Department of Information and Communication Technology

Machine Learning in Medicine

Report - Practice 1

ID - Student Name:
22BI13472 - Nguyen Ba Vinh - Data Science

Lecturer:
Prof. Tran Giang Son

Academic year: 2022 - 2025
Hanoi, February 2025

1 Introduction

This report covers Practical 1 of Machine Learning in Medicine, focusing on analyzing a publicly available Kaggle dataset to explore its structure and features before creating a classification model. The study outlines the dataset's characteristics, sources, and preprocessing methods. A random forest model is developed for classification, and its performance is carefully evaluated. The results are then compared with those from the original research paper to evaluate the model's effectiveness.

2 About Dataset

The Arrhythmia Dataset, sourced from PhysioNet's MIT-BIH Arrhythmia Database, contains 109,446 samples with a sampling frequency of 125 Hz. The dataset is categorized into five classes representing different heartbeat types: normal beats (N-encoded 0), supraventricular ectopic beats (S-encoded 1), ventricular ectopic beats (V-encoded 2), fusion beats (F-encoded 3), and unclassified beats (Q-encoded 4). This dataset is widely used for training and evaluating machine learning models for arrhythmia detection, serving as a benchmark for heartbeat classification tasks in medical research. The data is split into 80% for training and 20% for testing. The distribution of data points across these classes is shown in the **Figure 1** below.

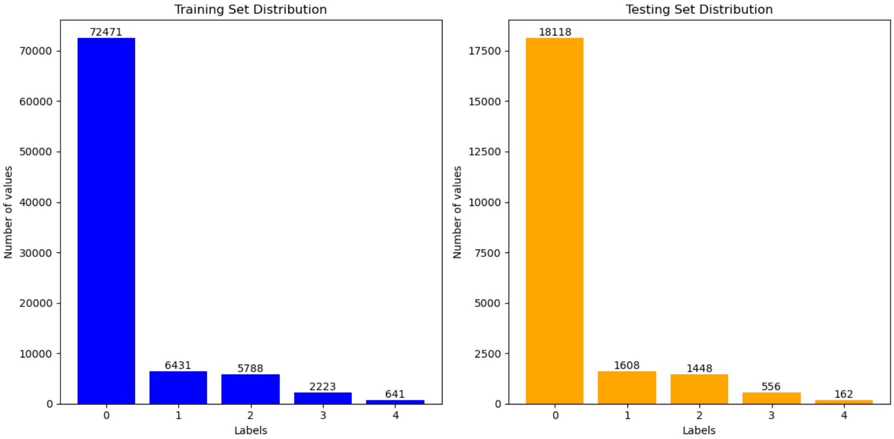


Figure 1: Distribution of data points in the dataset

3 Preprocessing

During the preprocessing stage the examination of the distribution of data points across the five heartbeat classes in the Arrhythmia Dataset revealed that the dataset is imbalanced, with some classes containing significantly fewer samples than others. This class imbalance could bias the model toward majority classes, reducing its ability to accurately classify minority classes. To mitigate the problem, the SMOTETomek technique — a hybrid approach combining Synthetic Minority Over-sampling Technique (SMOTE) and Tomek links is applied in the training set. **Result** after resampling the dataset is shown in the **Figure 2** below.

- SMOTE generates synthetic samples for minority classes by interpolating between existing data points.
- Tomek links remove noisy and borderline samples that may overlap between classes

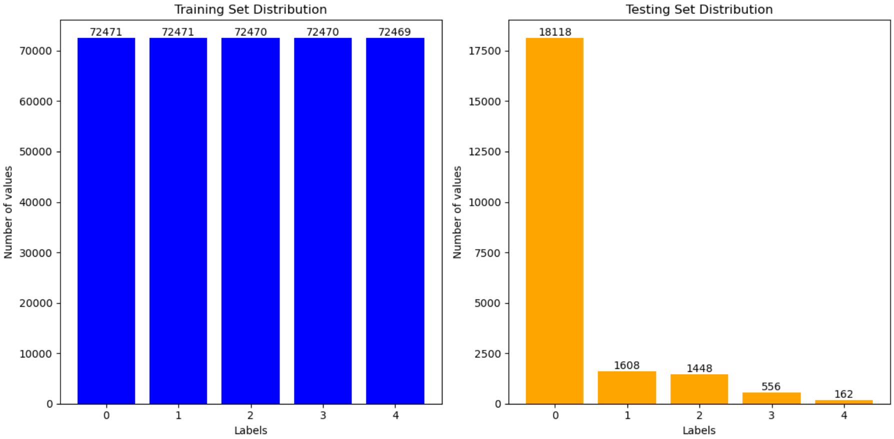


Figure 2: Distribution of data points in the dataset after resampling

4 Model Building

After applying SMOTETomek, the class distribution became more balanced, enhancing the model's ability to learn meaningful patterns across all classes. A **Random Forest Classifier** is then employed for the classification task, chosen for its robustness, ability to handle high-dimensional data, and effectiveness in reducing overfitting through ensemble learning. The model was implemented using the default hyperparameter settings with:

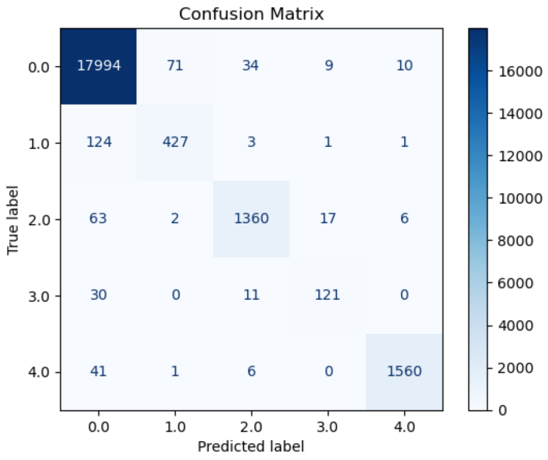
- Number of trees: 100
- Split criterion: Gini impurity
- Bootstrap sampling
- Stopping criteria: Splits until pure or too few samples

5 Results

After applying the machine learning model, the performance of the model is shown in **Table 1** and **Figure 3** below.

Class	Precision	Recall	F1-score	Support
0	0.99	0.99	0.99	18118
1	0.85	0.77	0.81	556
2	0.96	0.94	0.95	1448
3	0.82	0.75	0.78	162
4	0.99	0.97	0.98	1608
Accuracy			0.98	21,892
Macro Avg	0.92	0.88	0.90	21,892
Weighted Avg	0.98	0.98	0.98	21,892

Figure 3: Classification Report



	0.0	1.0	2.0	3.0	4.0
0.0	17994	71	34	9	10
1.0	124	427	3	1	1
2.0	63	2	1360	17	6
3.0	30	0	11	121	0
4.0	41	1	6	0	1560

Figure 4: Model Performance Visualization

The results indicate that the Random Forest model performs remarkably well, achieving an overall accuracy of 98%. The model gains near-perfect classification for the majority classes (0, 2, and 4), but still faces challenges with minority classes (1 and 3), as reflected in their lower recall scores. This suggests that despite applying SMOTETomek to address class imbalance, the model retains some bias toward the more frequent classes. Notably, when compared to the deep convolutional neural network approach from the original paper, which achieved an accuracy of 93.4%, the Random Forest Classifier outperforms it, highlighting the effectiveness of the chosen machine learning technique for this dataset.