Heart Rhythms Revealed: Exploring Arrhythmia via Classification and Clustering

Contents

1 Introduction

The heart is the most important organ in the circulatory system and serves as the vital pump of our body, ensuring the continuous circulation of blood. It is comprised of four chambers, two atria for receiving blood and two ventricles for pumping it out, connected through a system of valves that ensure blood flows in the correct direction.

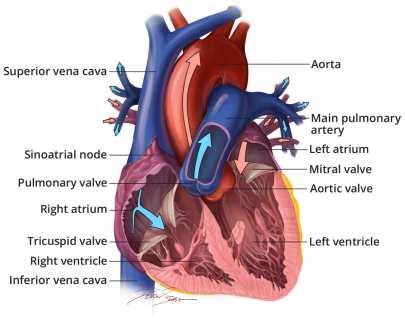

Figure 1: Anatomy of the heart
Source by: `https://www.nhlbi.nih.gov`

Each heartbeat consists of an electrical signal originated by a group of cells, also known as natural pacemaker, located in the sinoatrial (SA) node in the right atrium. This first signal travels firstly through the atria, causing them to contract and push blood into the ventricles, and then passes through the atrioventricular (AV) node, which briefly delays it, and finally transmits to the ventricles, causing them to contract and pump blood out to the body. This process repeats in a coordinated cycle, resulting in the heartbeat.

The propagation of electrical impulse of a heartbeat is tightly connected to the process of depolarization and repolarization. The former occurs when the cardiac cells receive signals from the SA node, causing a shift in electrical charge that triggers contraction. Following contraction, repolarization begins, restoring the cell's electrical charge to its resting state. This sequential depolarization and repolarization process propagates through the heart, coordinating its rhythmic contractions and ensuring efficient blood circulation.

Recording a sequence of impulses creates the distinct waveform observed in an electrocardiogram (ECG), reflecting the regular rhythm of the muscular contractions of the heart. Different types of ECGs, such as the resting ECG or the exercise stress test, can be recorded under different conditions in order to address specific purposes.

Moreover, an ECG is typically presented in a decomposed way. The most common is the 12-lead ECG (Figure 2), which is divided into 12 channels. Hence, each of the above presented parameters can be computed for each channel, which represents a specific angle of view or perspective of the heart's electrical activity and offers a unique perspective on the heart's depolarization and repolarization processes, allowing clinicians to detect abnormalities such as arrhythmias.

The term arrhythmia refers to an irregular heartbeat, where the heart may beat too fast (tachycardia), too slow (bradycardia), or irregularly. Early detection of arrhythmias is crucial due to their potential to trigger a series of cardiovascular events, including the serious risk of Sudden Cardiac Death (SCD) [3]. Overall, detecting arrhythmias is essential for early intervention, proper management, and prevention of complications, ultimately improving patient outcomes and quality of life. The complexity of interpreting ECG patterns is exacerbated by demographic factors such as age, gender, weight, and height, as well as underlying health conditions. Therefore, the potential impact of precise and automated machine learning algorithms cannot be overstated. These advancements offer the possibility of transforming cardiac care by facilitating continuous real-time monitoring of patients using devices capable of analyzing ECG data.

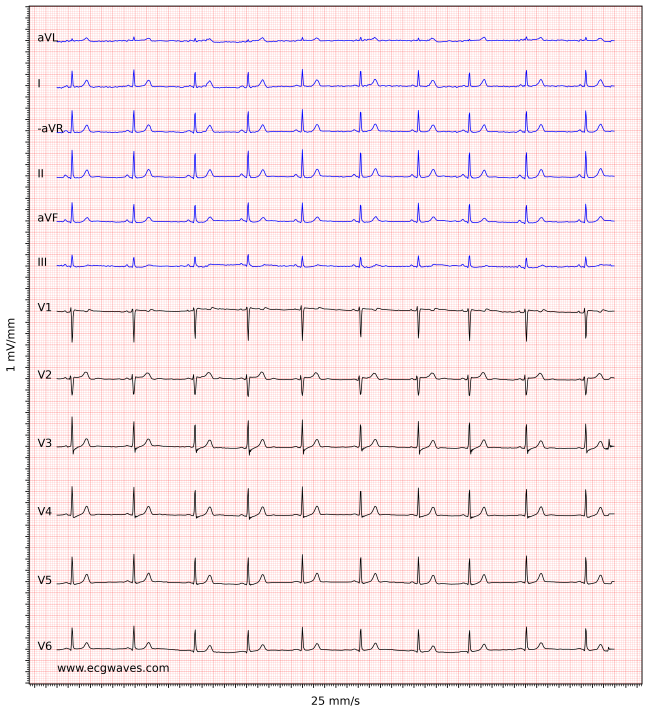

Figure 2: Example of a 12-lead ECG that shows a normal sinus rhythm from
Source by: `https://ecgwaves.com`

2 Dataset

The Arrhythmia dataset is a collection of medical data from UCI repository aimed at facilitating the detection and classification of various types of cardiac arrhythmias. It includes attributes such as age, sex, weight, height and a series of ECG measurements for each of the 12 channels and 452 istances.

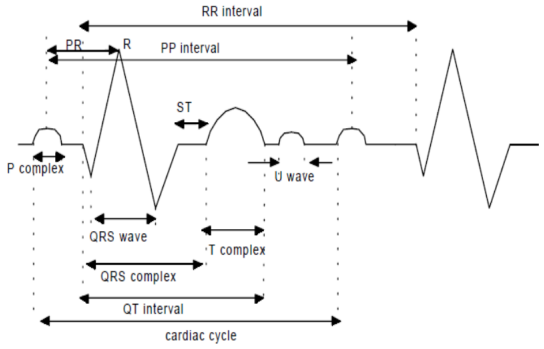

Figure 3: Main parameters of an ECG
Source by: [4]

Indeed, from the intricate pattern of an ECG, several key parameters are derived (Figure 3), providing valuable insights into cardiac function. Firstly, the ECG waveforms serve as primary indicators of cardiac activity. The P wave signifies atrial depolarization, while the QRS complex represents ventricular depolarization, and the T wave denotes ventricular repolarization. Secondly, intervals measured on the ECG offer critical information about the timing of various cardiac events.

The PR interval reflects the time from atrial to ventricular depolarization, while the QT interval regards ventricular depolarization and repolarization. Additionally, the QRS duration measures the time needed for ventricular depolarization. Furthermore, segments on the ECG tracing provide further insights into specific phases of the cardiac cycle. Heart rate, a vital parameter, is derived from intervals such as the RR interval, representing the time between two consecutive R waves. Lastly, the QRS axis represents the overall direction of ventricular depolarization, providing information about the heart's structural integrity and potential abnormalities.

The dataset is labeled, so we can see in figure 4 that the distribution of the normal (absence of arrhythmias) and abnormal (presence of arrhythmias) ECGs is quite balanced.

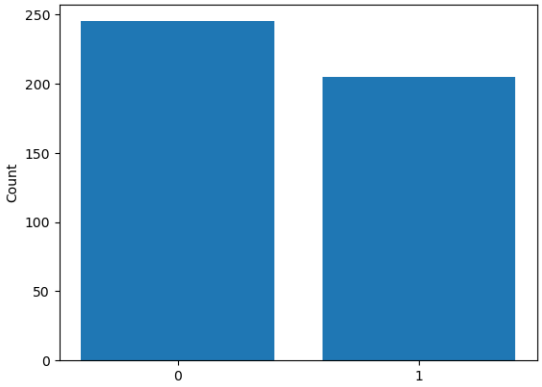

Figure 4: Distribution of the normal ECGs vs abnormal ones

Zooming in the abnormal classified ECG, we have 13 classes of arrhythmias, whose distribution is showed in figure 5.

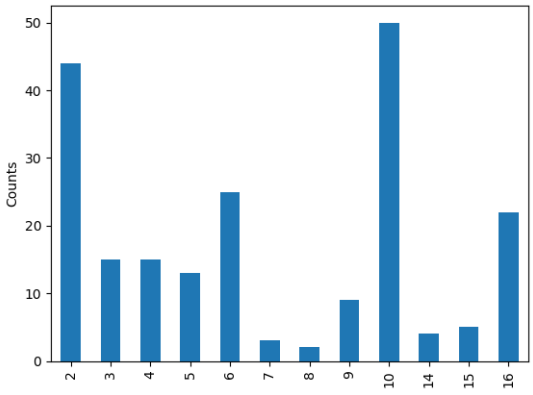

Figure 5: Distribution of the abnormal class

3 Classification

Given the considerable number of features in the dataset, dimensionality reduction is considered to enhance the models performances without loss of information. We opted to perform a Principal Component Analysis (PCA), selecting the number of principal components necessary to capture 95% of the total variance in the data.

PCA is performed after splitting the data into training (80%) and test (20%) sets and leads to the choice of 94 principal components. This ensures that PCA is not influenced by information in the test set, which could lead to data leakage and overfitting.

We are going to consider a two-class classification problem: presence (y=1) and absence (y=0) of cardiac arryhthmia. Six classification algorithms are considered: Logistic Regression (LR), Bayesian Logistic Regression (BLR), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), K-Nearest Neighbors (KNN) and Random Forest (RF).

The accuracies and the values of the AUC achieved on the test set by each model are reported in Table 1:

Table 1: Test accuracies

	LR	BLR	LDA	QDA	RF	KNN
Accuracy	73.3%	72.2%	74.4%	67.8%	72.2%	65.6%
AUC	0.73	0.81	0.74	0.67	0.72	0.65

The LDA method achieves the highest accuracy, while BLR performs best in terms of AUC. Additionally, the ROC curve in the figure 6 visually illustrates the comparison between the AUC values.

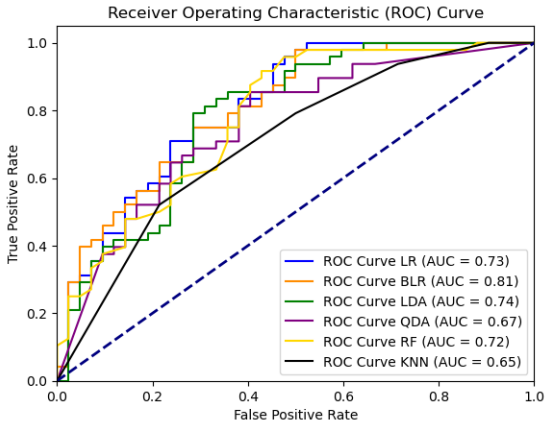

Figure 6: Plot of the ROC curves

To gain deeper insights into the models' performances, examining the confusion matrices and calculating sensitivities is valuable. In the context of medical diagnosis, a false negative—wherein the model incorrectly predicts the absence of an arrhythmia when it is actually present—is more dangerous than a false positive. Indeed, this type of error poses a greater risk to the patient's well-being as it could potentially lead to undetected health issues or delayed treatment. Upon scrutinizing all six confusion matrices (figure 7), it is evident that the models exhibit challenges in accurately identifying true instances of arrhythmias, as indicated by the presence of false negatives across the board. This collective struggle highlights a critical area of concern that necessitates further investigation and refinement in the models' performance, particularly with regard to minimizing the occurrence of false negatives to enhance the reliability and effectiveness of arrhythmia detection in clinical settings.

To see which model performs better in these terms, specificity can be computed. Sensitivity, also known as recall, quantifies the proportion of actual positive cases that are correctly identified by a classification model. In the context of medical diagnosis, such as detecting arrhythmias, sensitivity indicates the model's ability to correctly identify patients with the condition (true positives) out of all individuals who actually have the condition (true positives + false negatives).

In simpler terms, sensitivity answers the question: "Of all the individuals who truly have arrhythmias, how many did the model correctly identify?".

Referring to table 2, it is apparent that, except for KNN, all models exhibit similar levels of sensitivity in accurately detecting true arrhythmia cases. However, there remains room for improvement

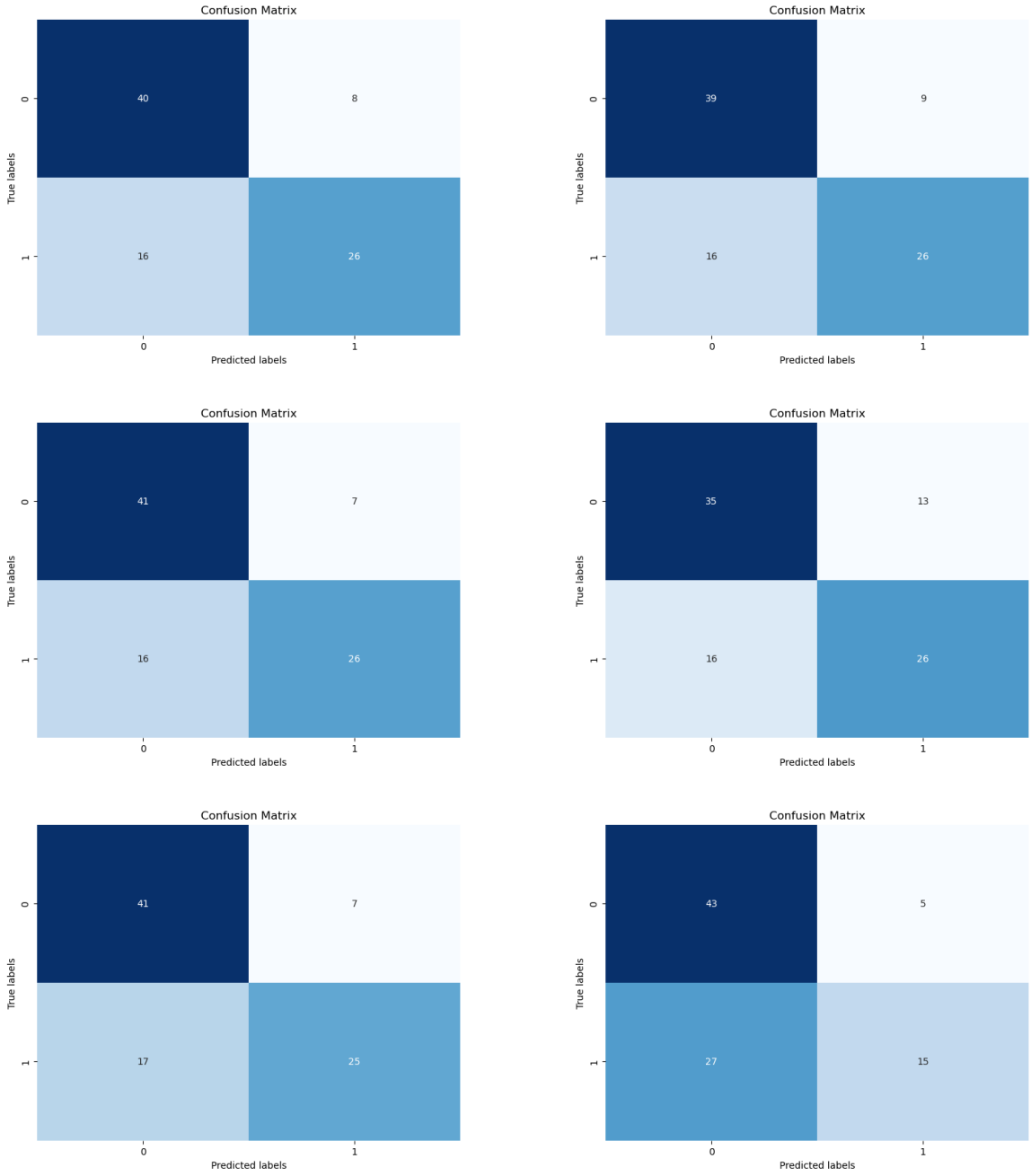

Figure 7: Confusion matrices for each model

Table 2: Sensitivity

	LR	BLR	LDA	QDA	RF	KNN
Sensitivity	0.62	0.62	0.62	0.62	0.60	0.36

across all models, indicating the potential for enhancing their performance in this aspect.

4 Cluster analysis

In this section, we are employing a statistical model called Gaussian Mixture Model (GMM). We are using demographic data along with features related to the PQRST complex from ECG signals. We are excluding consideration of the parameters specific to the 12 channels of the ECG. Indeed, the relevant diagnostic information in the ECG signals is usually represented via PQRST complex features, which include the wave location, duration, amplitudes and shapes [5].

Our objective here is twofold: first, to group individuals into homogeneous clusters based on the selected variables, and second, to capture any discernible patterns within these clusters that might aid in identifying different types of arrhythmia.

The algorithm determined that the optimal number of groups is 14, as indicated by the lowest Bayesian Information Criterion (BIC). When visualized in Figure 10, it is evident that these clusters lack clear separation when plotted against pairs of variables. This highlights the difficulties involved in diagnosing arrhythmias. It emphasizes the requirement for distinct channels and specific criteria tailored to each class of arrhythmia. However, a thorough analysis can reveal certain patterns more clearly. In Figure 8, the yellow group exhibits distinct separation from the other groups, characterized by a PR interval of 0. This strongly indicates that this group represents individuals with atrial fibrillation [1].

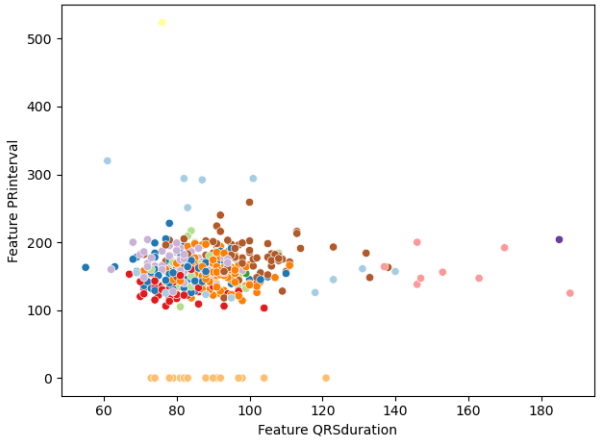

Figure 8: Plot of the clusters against QRS duration and PR interval

Figure 9 clearly identifies the dark green group, distinguished by a QRS duration exceeding 120 milliseconds. This criterion is commonly linked with left and right branch blocks [2].

5 Limitation and further reasearch

The Arrhythmia dataset, while comprehensive, may not fully capture the complexity and diversity of real-world arrhythmia cases. The dataset's relatively small size (452 instances) could limit the generalizability of our findings. Additionally, the absence of certain demographic and clinical variables may impact the robustness of the models.

Our classification models, though performing reasonably well, still exhibit limitations in accurately detecting true instances of arrhythmias, particularly in minimizing false negatives. Further refinement of these models, potentially through the incorporation of additional features or the use of more advanced machine learning techniques, may improve their performance.

Similarly, while our clustering analysis provides insights into grouping individuals based on selected variables, the lack of clear separation between clusters highlights the challenges in diagnosing arrhythmias solely based on demographic and ECG features. Future research could explore alternative clustering algorithms or incorporate additional data sources to enhance cluster identification and interpretation.

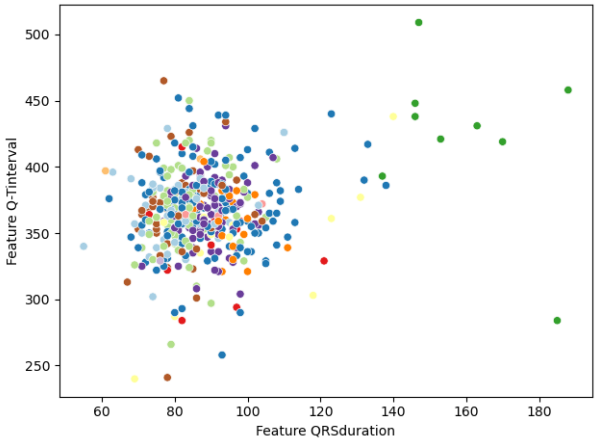

Figure 9: Plot of the clusters against QRS duration and Q-Tinterval

Furthermore, deploying machine learning algorithms for arrhythmia detection and classification in real-world healthcare settings requires careful validation and integration into existing clinical workflows. Moreover, the interpretation of algorithmic predictions alongside clinical expertise is essential to ensure accurate diagnosis and appropriate patient management.

6 Conclusion

In conclusion, this study investigates the intricate realm of arrhythmia detection and classification using machine learning techniques. Through an in-depth analysis of ECG data and demographic variables, we aimed to enhance the understanding of cardiac arrhythmias and develop effective models for their identification.

Our findings demonstrate the potential of machine learning algorithms in accurately classifying arrhythmias and grouping individuals based on similar characteristics. Despite the challenges posed by the complexity and diversity of arrhythmia patterns, models achieved respectable performance metrics, highlighting their utility in assisting healthcare professionals in early detection and diagnosis.

Additionally, it is important to keep in mind that the integration of machine learning algorithms into clinical practice requires careful validation and collaboration between data scientists and healthcare professionals to ensure accurate diagnosis and improved patient outcomes.

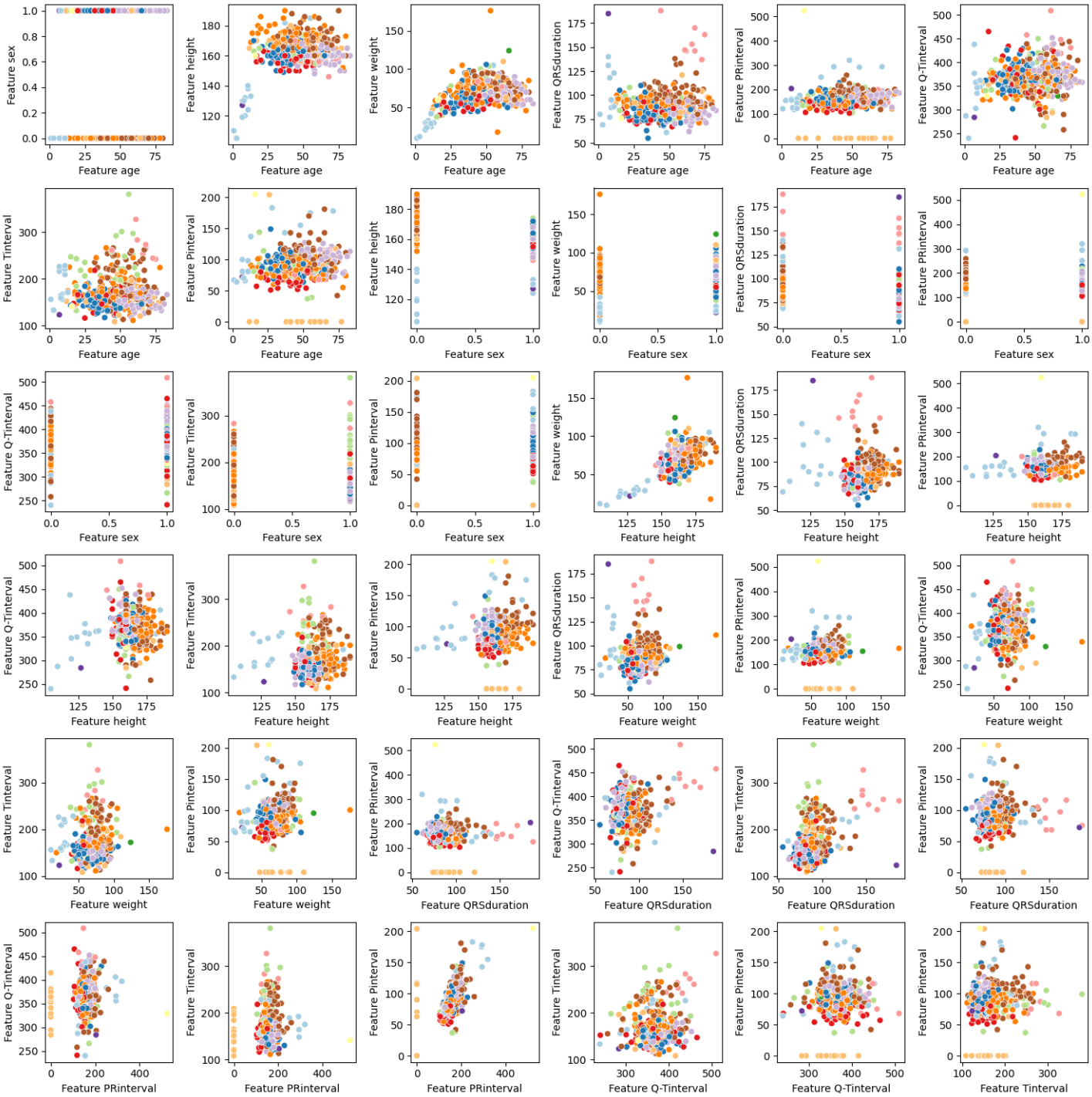

Figure 10: Plot of the clusters for each pair of variables

References

[1] A. L. Goldberger, Z. D. Goldberger, and A. Shvilkin. Chapter 23 - interpreting ecgs: An integrative approach. In A. L. Goldberger, Z. D. Goldberger, and A. Shvilkin, editors, *Goldberger's Clinical Electrocardiography (Ninth Edition)*, pages 240–246. Elsevier, ninth edition edition, 2018.

[2] R. S. Mane, A. Cheeran, V. D. Awandekar, and P. Rani. Cardiac arrhythmia detection by ecg feature extraction. *Int. J. Eng. Res. Appl*, 3(2):327–332, 2013.

[3] S. Sahoo, M. Dash, S. Behera, and S. Sabut. Machine learning approach to detect cardiac arrhythmias in ecg signals: A survey. *IRBM*, 41(4):185–194, 2020.

[4] S. Samad, S. A. Khan, A. Haq, and A. Riaz. Classification of arrhythmia. *International Journal of Electrical Energy*, 2(1):57–61, 2014.

[5] Y.-C. Yeh, C. W. Chiou, and H.-J. Lin. Analyzing ecg for cardiac arrhythmia using cluster analysis. *Expert Systems with Applications*, 39(1):1000–1010, 2012.