# 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

## 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.

The data distribution of this dataset is represented by the plot. The dataset shows a highly imbalanced class 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 underperforming 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 Undersampling** is used to reduce the number of

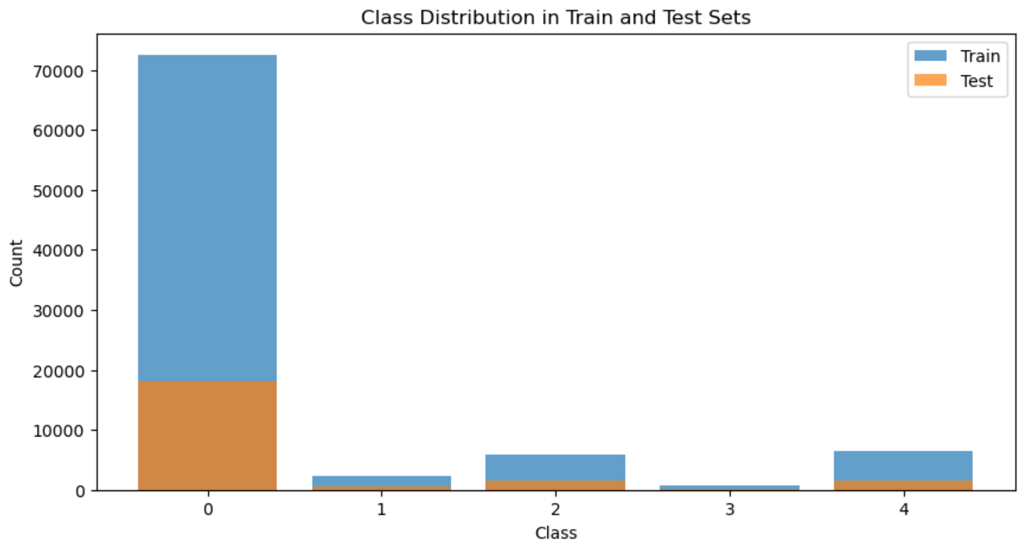


Figure 1: Class distribution graph

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 hyper-parameters, 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 Random Forest Classifier achieved an overall accuracy of 0.97 on the test dataset, below is a detailed analysis of the classification metrics:

- **Accuracy** (0.9743): The model correctly classified approximately 97.43 percent of the total test samples.

- **Precision**: The precision scores are 0.97(class 0.0),0.99(class 1.0), 0.98(class 2.0), 0.88 (class 3.0), and 1.00 (class 4.0), indicating that most classes are well-classified, with minimal false positives.

- **Recall** is high across most classes( 1.0 for class 0.0, 0.88 for class 2.0 and 0.94 for class 4.0), class 1.0 (0.60) and class 3.0 (0.62) having lower recall values, suggesting some mis-classifications in these categories.

- The **weighted F1-score** is 0.97, indicating strong overall classification performance

## 3.2 Comparison

The performance in given paper is **Accuracy:** 95.1%, **Precision:** 95.2%, **Recall:** 95.9%. These metrics in work are 97.43%, 96% and 81% respectively. My model outperforms the reference work in overall accuracy and precision. However our recall is 81.0% vs 95.1% in the reference work, meaning our model struggles with false negatives, especially in classes 1.0 and 3.0, where misclassification rates are higher. Given the different in size of input data, the original paper works on the raw MIT-BIH Arrhythmia Database, whereas our model is trained on a processed version of the dataset. The reference work's model appears to generalize better across smaller classes, due to a deeper and more comprehensive feature extraction process.