Classification and Detection of Arrhythmia using Electrocardiogram Features

Alan Tang
University of Massachusetts, Amherst

Arham Choraria
University of Massachusetts, Amherst

Shaham Zahir
University of Massachusetts, Amherst

***Abstract*—An arrhythmia is a problem with the rate or rhythm of a heartbeat. This could be a heart that beats too quickly, too slowly, or with an irregular pattern with a multitude of complications including a heart stroke or even sudden death. Provided with a data set from the UC Irvine Machine Learning Repository, the data set was refined to suit the task more appropriately. The objective of this project was to train a model to classify regular and irregular heartbeats from electrocardiogram signals. Using features extracted from the ECG signals, the models differentiated between regular and irregular heartbeats to find out whether arrhythmia was present or not. K Nearest Neighbors, Decision Trees and Random Forest were used and were found to have perfect accuracy in their binary classifications. The various metrics used such as accuracy, f1 score, precision recall curves, ROC curves, classification reports and confusion matrices all indicated perfect results, indicating some very discriminating features within the data set enabling such high performance. From a feature importance standpoint, the decision tree and random forest both placed a high importance on similar variables, with random forest considering a few more variables compared to the Decision Tree. Overall, with 278 initial attributes and decision tree using 1 feature and random forest using 18 features, it is clear a few features are extremely important to the data set in order to detect the presence of an arrhythmia.**

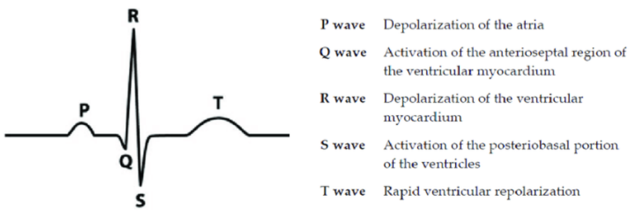
I. Introduction

The heart is the electrical powerhouse of the human body. It pumps thousands of gallons of blood every day and beats 100,000 times keeping the body alive, thanks to the body's internal electrical system. This typical beat goes on until a kink in the electrical system causes it to malfunction [7].

When electrical activity interrupts the heart, it's called an arrhythmia. Arrhythmia is related to the improper beating of a heart. It is a case where the rate or the rhythm of the heartbeat is irregular and it occurs when the electrical impulses in the heart don't work properly. This means that the heart beats too quickly, too slowly, or basically with an irregular pattern. When the heart beats faster than normal, it is called tachycardia and when the heart beats too slowly, it is referred to as bradycardia [8]. In general, arrhythmias aren't harmful unless they interfere with the heart's ability to pump blood, which can become life-threatening. Moreover, arrhythmias are associated with an increased risk of blood clots. Hence, if a clot breaks loose, it can travel from the heart to the brain, causing a stroke. It is important that knowledge around arrhythmia increases so people can be more easily diagnosed and treated before potentially fatal incidents occur. Arrhythmia is also a very common disease, especially in the United States. About 14 million people in the USA have arrhythmias [3]. The most common disorders are atrial fibrillation and flutter and the incidence is highly related to age and the presence of an underlying heart disease. The incidence approaches 30% following open heart surgeries.

It is important to know how this disease is normally detected through the electrocardiogram signals. Electrocardiography (ECG) is a procedure used to evaluate the electrical activity of the heart with reference to time by insertion of electrodes on the skin [2]. These electrodes can recognize trivial electrical changes in skin and are placed in several areas of the human body. ECG detects physical cardiac activities which are shaped by the polarization and depolarization of the atria and ventricles of the heart and that is the basic procedure of how an ECG signal is produced [4]. These signals generally follow a pattern that indicates if the heart is beating correctly. However, when these signals seem to be improper or irregular, it is an indication of having arrhythmia. Fig 1 shows an example of an ideal ECG signal with labeled features along with how they are detected using the electrodes. The ECG signals consist of several features such as the P, the QRS complex (the Q, R and S wave segment as shown) and the T waves and studying such features plays an imperative part in the diagnosis of various arrhythmias as different rhythm disturbances need different treatments so diagnosing the precise type of arrhythmia is important [1]. Furthermore, these features also play an important role in diagnosing abnormalities of heart signals as an irregular pattern in these wave features is what helps detect the presence of arrhythmia. These features have also been used throughout this project in order to detect and classify arrhythmia.

Fig. 1. Ideal ECG Signal Diagram [1]



P wave	Depolarization of the atria
Q wave	Activation of the anterioseptal region of the ventricular myocardium
R wave	Depolarization of the ventricular myocardium
S wave	Activation of the posteriobasal portion of the ventricles
T wave	Rapid ventricular repolarization

II. Methods

The aim of the project was to be able to create a model that would be able to accurately classify electrocardiogram data into that of a regular heartbeat or one with arrhythmia. As such, this was interpreted as a goal that is achievable with binary classification. The dataset chosen for this curated project was the arrhythmia dataset from the University of California Irvine's Machine Learning Repository [5]. Originally the data was a 452 row dataset with 279 attributes before it was deemed necessary to remove an attribute that was missing for almost every sample, as well as those rows with some missing attributes and finally to remove results that would have inconclusive results. In the modified dataset, there are 402 rows which each represent the electrocardiogram data for an individual, and there are also 278 attribute columns with 206 of them being linear values. The final 279th column is the arrhythmia classification, in which class 01 refers to 'normal' ECG, while classes 02 to 15 refers to different classes of arrhythmia and class 16 refers to the rest of the unclassified ones. Class 16 would have led to an inability to train or validate a model due to these pieces of data lacking actual results. The majority of the dataset was standardized for the purposes of the objective to the value of '1' while classes 2-15 of varying types of arrhythmia were all changed to a value of '0' for a true binary classification. For the methods of classification 3 models were selected: K-Nearest Neighbors, Decision Trees and Random Forest. A typical stratified split of the data was done into a 15/15/70 split for testing, validation and training, respectively. There was also usage of the StandardScale function to standardize features by removing mean and scaling to unit variance

A. K-Nearest Neighbors

K-Nearest Neighbors was selected for its classification purposes as a very simple and standard supervised classification machine learning model. The accuracy of the model was verified using the sklearn metrics python library's accuracy function, a confusion matrix and a classification report [6]. The precision recall graph was also examined for an additional measure. The variation within this model type could potentially stem from the chosen number of neighbors due to the model's usage of proximity to neighbors and their similarities in the dataset. As such, the number of neighbors was treated as a hyperparameter and a myriad of numbers were input into the KNN model on the arrhythmia dataset. Interestingly enough, all of the numbers of neighbors led to the exact same confusion matrix, precision recall plot, classification report and f1 score. As such, the number of neighbors was found to be irrelevant for this particular dataset's usage in a binary classification. Additionally the ROC curve was plotted and the AUC was measured as well.

B. Decision Trees

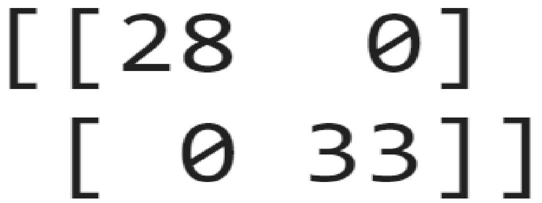
The second model implemented in order to classify electrocardiogram data into normal or arrhythmia classification was the decision tree. The decision made was to use decision trees in order to gain insight into which features of the dataset may be of more importance in splitting the data. Since decision trees split the data on binary nodes of the tree, it can easily be visualized how the model does feature importance ranking and prioritization, enabling the ability to clearly comprehend which features lead to certain classification outcomes or which may compound upon one another to be sorted. The python library sklearn's DecisionTreeClassifier with entropy criterion on the modified dataset with the binary outcomes of normalcy and arrhythmia was used in order to perform the model. Then the accuracy of the model was examined with the sklearn metrics accuracy function and a precision recall curve.

C. Random Forest

The final model chosen was the Random Forest model. The random forest implements an ensemble of multiple decision trees in order to avoid the possible overfitting or bias of a single decision tree, enabling us to compare and contrast the outputs of the decision tree with those of the random forest model. Specifically, it was desired to be able to examine the feature importances of each in order to determine if there was a shared feature or set of features that clearly determined the status or classification of a patient's data. Then there was an examination of the feature importance to determine the higher weights placed on certain attributes/features using the sklearn library's RandomForestClassifier.feature_importance_ function. The accuracy of the model was verified using the sklearn metrics accuracy function along with a precision recall curve.

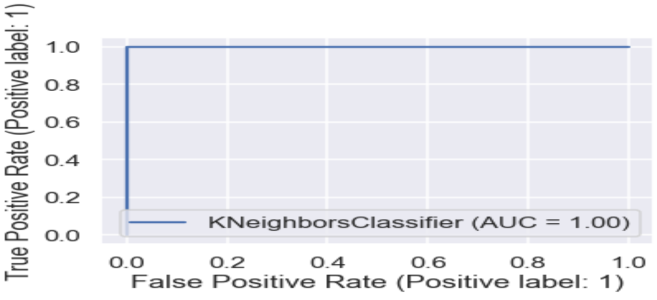
III. Results

Fig. 2. KNN Confusion Matrix



```
[[28  0]
 [ 0 33]]
```

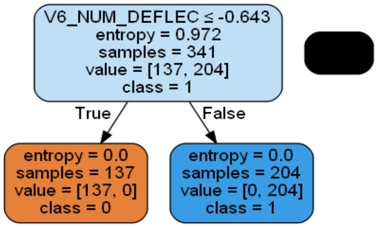
Fig. 3. KNN ROC Curve



In normal practice, the optimal value of K for the K-Nearest Neighbors model is usually the square root of the total number of samples but while testing the models on the given dataset, it was found to be fully accurate and perfectly predictive over a wide range of neighbor values which suggested that the error rates of the K values used on the dataset were the same. The confusion matrix shown in Figure 1 was also consistent over all values of the neighbors in a test size of 61 (15% of the dataset), with 28 true positives and 33 true negative values, indicating that there were no false negatives nor false positives. Additionally, the precision recall curves were the same over these number of neighbor distributions along with the same f1 score and accuracy metric. Finally the AUROC shown in figure 2 showed a similar trend of perfect classification by the KNN model. Essentially, the K-Nearest Neighbors model was able to perfectly and accurately predict the correct resulting binary classification of the electrocardiogram data based upon its features.

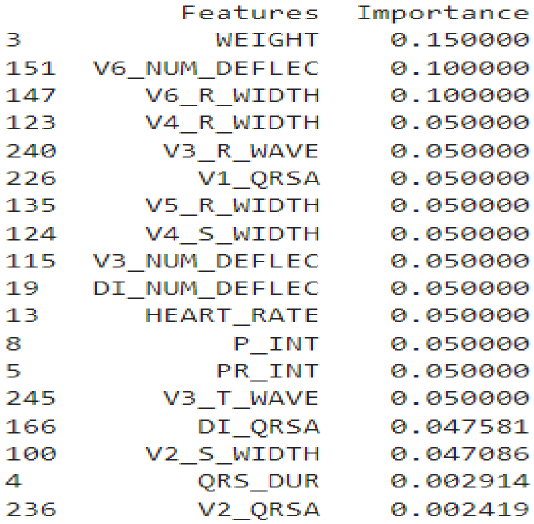
Moving on to our second model, the Decision Tree was also found to be fully accurate according to the metrics chosen to measure its accuracy. The entire dataset was found to be split upon a single variable, which is the ‘V6_NUM_DEFLEC’ variable in the dataset which means the number of intrinsic deflections in channel V6. This variable split the entire dataset with a value of -0.643. Values greater than that were put in one branch and values less than or equal to were put in the other branch as shown in the tree figure. The f1 score, accuracy score, and precision recall plot all had a perfect value of 1.0 for this model as well..

Fig. 4. Decision Tree split



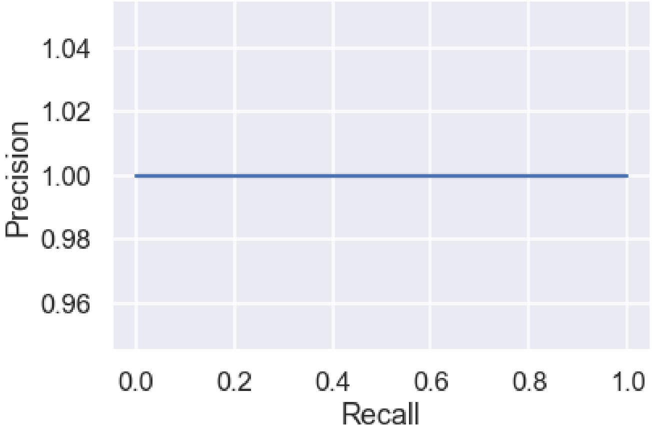
Finally, and not so surprisingly anymore, the Random forest model also had perfect accuracy on the dataset. However, the results were slightly different. Due to the nature of random forest’s voting on feature selection, the feature selection function did not prioritize V6_NUM_DEFLEC as the highest weighted variable like the Decision Tree model did but instead, WEIGHT was at 0.15 being weighted the highest.Variables V6_R_WIDTH (the average width of the R wave for channel V6) and V6_NUM_DEFLEC were tied for the second highest weighted variable at 0.1, with the other 15 variables out of the initial 278 having very low weights. A total of 18 variables were judged to have weights by the Random Forest model, meaning 260 of the variables had a weight of 0.

Fig. 5. Random Forest Feature Importance Table


```
           Features  Importance
3            WEIGHT    0.150000
151   V6_NUM_DEFLEC    0.100000
147      V6_R_WIDTH    0.100000
123      V4_R_WIDTH    0.050000
240       V3_R_WAVE    0.050000
226         V1_QRSA    0.050000
135      V5_R_WIDTH    0.050000
124      V4_S_WIDTH    0.050000
115   V3_NUM_DEFLEC    0.050000
19    DI_NUM_DEFLEC    0.050000
13       HEART_RATE    0.050000
8             P_INT    0.050000
5            PR_INT    0.050000
245       V3_T_WAVE    0.050000
166         DI_QRSA    0.047581
100      V2_S_WIDTH    0.047086
4           QRS_DUR    0.002914
236         V2_QRSA    0.002419
```

IV. Discussion

Fig. 6. Identical Precision Recall Curve For All Models



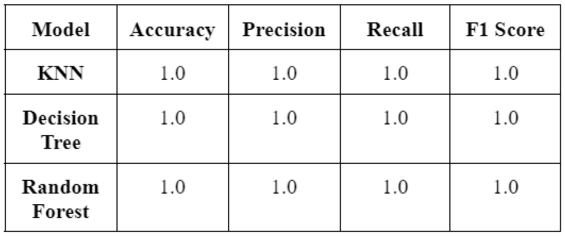
The K-Nearest Neighbor model outputted perfect results, where the precision, recall, and F1-score all totalled 1.0. No matter the number of neighbors, these scores remained perfect and the confusion matrix did not change as well. The confusion matrix values stayed consistent at 28 true positive values and 33 true negative values. The surprising part was that even when K was equal to 1 or when K was equal to any other number, the results stayed exactly the same. This indicates that some features in the dataset are extremely distinct and lead to being easily classified, especially in a binary classification of whether a patient has arrhythmia or not.

The decision tree model was perfectly split into two classifications, having arrhythmia and not having arrhythmia, based on 1 of the 278 features. This split was based on the variable ‘V6_NUM_DEFLEC’ and on the value of -.643. As explained in the results, this variable showed the number of intrinsic deflections of channel V6 in the heart. Channel

V6 indicates that this was the channel that was connected to the left part of the heart of a patient while measuring the ECG signals. The intrinsic deflections in this case mean the time period from the onset of the QRS complex to the peak of the R wave in an ECG which was shown in the ECG signal diagram in Fig. 1 above. This suggests that almost all the patients that had arrhythmia had an irregular pattern in their ECG signal between the starting of the Q wave to the peak of the R wave. That is why the Decision Tree model was able to perfectly split the dataset in two parts.

The Random Forest model allowed us to rank each feature by its importance. The output feature importance table showed that the variable used to split the decision tree, 'V6_NUM_DEFLEC' , was not the most important variable. WEIGHT was given more importance and V6_R_WIDTH was given the second highest importance tied with the V6_NUM_DEFLEC variable. This means if the variable V6_NUM_DEFLEC was removed from the dataset, the dataset would probably still be split perfectly because of these other important variables. This means that patients who had arrhythmia in the given dataset had an irregular pattern in the width of the R wave in their ECG signal and a certain weight that could be differentiated from patients who did not have arrhythmia. However, even though WEIGHT had a higher importance rank, V6_NUM_DEFLEC was still extremely important with an importance of .10 according to the Random Forest model which supported the results produced by the Decision Tree model. Overall in the feature importance table, out of all 278 features, only 18 features had an importance of greater than 0 and of those 18 features only 15 of those features had enough negligible weight to alter the model.

Fig. 7. Peformance Table of All Models



Model	Accuracy	Precision	Recall	F1 Score
KNN	1.0	1.0	1.0	1.0
Decision Tree	1.0	1.0	1.0	1.0
Random Forest	1.0	1.0	1.0	1.0

Overall, it was surprising to see such perfect results on our dataset in every model that was used. However, this was only possible because the given dataset was simple and extremely clean. It is obviously a very rare and unlikely case that a dataset is this clean in the real world but this helped confirm the results that all models produced as they delivered similar performances. The dataset was simplified from a multiclass classification with 16 classes to a binary classification. This allowed each of our models, Random Forest, Decision Tree and KNN, to perfectly classify whether or not each patient had arrhythmia. Across all 3 of the models, the precision recall curve was identical, with a horizontal line at 1.0 for precision. This was due to the fact that each model calculated the output with perfect accuracy and F1 score as seen in the table of the metric values produced by each model.

V. Conclusion

In conclusion, as seen by the findings and performance of all our 3 models, KNN, Decision Tree and the Random Forest, we were able to perfectly classify whether each patient had arrhythmia or not. Getting similar results over a number of models helped us confirm the results and make our findings and discussion more reliable. We were able to use binary classification of 1 and 0, for a normal ECG and different classes of arrhythmia respectively, thanks to the simplicity of the provided dataset. Moreover, we had identical precision recall scores, F1 scores and accuracy of 1.0 for all the models. This is a perfect example of a dataset being too clean, refined and simple where one single attribute, for example like the V6_NUM_DEFLEC, was able to perfectly classify each patient for having arrhythmia or not. Even though other studies showed that the incidence of arrhythmia was highly related to age, the Random Forest model suggested that a patient's weight held the most importance [3]. The RF model was able to distinguish the difference between the presence and absence of arrhythmia within a patient and narrowed down the attributes with importance to this classification from 278 different attributes to only 18. This could help researchers in the future to understand which attributes to pay attention to during screenings that best help determine arrhythmia and the reason behind a patient having it.

References

[1] Savalia, Shalin Emamian, Vahid. (2018). Cardiac Arrhythmia Classification by Multi-Layer Perceptron and Convolution Neural Networks. Bioengineering. 5. 35. 10.3390/bioengineering5020035. https://www.researchgate.net/publication/324988448_Cardiac_Arrhythmia_Classification_by_Multi-Layer_Perceptron_and_Convolution_Neural_Networks/download

[2] Heart arrhythmia - Diagnosis and treatment - Mayo Clinic. (2021, October 1). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/heart-arrhythmia/diagnosis-treatment/drc-20350674

[3] CV Physiology — Arrhythmias. (2012, November 10). Cvphysiology. https://www.cvphysiology.com/Arrhythmias/A008:

[4] Luz, E. J. da S., Schwartz, W. R., Cámara-Chávez, G., Menotti, D. (2015, December 30). ECG-based Heartbeat Classification for Arrhythmia Detection: A Survey. Computer Methods and Programs in Biomedicine. Retrieved April 26, 2022, from https://www.sciencedirect.com/science/article/pii/S0169260715003314

[5] Guvenir, H. and Acar, B. and Muderrisoglu, H. (2019). UCI Machine Learning Repository Arrhythmia Dataset [https://archive.ics.uci.edu/ml/datasets/Arrhythmia]. Irvine, CA: University of California, School of Information and Computer Science.

[6] Supervised learning. (2022). Scikit-Learn. https://scikit-learn.org/stable/supervised_learning.html

[7] 5 reasons why a heart arrhythmia can be so dangerous. Eastern Idaho Regional Medical Center (EIRMC). (n.d.). Retrieved May 3, 2022, from https://eirmc.com/blog/entry/5-reasons-why-a-heart-arrhythmia-can-be-so-dangerous

[8] Mayo Foundation for Medical Education and Research. (2022, April 30). Heart arrhythmia. Mayo Clinic. Retrieved May 3, 2022, from https://www.mayoclinic.org/diseases-conditions/heart-arrhythmia/symptoms-causes/syc-20350668: :text=A