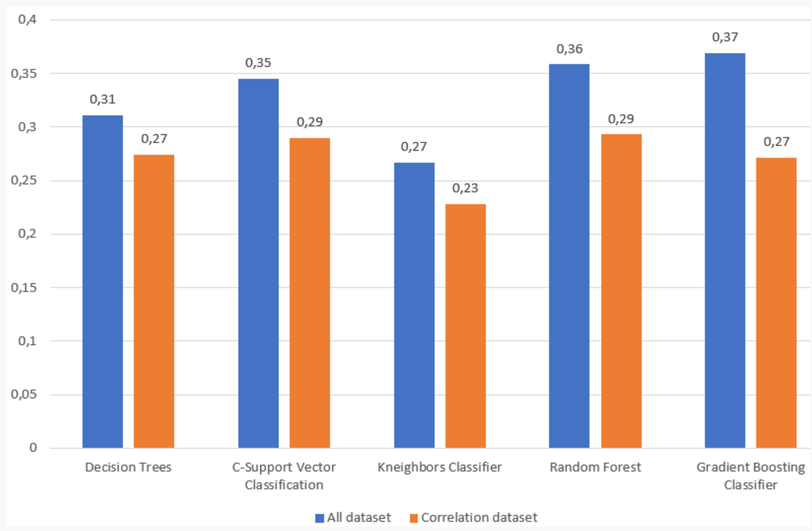


**Figure 2.** F1 scores for all algorithms

## 4 Conclusion

The problem of poor performance in this study is primarily related to the dataset. There is too much data concentration in one class (Normal class), while the others have little to no data. Therefore, the model fails to recognize other classes due to there not being enough samples to learn which features result with other class. The results would certainly be even better if a larger dataset was used.