Labwork 1 Report - ECG Heartbeat Categorization Using ML

Ngo Thanh Dat
Student ID: 23BI14090
University of Science and Technology of Hanoi
Email: DatNT.23BI14090@usth.edu.vn

I. Introduction

Electrocardiogram (ECG) signals are widely used for monitoring cardiac activity and diagnosing heart diseases. Automatic heartbeat classification plays an important role in assisting clinicians by reducing manual analysis effort. In this practical work, a machine learning approach is applied to classify ECG heartbeats using a publicly available dataset. The objective is to explore the dataset, implement a classification model, and evaluate its performance.

II. Dataset Description

The dataset used in this study is the ECG Heartbeat Categorization Dataset derived from the MIT-BIH Arrhythmia Database and obtained from Kaggle. It consists of preprocessed electrocardiogram (ECG) heartbeat segments extracted from long-term ambulatory recordings.

Each sample in the dataset represents a single heartbeat and is encoded as a one-dimensional time series of 187 numerical values corresponding to ECG signal amplitudes over time. These values capture key cardiac waveform components such as the P wave, QRS complex, and T wave. The final attribute of each sample is a categorical label indicating the heartbeat class.

The dataset contains five heartbeat categories:

- Class 0: Normal heartbeat
- Class 1: Supraventricular ectopic beat
- Class 2: Ventricular ectopic beat
- Class 3: Fusion beat
- Class 4: Unknown beat

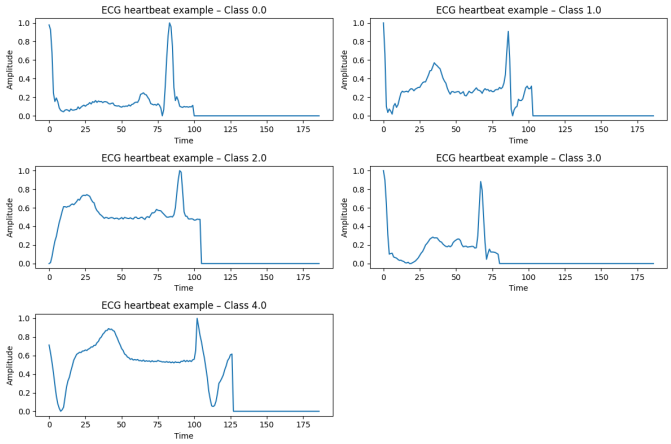

Fig. 1. ECG heartbeat examples for the five heartbeat classes.

Figure 1 presents representative ECG heartbeat examples for each class. Normal heartbeats exhibit a regular waveform with a prominent QRS complex, while abnormal heartbeats show noticeable morphological variations. These differences highlight the importance of temporal signal patterns for heartbeat classification.

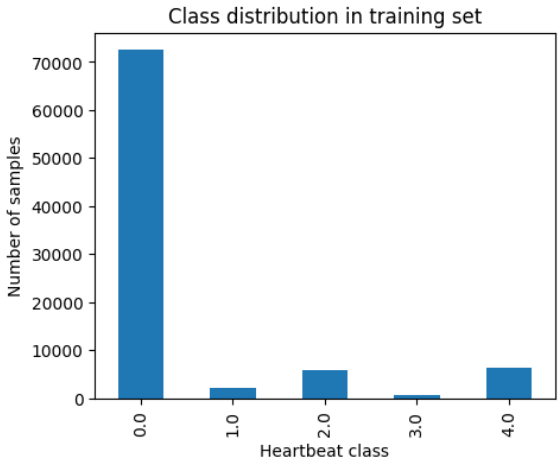

Fig. 2. Distribution of ECG heartbeat classes in the training dataset.

Figure 2 shows the distribution of heartbeat classes in the training dataset. The dataset is highly imbalanced, with normal

heartbeats accounting for the majority of samples, while abnormal heartbeat classes are significantly underrepresented. This imbalance poses a major challenge for classification models.

The dataset does not contain missing values, and the ECG signals have been preprocessed and normalized prior to distribution. Despite its clean structure, the combination of high dimensionality and class imbalance makes ECG heartbeat classification a non-trivial task.

III. Model Implementation

To address the ECG heartbeat classification task, a Random Forest classifier was employed. Random Forest is an ensemble learning method that combines multiple decision trees to improve classification robustness and reduce overfitting. This model is well-suited for tabular data and can effectively handle non-linear relationships between input features.

Each ECG heartbeat sample, represented by 187 numerical features, is used directly as input to the model without manual feature extraction. The dataset is split into training and testing sets to evaluate generalization performance. During training, multiple decision trees are constructed using bootstrap sampling, and final predictions are obtained through majority voting.

Random Forest was chosen for its simplicity, interpretability, and strong baseline performance, making it appropriate for an exploratory study on ECG heartbeat classification.

IV. Hyperparameter Experiments

Several hyperparameters of the Random Forest model were experimentally adjusted to study their influence on classification performance. Key hyperparameters include the number of trees (*n_estimators*), maximum tree depth, and minimum number of samples required to split a node.

Increasing the number of trees generally improved model stability but also increased computational cost. Limiting the maximum depth helped reduce overfitting, particularly given the strong class imbalance in the dataset. These experiments demonstrate that appropriate hyperparameter tuning is essential to balance model complexity and generalization performance.

V. Results and Discussion

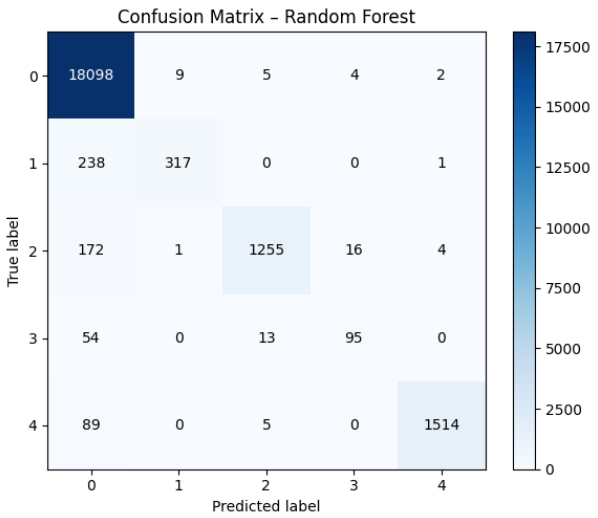

Fig. 3. Confusion matrix of the Random Forest classifier on the test dataset.

Figure 3 presents the confusion matrix obtained on the test dataset. The model shows strong performance for Class 0 (normal heartbeat), which is expected due to its dominance in the training data.

However, misclassifications are more frequent for minority classes such as Class 1 and Class 3. A noticeable number of abnormal heartbeats are incorrectly predicted as normal, highlighting the impact of class imbalance. Despite this limitation, the model is still able to correctly identify a significant portion of ventricular ectopic beats (Class 2) and unknown beats (Class 4).

These results indicate that while the model effectively captures general ECG signal patterns, its sensitivity to underrepresented classes remains limited. This behavior is commonly observed in imbalanced classification problems.

VI. Conclusion

In this work, an ECG heartbeat classification task was studied using a publicly available dataset. The dataset was analyzed in detail, including visualization of ECG signals and examination of class imbalance. A Random Forest classifier was implemented, and the impact of hyperparameter choices was investigated.

Experimental results show that the model performs well on normal heartbeats but faces challenges when classifying minority heartbeat categories. This limitation suggests that future work could explore class balancing techniques, such as resampling or cost-sensitive learning, as well as more advanced deep learning models to improve classification performance.