# Machine Learning in Medicine

Report 1

*Student Name - ID* Dinh Tuan Kiet - 22BI13229
*Lecturer:* Prof. Tran Giang Son

## 1 Introduction

Machine learning in medicine has the potential to transform patient outcomes, improve quality of care, streamline healthcare delivery, and more importantly, make healthcare more accessible and cost-effective. In this report, we will explore and analyze the dataset **ECG Heartbeat Categorization**, which contains the heart rate of the entire patient, using 2 machine learning models such as **Random Forest** and **Support Vector Machine** to classify the patients. 5 different heart rhythm types: 0 normal, 1 supraventricular, 2 ventricular, 3 mixed, 4 unknown.

## 2 Dataset

The ECG Heartbeat Categorization dataset is an important resource in cardiovascular disease research, compiled from two major sources. MIT-BIH Arrhythmia Dataset and PTB Diagnostic ECG Database. This dataset contains electrocardiogram (ECG) signals segmented into individual heartbeats, to aid in the classification and detection of arrhythmias. However, we will focus to the MIT-BIH dataset.

### 2.1 MIT-BIH Arrhythmia Dataset

The MIT-BIH dataset contains approximately 108,000 heartbeats recorded from 47 patients, categorized into five main classes:

- **0 - Normal Beat:** Common normal heartbeats (approximately 80-90% of the dataset).
- **1 - Supraventricular Ectopic Beat:** Irregular beats originating above the ventricles.
- **2 - Ventricular Ectopic Beat:** Abnormal beats generated in the ventricles.
- **3 - Fusion Beat:** A combination of normal and abnormal beats.
- **4 - Unclassifiable Beat:** Beats that do not fit into the predefined categories.

The data is split into a training set (*mitbih_train.csv*, containing around 87,000 heartbeats) and a test set (*mitbih_test.csv*, with approximately 21,000 heartbeats). Each heartbeat is represented by 186 numerical features extracted from the ECG signal and a corresponding label column indicating the class.

## 2.2 Data Characteristics and Challenges

All ECG signals in the dataset are preprocessed and normalized to a range of 0 to 1 to ensure consistency. However, one major challenge of this dataset is its **class imbalance**, where normal beats significantly outnumber abnormal ones. This imbalance complicates the classification of rare arrhythmic beats and requires techniques such as resampling, or data augmentation to improve model performance.

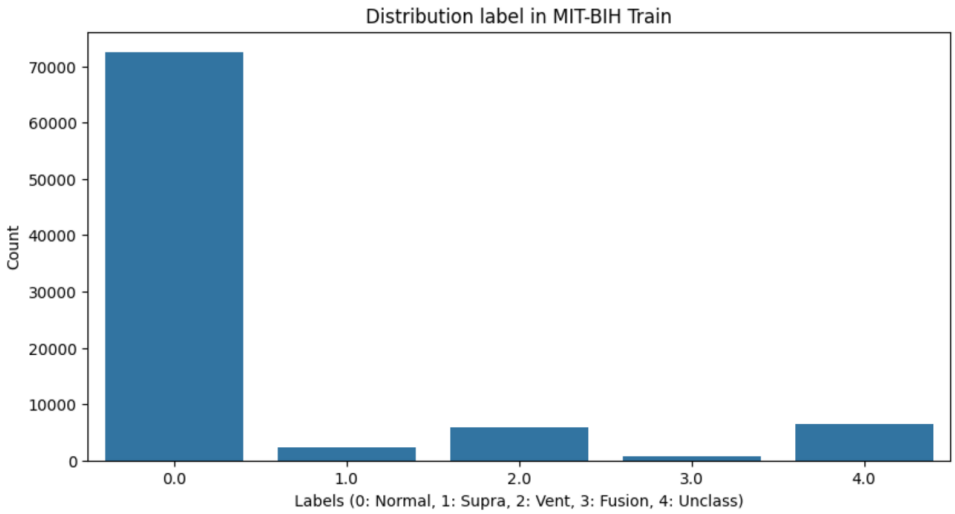


Figure 1: Distribution label in MIT-BIH training set

## 2.3 Data Preprocessing Technique: SMOTE

We use SMOTE to address the class imbalance in the MIT-BIH dataset, where normal heartbeats (label 0) dominate ( 80-90%) while abnormal heartbeats (labels 1, 2, 3, 4) approximate ( 10-20%). SMOTE generates synthetic samples for minority classes by interpolating between existing minority instances in the feature space, balancing the distribution across all five classes (0, 1, 2, 3, 4). This technique was applied exclusively to the training set to ensure a realistic test set evaluation.

# 3 Model Evaluation

We evaluated two machine learning models to classify ECG heartbeats: Random Forest (RF) and Support Vector Machine (SVM), both trained on the SMOTE-balanced training data and evaluated on the original test set.

### 3.1 Random Forest

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| 0.0 | 0.99 | 0.99 | 0.99 | 18118 |
| 1.0 | 0.87 | 0.77 | 0.82 | 556 |
| 2.0 | 0.97 | 0.93 | 0.95 | 1448 |
| 3.0 | 0.80 | 0.73 | 0.77 | 162 |
| 4.0 | 0.99 | 0.97 | 0.98 | 1608 |
| Accuracy | - | - | 0.98 | 21892 |
| Macro avg | 0.92 | 0.88 | 0.90 | 21892 |
| Weighted avg | 0.98 | 0.98 | 0.98 | 21892 |

Table 1: Classification Report for RF

### 3.2 Support Vector Machine

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| 0.0 | 0.99 | 0.94 | 0.96 | 18118 |
| 1.0 | 0.42 | 0.81 | 0.55 | 556 |
| 2.0 | 0.89 | 0.94 | 0.91 | 1448 |
| 3.0 | 0.31 | 0.90 | 0.46 | 162 |
| 4.0 | 0.97 | 0.98 | 0.97 | 1608 |
| Accuracy | - | - | 0.94 | 21892 |
| Macro avg | 0.71 | 0.91 | 0.77 | 21892 |
| Weighted avg | 0.96 | 0.94 | 0.95 | 21892 |

Table 2: Classification Report for SVM

## 4 Result

When comparing our models to the original paper's MIT-BIH results, the paper reported **95.9%** accuracy. Our Random Forest model achieved **98%** accuracy, outperforming the paper , especially for majority classes, but it struggles with minority ones. Our SVM model scored **93.9%** accuracy, due to lower precision but higher recall for rare heartbeats. Random Forest is better overall for accuracy, but SVM might need optimization, like SMOTE, to match the paper's balanced performance for detecting abnormalities.