Department of Information and Communication Technology

Machine Learning in Medicine

ECG Heartbeat Classification Report

ID - Student Name:
22BI13103 - Le Duc Dung - Data Science

Lecturer:
Prof. Tran Giang Son

Academic year: 2022 - 2025

Hanoi, February 2025

1 Introduction

This report presents Practical 1 of Machine Learning in Medicine, where a publicly available Kaggle dataset is analyzed to understand its structure and features before developing a classification model. The study involves a detailed description of the dataset, including its features, sources, and preprocessing steps. A machine learning approach is implemented to build a classification model such as the Support Vector Machine and its performance is thoroughly evaluated. The results are then compared with those reported in the original research paper to assess the effectiveness of the proposed model.

2 Dataset

The MIT-BIH Arrhythmia Dataset is a widely used benchmark for ECG classification tasks. It contains 187 features that describe heartbeat signals recorded at a 125 Hz sampling rate. The dataset categorizes heartbeats into five classes.. Normal, Atrial Premature, Premature Ventricular Contraction, Fusion of Ventricular and Normal, and Paced beats. It consists of 87,554 training samples and 21,892 testing samples but suffers from a class imbalance, with the Normal class being the most prevalent.

2.1 Arrhythmia Dataset

- **Number of Samples:** 109,446
- **Number of Categories:** 5
- **Sampling Frequency:** 125Hz
- **Data Source:** Physionet's MIT-BIH Arrhythmia Dataset
- **Classes:**
  - **N**: 0
  - **S**: 1
  - **V**: 2
  - **F**: 3
  - **Q**: 4

2.2 The PTB Diagnostic ECG Database

- **Number of Samples:** 14,552
- **Number of Categories:** 2
- **Sampling Frequency:** 125Hz
- **Data Source:** Physionet's PTB Diagnostic Database

2.3 Class Distribution

This dataset suffers from class imbalance with Normal class taking the highest distribution among others on both train-set and test-set.

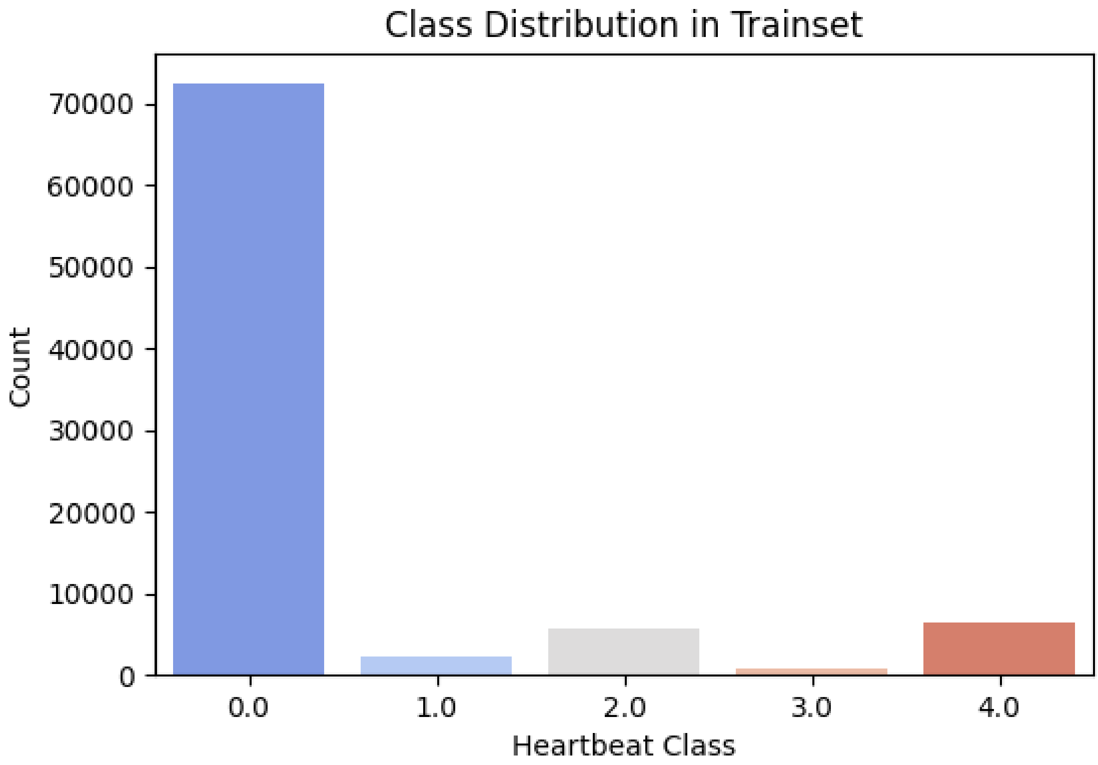

Figure 1: Train-set Label Distribution

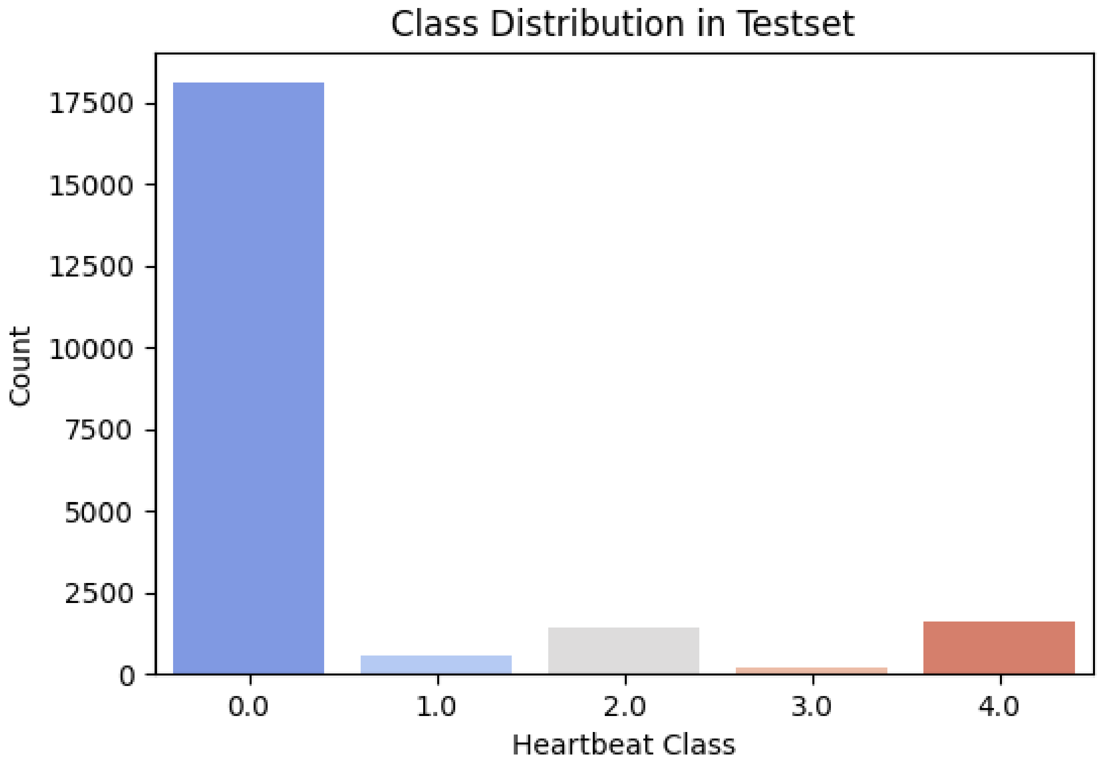

Figure 2: Test-set Label Distribution

3 Methodology

This section is about how the Support Vector Machine is trained with the preprocessed dataset.

3.1 Preprocessing

Since the training dataset is imbalanced, we use SMOTE (Synthetic Minority Oversampling Technique) from Imblearn Library to resample the class distribution for achieving better results.

$$x_{\text{new}} = x_{\text{current}} + \lambda \times (x_{\text{neighbor}} - x_{\text{current}}) \tag{1}$$

where λ is a random number between 0 and 1.

3.2 Model Architecture

Support Vector Machine is a machine learning model with low computational-cost. And SVC (Support Vector Classification) is particularly effective for high-dimensional datasets and cases where decision boundaries are complex.

Given a training dataset (x_i, y_i), where:

- x_i is a feature vector,
- $y_i \in \{-1, 1\}$ represents class labels,

The goal is to find a hyperplane:

$$w \cdot x + b = 0 \tag{2}$$

where w is the weight vector and b is the bias.
The margin is maximized by solving:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \tag{3}$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1, \quad \forall i \tag{4}$$

Param	Value
C	1.0
break_ties	False
cache_size	200
class_weight	None
coef0	0.0
decision_function_shape	ovr
degree	3
gamma	scale
kernel	rbf
max_iter	-1
probability	False
random_state	None
shrinking	True
tol	0.001
verbose	False

Table 1: SVC Hyperparameters

4 Results

This section will show the classification report and confusion matrix of SVC model on test dataset.

Class	Precision	Recall	F1-score	Support
0	0.97	1.00	0.98	18,118
1	0.96	0.56	0.71	556
2	0.97	0.86	0.91	1,448
3	0.76	0.46	0.57	162
4	1.00	0.90	0.95	1,608
Accuracy			0.97	21,892
Macro Avg	0.93	0.76	0.82	21,892
Weighted Avg	0.97	0.97	0.96	21,892

Table 2: Classification Report

The proposed model outperforms the previous benchmark in all key metrics. It achieves a precision of 97%, significantly higher than the prior 95. 9%, demonstrating an improved overall performance. Its 98% precision surpasses 95.2%, indicating better positive prediction reliability. Furthermore, its 98% recall exceeds 95.1%, showing improved sensitivity in detecting all classes. These enhancements suggest better class balance handling and machine learning algorithms that can handle data imbalance.

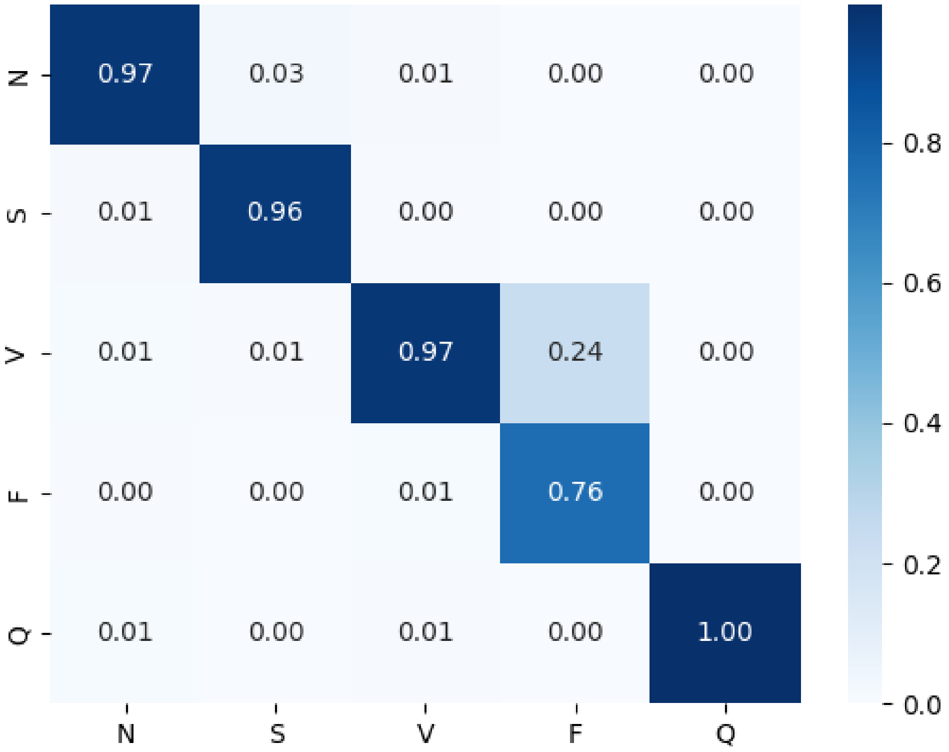

Figure 3: Confusion Matrix

The confusion matrix has strong diagonal elements in class N(97%), S(96%), V(97%), Q(100%) which indicates that the SVC model performs quite well in distinguishing these classes. However, there are some overlapping characteristics (class V has 24% misclassification rate into class F) due to the imbalance in train dataset.