ECG Heartbeat Classification using an LSTM Network on the MIT-BIH Dataset

Ho Huyen Chau - 23BI14067
Machine Learning in Medicine 2026
University of Science and Technology of Hanoi (USTH)
Email: chauhh.23bi14067@usth.edu.vn

I. Introduction

Electrocardiogram (ECG) signals are widely used in clinical practice to analyze the electrical activity of the heart and to detect cardiac abnormalities such as arrhythmias. Automatic ECG heartbeat classification has therefore become an important research topic in medical signal processing, as it can support clinicians by reducing manual workload and improving diagnostic consistency.

Despite significant progress, the classification of heartbeats on the ECG remains a challenging task. Real-world ECG datasets typically exhibit a strong class imbalance, where normal heartbeats dominate the data, while abnormal events occur much less frequently. These challenges make it difficult to design models that perform well in all heartbeat classes.

In this work, I study the problem of ECG heartbeat classification using the MIT-BIH Arrhythmia dataset. Each heartbeat is represented as a fixed-length one-dimensional signal, and the task is formulated as a multi-class classification problem with five heartbeat categories. Instead of relying on deep learning models, which often require substantial computational resources, I focus on a classical machine learning approach based on gradient boosting. This choice allows me to analyze the impact of data imbalance and training strategies while maintaining a relatively simple and interpretable classification pipeline.

II. Dataset Description

A. MIT-BIH Arrhythmia Dataset

The primary dataset used in this study is the MIT-BIH Arrhythmia dataset, originally collected by the Massachusetts Institute of Technology and Beth Israel Hospital and made publicly available via PhysioNet. In this work, I use a preprocessed version of the dataset provided on Kaggle, where individual heartbeats have been segmented and normalized.

Each heartbeat is represented by 187 sampled amplitude values corresponding to approximately 1.5 seconds of ECG recording at a sampling frequency of 125 Hz. Consequently, each sample consists of 188 columns, where the first 187 columns contain the ECG signal values and the last column corresponds to the class label. The dataset is divided into a training set containing 87,554 samples and a test set containing 21,892 samples.

The classification task involves five heartbeat categories: Normal (N), Supraventricular ectopic (S), Ventricular ectopic (V), Fusion (F), and Unknown (Q).

B. Class Distribution and Imbalance

A major characteristic of the MIT-BIH dataset is its severe class imbalance. Normal heartbeats account for more than 80% of the samples in both the training and test sets, while some abnormal classes, such as Fusion beats, represent less than 1% of the data. This imbalance poses a significant challenge for classification models, as high accuracy can be achieved by predicting only the majority class.

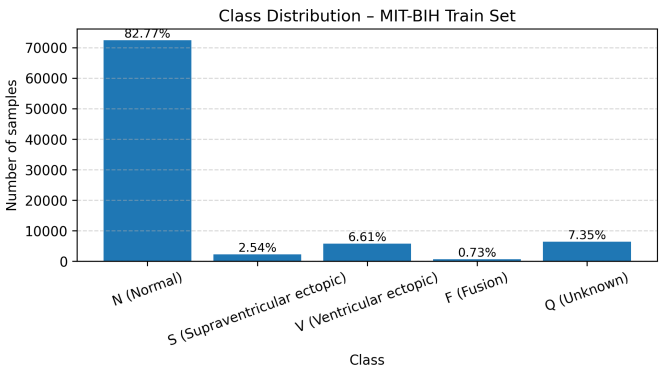

Fig. 1: Class distribution of the MIT-BIH training set.

Figures 1 and 2 present the class distributions of the training and test sets, respectively, and clearly illustrate the dominance of the Normal class over the minority classes.

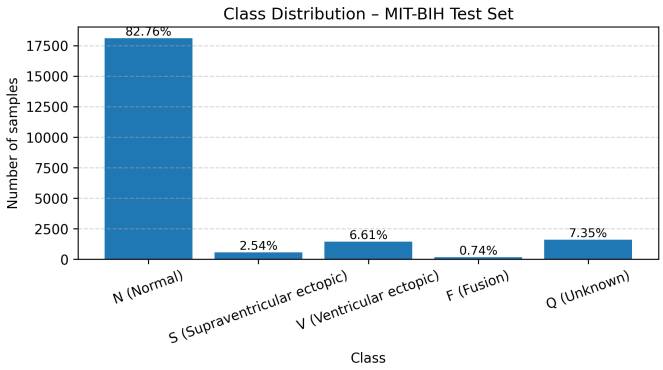

Fig. 2: Class distribution of the MIT-BIH test set.

Figure 3 also indicates a noticeable class imbalance in the PTB dataset although it is less imblanced than MIT-BIH one.

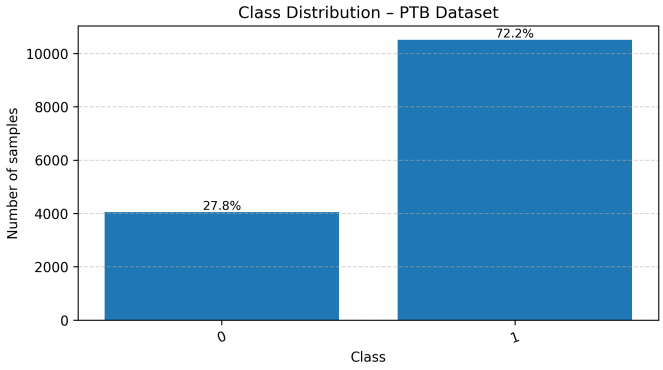

Fig. 3: Class distribution of the PTB dataset.

C. Visual Exploration of ECG Signals

To better understand the morphological differences between heartbeat classes, visual inspection of ECG waveforms is performed. Several example heartbeats from each class are shown in Figure 4. These examples highlight noticeable differences in waveform shape and peak structure among the heartbeat categories.

In addition, the mean waveform and the corresponding ± 1 standard deviation band are computed for each class, as illustrated in Figure 5. These visualizations provide insights into both the typical heartbeat morphology and the variability within each class.

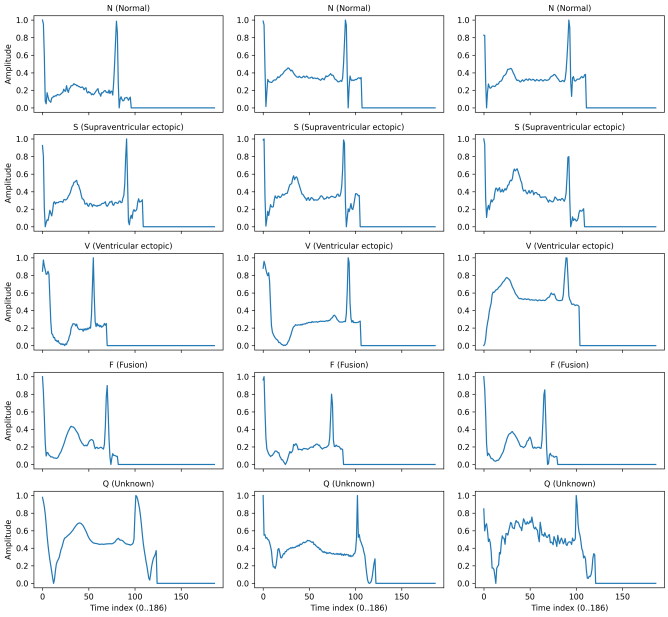

Fig. 4: Example ECG heartbeats for each class in the MIT-BIH dataset.

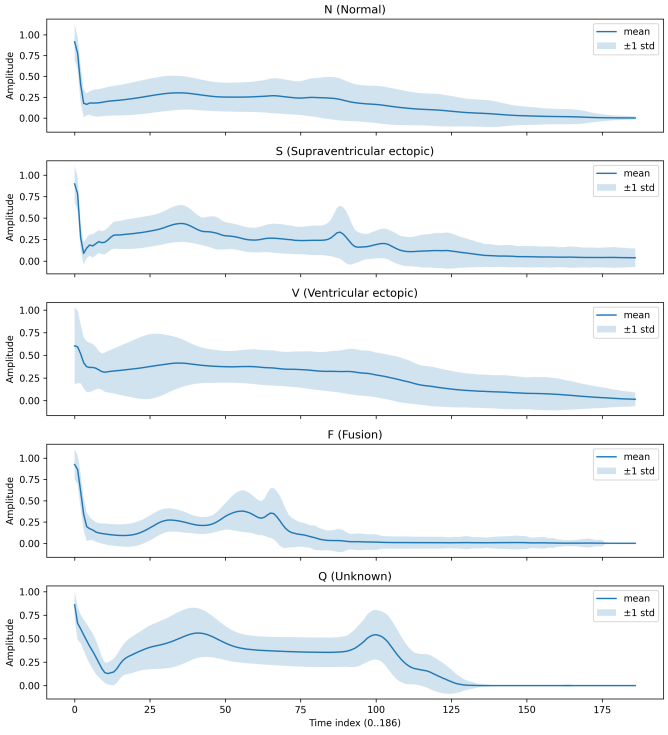

Fig. 5: Mean ECG waveform and ± 1 standard deviation for each heartbeat class.

III. Methodology

A. Overall Pipeline

The overall ECG classification pipeline consists of four main steps: data preprocessing, train–validation splitting, model training, and performance evaluation. Each ECG heartbeat is treated as a fixed-length feature vector of 187 signal values. No handcrafted features are extracted; instead, the raw normalized waveform samples are directly used as input to the classifier.

The training set is further divided into training and validation subsets using a stratified split to preserve the original class distribution. The validation set is used to monitor model performance and guide hyperparameter selection, while the test set is kept strictly unseen until final evaluation.

B. Gradient Boosting Classifier

For heartbeat classification, we employ a gradient boosting model based on decision trees, specifically the `HistGradientBoostingClassifier` from the scikit-learn library. Gradient boosting builds an ensemble of weak learners in a sequential manner, where each new tree attempts to correct the errors made by the previous ensemble. This method is well suited for tabular data and can capture nonlinear relationships between input features.

The model is trained using the multinomial log-loss objective, appropriate for multi-class classification. The main hyperparameters include the learning rate, maximum number of boosting iterations, tree depth, and regularization terms controlling model complexity.

C. Hyperparameter Configuration

After empirical experimentation, the final model configuration uses a learning rate of 0.1 and a maximum of 300 boosting iterations. Tree complexity is controlled through a maximum depth of 6, a maximum of 31 leaf nodes, and a minimum of 20 samples per leaf. An L2 regularization term is applied to reduce overfitting.

Early stopping is enabled based on the validation loss, with a validation fraction of 10% and a patience of 20 iterations. This mechanism stops training when no significant improvement is observed, preventing unnecessary model complexity growth.

D. Handling Class Imbalance

Due to the strong class imbalance in the dataset, multiple sample weighting strategies are explored during training. These strategies aim to increase the contribution of minority classes during optimization. Model selection is primarily guided by the macro-averaged F1 score on the validation set, which equally weights all classes regardless of their frequency and provides a more informative measure than accuracy alone in imbalanced settings.

E. Evaluation Metrics

Model performance is evaluated using accuracy, macro-averaged F1 score, and class-wise precision and recall. In addition, confusion matrices are analyzed to better understand misclassification patterns and to identify whether the model disproportionately favors the majority class.

IV. Results

A. Validation Performance

On the validation set, the gradient boosting classifier achieves a high accuracy of 95.23% and a macro-averaged F1 score of 81.25%. This result initially suggests that the model is able to fit the training data and capture relevant patterns in the ECG signals. However, given the severe class imbalance of the dataset, these metrics must be interpreted with caution.

B. Test Set Performance

When evaluated on the independent test set, the model achieves an overall accuracy of 81.6%. However, the macro-averaged F1 score drops significantly to 18.17%, indicating poor performance on minority classes. This discrepancy between accuracy and macro-F1 highlights the limitations of accuracy as a performance metric in imbalanced classification problems.

The detailed classification report reveals that the model predicts the majority Normal class with very high recall, while the minority classes are almost entirely misclassified. Precision and recall for classes S, V, F, and Q are close to zero, indicating that the model effectively collapses to predicting the majority class.

C. Confusion Matrix Analysis

The confusion matrix further confirms this behavior. Most test samples from minority classes are incorrectly predicted as Normal heartbeats. Very few samples are assigned to non-Normal classes, resulting in extremely low recall for abnormal heartbeat categories. This outcome demonstrates that the model has learned a biased decision boundary favoring the dominant class.

D. Discussion of Overfitting and Class Collapse

Although early stopping and regularization are applied, the model still exhibits a strong tendency to overfit the majority class. The high validation macro-F1 score suggests that the validation split may not fully represent the difficulty of the test distribution, especially for rare classes. As a result, the model fails to generalize its decision boundaries to minority heartbeat types.

These results emphasize the importance of robust imbalance handling strategies and careful validation design when applying classical machine learning models to highly imbalanced medical datasets. While gradient boosting provides strong performance on the dominant class, additional techniques such as more advanced resampling strategies or alternative model architectures may be required to achieve balanced performance across all heartbeat categories.