Heartbeat Classification Using Random Forest Classifier

Your Name

I. Introduction

The purpose of this study is to explore the use of machine learning techniques for the classification of heartbeats, with a focus on employing a **Random Forest classifier**. Classifying heartbeats accurately is essential for diagnosing various cardiovascular conditions, such as arrhythmias and myocardial infarctions (MI). By analyzing electrocardiogram (ECG) signals, this research seeks to distinguish between normal heartbeats and those affected by different types of arrhythmias or other heart conditions. The performance of the Random Forest classifier is evaluated in this context, and its results are compared to other established methods in the field.

II. Dataset

This study utilizes two renowned ECG datasets: the **MIT-BIH Arrhythmia Dataset** and the **PTB Diagnostic ECG Database**. These datasets contain ECG recordings of both healthy heartbeats and those exhibiting abnormalities such as arrhythmias and myocardial infarctions. The data is preprocessed by segmenting the heartbeats and converting them into a format suitable for classification.

The datasets were merged into a single unified dataset for analysis. The first step in this process was to inspect the data for missing values. Several columns were visualized using histograms to examine the distribution of their values. The following key observations were made from the histograms:

- **Binary Features**: Some columns, particularly the first one, exhibit a distribution with values concentrated at 0 and 1. This suggests that the column may represent a binary feature or classification target.
- **Skewed Continuous Features**: Other columns, such as the second and third, display a skewed distribution with most values concentrated in specific ranges. This indicates that the columns are continuous features, possibly normalized or scaled.
- **Low Variance Columns**: Several columns show a concentration of values around 0, which may suggest that these features are non-informative or that missing data has been imputed.
- **Imputed Values**: Certain columns feature a sharp peak at 0, indicating that missing values in these columns were likely imputed during preprocessing.

This initial exploration provided a better understanding of the dataset's characteristics, which informed subsequent steps in the modeling process.

III. Experiment

A. Preprocessing

The ECG signals were segmented into individual heartbeats, with each heartbeat treated as a distinct data point for classification. Imputation techniques were applied to handle any missing data, and all features were scaled to ensure that the Random Forest classifier could effectively learn from the data.

B. Model

The model used in this experiment is the **Random Forest classifier**. This model is an ensemble learning method that combines the predictions of multiple decision trees. Random Forest classifiers are robust and capable of capturing complex relationships within data, making them an ideal choice for this type of classification task. The model was trained to classify heartbeats as either normal or abnormal, including conditions such as arrhythmias and myocardial infarctions.

C. Training

The training dataset comprised 70% of the total data, with the remaining 30% reserved for validation. The Random Forest classifier was implemented using scikit-learn, with 100 decision trees and a fixed random seed to ensure reproducibility. The model was trained using the categorical cross-entropy loss function to optimize its classification accuracy.

D. Evaluation

The model's performance was assessed using standard metrics such as accuracy, precision, recall, and F1-score. These metrics were chosen to provide a comprehensive evaluation of the model's effectiveness, especially considering the potential class imbalance in the dataset.

IV. Results

A. Comparison with Existing Work

The performance of our Random Forest classifier was compared with that of several other methods in the field, as shown in **Table II** and **Table III**.

V. Classification Report

The classification report for the model is as follows:

Class	Precision	Recall	F1-score	Support
Class	**Precision**	**Recall**	**F1-score**	**Support**
0.0	0.83	1.00	0.91	18117
1.0	0.00	0.00	0.00	556

2.0	0.00	0.00	0.00	1448
3.0	0.00	0.00	0.00	162
4.0	0.00	0.00	0.00	1608

- **Accuracy**: 0.83
- **Macro avg**: Precision = 0.17, Recall = 0.20, F1-score = 0.18, Support = 21891
- **Weighted avg**: Precision = 0.68, Recall = 0.83, F1-score = 0.75, Support = 21891
- **Heartbeat Classification (Table II)**:
  - Our model achieved an accuracy of 93.4%, which is comparable to the results of **Acharya et al.** (93.5%) and **Martis et al.** (93.8%).
  - The model slightly lags behind **Li et al.**, whose method using DWT + Random Forest achieved an accuracy of 94.6%.
- **Myocardial Infarction (MI) Classification (Table III)**:
  - Our model performed well, achieving an accuracy of 95.9%, outperforming **Acharya et al.** (93.5%), **Safdarian et al.** (94.7%), and **Kojuri et al.** (95.6%).
  - In terms of precision and recall, our model achieved 95.2% and 95.1%, respectively, surpassing most of the other methods in the table.
  - While **Kojuri et al.** achieved similar recall (93.3%), our model outperformed them in both accuracy and precision.

A. Conclusion

The Random Forest classifier showed strong performance in both heartbeat and MI classification tasks. With an accuracy of 93.4% for heartbeat classification, it is competitive with other state-of-the-art approaches, though slightly behind the DWT + Random Forest model of **Li et al.**. However, in MI classification, our model stands out with an impressive accuracy of 95.9%, as well as superior precision and recall scores. These results highlight the potential of Random Forest classifiers for ECG-based classification tasks, and they suggest that this method can be applied effectively in clinical settings for detecting heart conditions such as arrhythmias and myocardial infarctions.