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

Nguyen The Cuong
22BA13058
Machine Learning in Medicine 2026
University of Science and Technology of Hanoi (USTH)
Email: cuongnt.22ba13058@usth.edu.vn

ABSTRACT

This report presents a deep learning baseline for classifying electrocardiogram (ECG) heartbeat segments into five categories using the MIT-BIH heartbeat dataset. We analyze class imbalance and visualize representative waveforms. The pipeline performs a stratified train/validation split, standardizes features, reshapes heartbeats into sequence format, and trains a lightweight Long Short-Term Memory (LSTM) network with early stopping. Performance is evaluated on a held-out test set using accuracy, macro F1-score, a per-class classification report, and a normalized confusion matrix. Results show strong overall accuracy, while minority classes remain more challenging due to imbalance. Future work includes class weighting, data augmentation, and hybrid architectures such as CNN-LSTM or attention mechanisms.

Index Terms—ECG, MIT-BIH, Arrhythmia, LSTM, Deep Learning, Classification

I. Introduction

Electrocardiograms (ECG) provide non-invasive measurements of cardiac electrical activity and are widely used to detect arrhythmias. Automatic heartbeat classification can support clinical decision-making by reducing manual annotation effort and enabling scalable screening.

In this practical work, we build a simple baseline model using a Long Short-Term Memory (LSTM) network to classify heartbeat segments into five classes. The objective is to establish a clear end-to-end pipeline including data exploration, preprocessing, model training, and evaluation on a held-out test set.

II. Dataset

A. Class Distribution

Fig. 1 shows the number of samples per class in the training and test sets. The dataset is strongly imbalanced: the *Normal* class dominates, while *Supraventricular* and *Fusion* are rare. This imbalance can bias the model toward predicting the majority class, so macro-averaged metrics are important during evaluation.

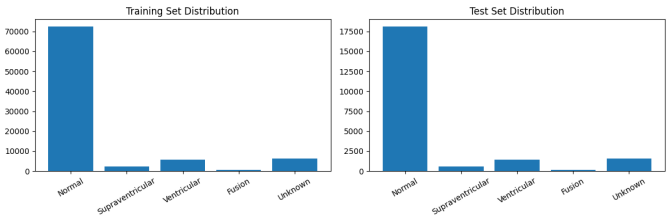

Fig. 1: Class distribution in the training set (left) and test set (right).

B. Representative ECG Heartbeats

Fig. 2 presents one example heartbeat waveform from each class. Each subplot shows the amplitude over timesteps for a single sample. We observe that waveform morphology differs across classes, but some minority classes have patterns that may resemble the majority class, making classification more difficult.

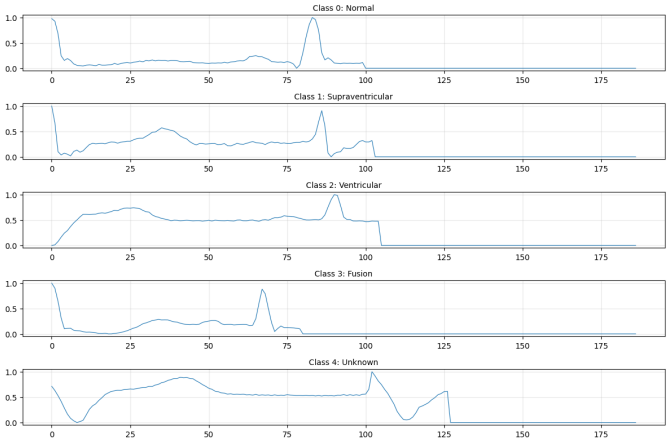

Fig. 2: One representative heartbeat waveform for each class (five subplots).

III. Methods

A. Preprocessing

We split the original training data into training and validation sets using an 85/15 stratified split. Features are standardized using `StandardScaler`, fitted on the training split only and then applied to validation and test sets. Each heartbeat segment is reshaped into sequence format $(N, T, 1)$ to match LSTM input requirements, where T denotes the number of timesteps.

B. Model Architecture

We use a lightweight LSTM classifier:

- One LSTM layer for temporal modeling
- Dropout for regularization
- A dense hidden layer with ReLU activation
- A softmax output layer for 5-class classification

TABLE I: LSTM architecture (example).

Layer	Output	Description
Input	$(T, 1)$	Heartbeat sequence
LSTM (64)	(64)	Temporal modeling
Dropout (0.3)	(64)	Regularization
Dense (32, ReLU)	(32)	Feature projection
Dense (5, Softmax)	(5)	Class probabilities

C. Training Setup

The model is trained using sparse categorical cross-entropy loss and the Adam optimizer (lr $= 10^{-3}$). Early stopping monitors validation loss and restores the best weights to reduce overfitting.

IV. Results

A. Learning Curves

Fig. 3 shows training and validation accuracy across epochs, and Fig. 4 shows the corresponding loss curves. The model improves during early epochs, indicating effective learning from the training data. After the best epoch, validation performance becomes less stable, suggesting that early stopping is important to prevent overfitting.

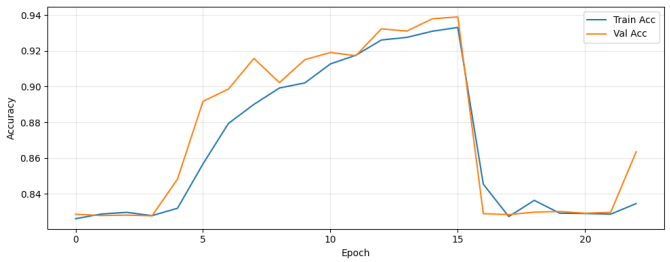

Fig. 3: Training and validation accuracy across epochs.

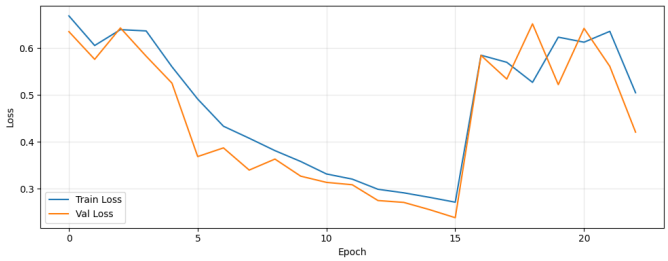

Fig. 4: Training and validation loss across epochs.

B. Confusion Matrix

Fig. 5 shows the normalized confusion matrix on the test set. The model predicts the *Normal* class with high accuracy. However, minority classes (especially *Supraventricular* and *Fusion*) are frequently confused with *Normal*, reflecting the effect of class imbalance and waveform similarity.

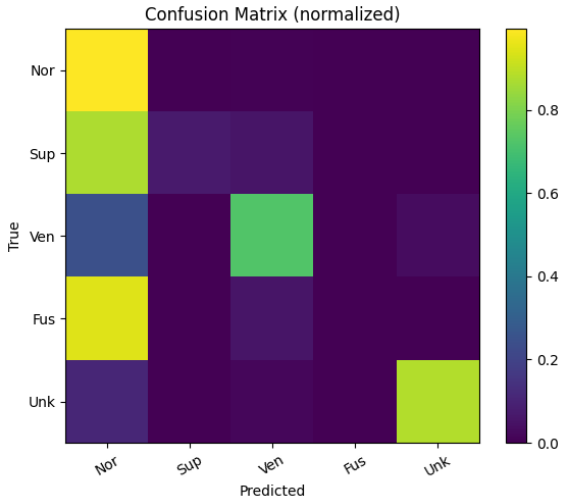

Fig. 5: Normalized confusion matrix on the test set (rows sum to 1).

C. Per-class Classification Report

Table II reports precision, recall, F1-score, and support for each class on the test set. The model achieves high performance on the majority class (*Normal*) and strong results on *Unknown*. However, performance drops substantially for minority classes, especially *Supraventricular* (low recall) and *Fusion* (near-zero detection), which is consistent with the strong class imbalance observed in Fig. 1.

TABLE II: Classification report on the test set.

Class	Precision	Recall	F1-score	Support
Normal	0.94	0.99	0.97	18118
Supraventricular	0.91	0.07	0.13	556
Ventricular	0.87	0.72	0.79	1448
Fusion	0.00	0.00	0.00	162
Unknown	0.95	0.88	0.91	1608
Accuracy		0.94		21892
Macro avg	0.73	0.53	0.56	21892
Weighted avg	0.93	0.94	0.92	21892

D. Prediction Confidence

Fig. 6 shows the distribution of predicted probability assigned to the true class, separated by class. The *Normal* class has very high confidence (most probabilities close to 1.0), while minority classes have lower and more spread-out confidence, which indicates higher uncertainty for these classes.

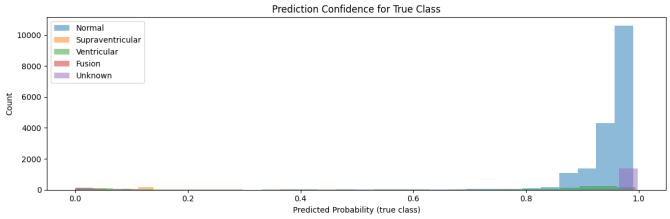

Fig. 6: Prediction confidence distribution for the true class, grouped by class.

V. Discussion

The LSTM baseline achieves strong overall accuracy, indicating that temporal patterns in heartbeat segments are informative for classification. However, due to class imbalance, minority classes typically show lower recall and are more likely to be confused with other categories. This behavior is commonly visible in the confusion matrix and macro-averaged metrics.

Possible improvements include:

- **Class weighting** or **focal loss** to reduce majority-class dominance.
- **Data augmentation** (noise injection, time shifting, scaling) for minority classes.
- **Hybrid architectures** such as CNN-LSTM to improve feature extraction.
- **Attention mechanisms** to focus on salient waveform regions.

VI. Conclusion

This work presented an end-to-end pipeline for ECG heartbeat classification on the MIT-BIH dataset using a simple LSTM neural network. We performed data exploration, preprocessing, model training, and evaluation on a held-out test set. While overall performance is strong, improving minority-class classification remains an important next step. Future work will explore imbalance-aware training and richer architectures to enhance robustness.