Practical Work 1: ECG Categorization

BI12-077 Hoang Ha Dang

March 6, 2024

1 Introduction

The electrocardiogram (ECG) is a record of the dynamic changes of the human heart's electrical activity that can be monitored. Base on ECG results, it can be used to diagnose cardiovascular diseases. Early diagnose by using ECG could be essential and helpful for reducing heart disease issues. For this practice, a supervised machine learning model was applied for this classification task which is Random Forest (RF).

2 Background

In this practice, we used datasets from the MIT-BIH Arrhythmia Dataset. This dataset has been used in exploring heartbeat classification using deep neural network architectures, and observing some of the capabilities of transfer learning on it. The signals correspond to electrocardiogram (ECG) shapes of heartbeats for the normal case and the cases affected by different arrhythmias and myocardial infarction. These signals are preprocessed and segmented, with each segment corresponding to a heartbeat.

This dataset contains ECG recordings at the sample rate of 125Hz and 5 categories which are labeled in N, S, F, V, Q.

Label	Description
N	Non-ecotic beats
S	Supraventricular ectopic beats
V	Ventricular ectopic beats
F	Fusion beats
Q	Unknown beats

Table 1: Label Description.

3 Method

For this classification task, it consists of two main steps: data preprocessing and classification. Because the dataset is imbalanced. Therefore, a random oversampling method is used to address this issue. Random oversampling is a technique for balancing imbalanced datasets by duplicating examples from both classes, which allows for effective insights to be drawn from the training sets.

In this practice, Random Forest for this. A Random Forest model is a machine learning algorithm that combines the output of multiple decision trees to reach a single result. By following the large number principle, a random forest classifier is considered to reduce the error and give better performance, compared to a single classifier.

The dataset was splitted into 2 sets: train set, test set with the ratio is 80 percent for train set and 20 percent for test set. The train set was taken into RF model then was trained with a set of number of trees parameter values: 10, 20, 50. The best parameter value based on the accuracy is 20.

4 Evaluation

The results of the RF model is summarized in Table 2. The assessment measures offered insights regarding the precision, sensitivity, and overall accuracy of the model.

Metric	Value
Accuracy	0.8264
Precision	0.8888
Recall	0.8251
F1 Score	0.8300

Table 2: Evaluation Results

5 Conclusion

From this practice,the efficiency of Random Forest Classification in categorizing heart conditions based on an ECG dataset. The encouraging findings indicate that machine learning models, especially ensemble techniques such as Random Forest, have the potential to make substantial contributions to the field of cardiology by automating and enhancing the accuracy of disease diagnosis.