[R]*Baltic Security Foundation* [L]

# Report for Labwork 1

Tran Trung Kien
*University of Science and Technology*

**Abstract**
An electrocardiogram (ECG or EKG) is a test to record the electrical signals in the heart. It shows the beating pace of your heart. ECG test results can help doctors diagnose irregular heartbeats (called arrhythmias),a previous heart attack or the cause of chest pain. MIT-BIH Arrhythmia Database is a result of the collaboration between Boston's Beth Israel Hospital and Massachusetts Institute of Technology. The database was the first generally available set of standard test material for evaluation of arrhythmia detectors. <br> In this labwork, the goal is to classification the heart's stage into 5 categories: - 0: Normal Beat - 1: Supraventricular Premature Beat - 2: Ventricular Premature Beat - 3: Fusion Beat - 4: Unknown Beat

## Contents

# 1 Dataset

In this section, we will discuss the dataset used for ECG heartbeat classification and details about the preprocessing steps applied to prepare the data for modeling

## 1.1 Data exploration

The MIT-BIH Arrhythmia Dataset consists of 109,446 samples, each represents an electrocardiogram (ECG) signal segment recorded at 187 time steps with a sampling frequency of 125 Hz. The dataset includes both normal heartbeats and heartbeats affected by various arrhythmias and myocardial infarction.

Each segment covers approximately 1.496 seconds, which is enough to capture a complete heartbeat cycle, given that a normal heart rate ranges between 60 and 100 beats per minute (BPM). A full cardiac cycle lasts about 0.6 to 1 second, the chosen frequency ensures that each heartbeat is well-represented across different heart rates.

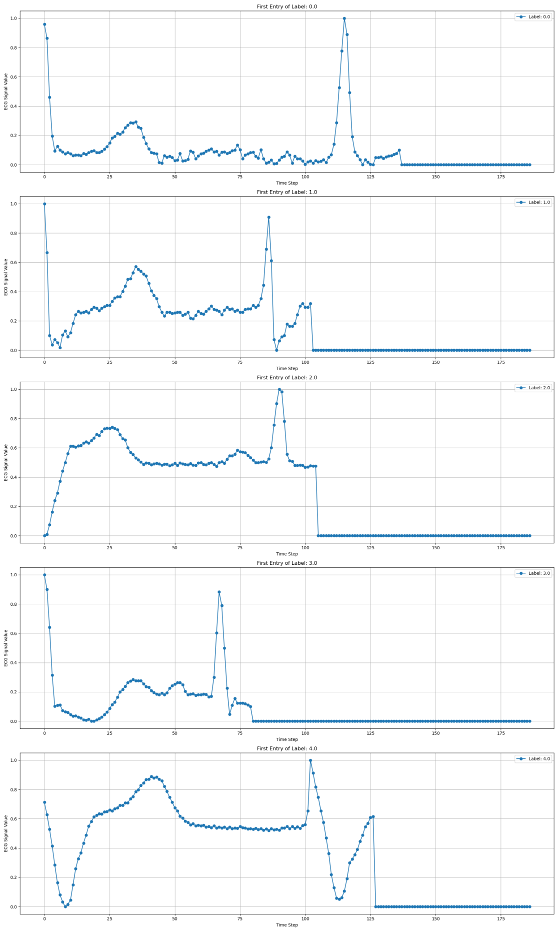

Figure 1: Class presentative graph

The graph above shows the presentative ECG signal for each class. There are differences in beat duration, peak positions and the shape of peak, highlights that there is distinct traits to each class. Some classes exhibit sharp, isolated peaks, while others show more gradual changes.

The data distribution of this dataset is represented by the plot below. The dataset shows a highly imbalanced class

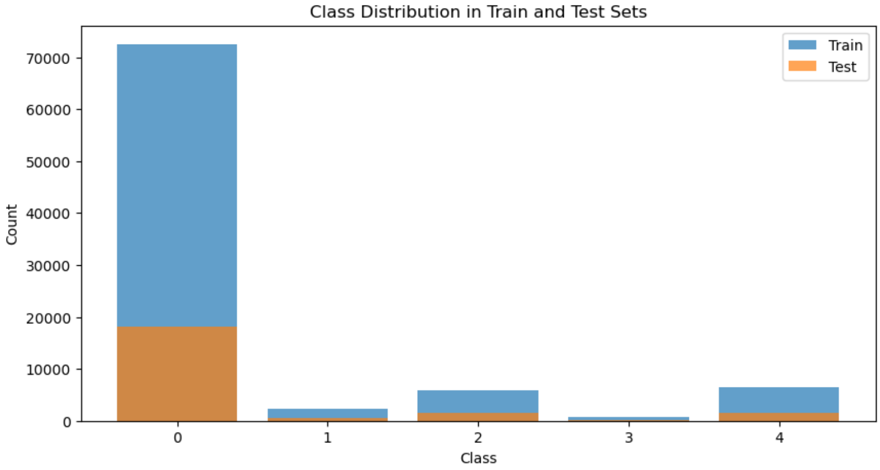

Figure 2: Class distribution graph

distribution with class 0 makes up 82 percent of the dataset with more than 90,000 entries. Class 3 is the smallest with only 803 entries, significantly lower than other classes. Such an imbalance could lead to model bias, where the classifier tends to favor the majority class while under-performing on minority classes.

## 1.2 Data processing

To resolve the problem mentioned above, multiple resampling techniques are employed. For the largest class (class 0), **Random Under-sampling** is used to reduce the number of entries to 20,000. The other classes are divided into two groups: the minority class (classes 2 and 4) and the extremely minority class(classes 1 and 3). For the first group, **SMOTE** (Synthetic Minority Oversampling Technique) generates synthetic samples by creating new points along the line between a selected sample and its **k** nearest neighbors. For the second group, **ADASYN** (Adaptive Synthetic Sampling) is used to generate more synthetic samples for harder-to-learn class instances instead of generating samples uniformly. Since the number of entries in the second group is low (2,779 and 803 for classes 1 and 3, respectively), **ADASYN** prioritizes regions where the classifier struggles. Additionally, to further refine the dataset and improve class separation, **Tomek Links** are applied after oversampling. Once identified, the majority-class sample in the Tomek Link is **removed**, reducing class overlap and making the decision boundary clearer. This helps eliminate borderline noise and ensures a better balance between classes, improving the classifier's performance. Class Distribution after resampling process:

- Class 0.0: 19,992
- Class 1.0: 15,188

- Class 2.0: 19,994
- Class 3.0: 9,970
- Class 4.0: 20,000

# 2 Model

The **Random Forest Classifier** is initialized with default hyperparameters, including 100 decision trees ($n\_estimators = 100$), allowing unrestricted tree depth ($max\_depth =$ None), and setting $min\_samples\_split = 2$ and $min\_samples\_leaf = 1$ to ensure that the trees grow fully. The $random\_state = 42$ ensures reproducibility, while $n\_jobs = -1$ allows the model to utilize all available CPU cores for parallel computation, improving efficiency. The training period is 5 minutes and 43 seconds.

# 3 Result

## 3.1 Metrics

The model was evaluated on 21,891 instances, achieving the overall accuracy of 97.43%. While this suggests strong predictive performance, the dataset exhibits significant class imbalance: class 0.0 (18,117 instances) dominates, while classes 1.0 (556 instances), 2.0 (1,448 instances), 3.0 (162 instances), and 4.0 (1,608 instances)
**Evaluation metrics:**

- **Class 0.0** (Majority): Precision = 0.97, Recall = 1.00, F1-score = 0.99
- **Class 1.0**: Precision = 0.99, Recall = 0.60, F1-score = 0.74
- **Class 2.0**: Precision = 0.98, Recall = 0.88, F1-score = 0.93
- **Class 3.0**: Precision = 0.88, Recall = 0.62, F1-score = 0.73
- **Class 4.0**: Precision = 1.00, Recall = 0.94, F1-score = 0.97

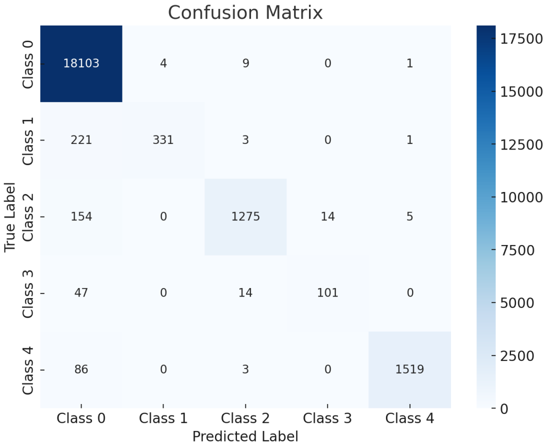

Figure 3: Class presentative graph

The confusion matrix reveals a tendency for misclassifications to favor the majority class (0.0). For example, 221 of 556 class 1.0 instances and 47 of 162 class 3.0 instances are incorrectly predicted as class 0.0. This bias underscores the impact of class imbalance on model predictions.

## 3.2 Comparison

The proposed model in the paper reported an accuracy of 95.9%, a precision of 95.2%, and a recall of 95.1%. Comparing that with the performance of Random Forest (accuracy of 97.43%, with a weighted average precision and recall of 97% each),the Random Forest model demonstrates superior performance.

The paper's model achieved an accuracy of 95% on a larger, more complex dataset and a wider range of features, suggesting robustness in handling intricate patterns. In comparison, the Random Forest model achieved a higher accuracy but on a smaller, simpler dataset, which may limit its generalization.