EGC Heartbeat Classification

NGUYEN Van Phu

February 25, 2025

1 Introduction

This report conducts experiments on the ECG heartbeat dataset provided on Kaggle using Machine Learning and Deep Learning methods. The results are based on the MIT-BIH Arrhythmia ECG Database, which serves as the data source for labeled ECG records. The report presents experiments and performance evaluations using three methods: SVM, FNN, and LSTM.

2 Exploratory Data Analysis

ECG is widely used by cardiologists and medical practitioners to monitor cardiac health. The main challenge in manually analyzing ECG signals, like many other types of time-series data, is detecting and categorizing the various waveforms and morphologies. The dataset comprises five types of heartbeats—Normal, Supraventricular, Ventricular, Fusion, and Unknown—encoded as 0, 1, 2, 3, and 4, respectively. Each sample contains 187 values representing the heartbeat signal. However, analysis shows that the MIT-BIH dataset is imbalanced, which may bias the model's results.

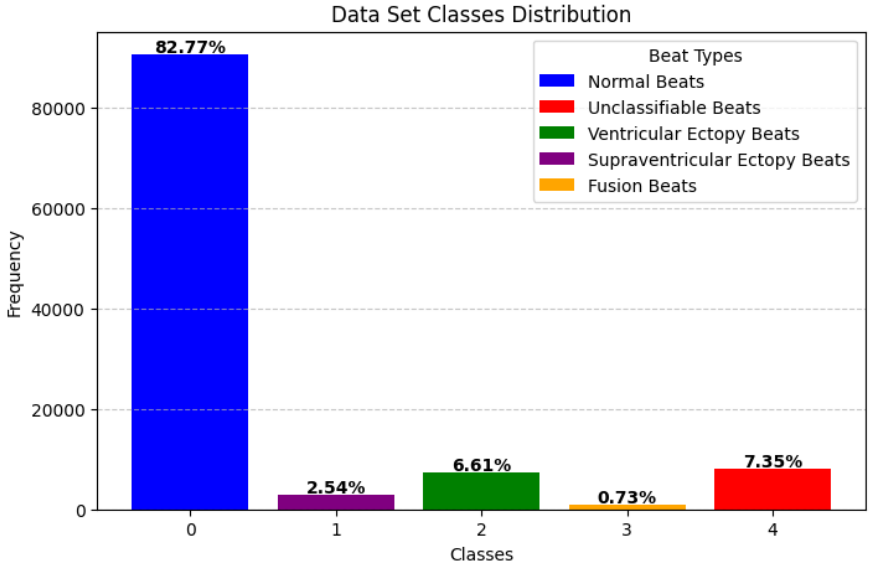


Figure 1: Data set class distributrion

2.1 Data Pre-processing

By using an up-scaling method, I added more samples to balance each class. Specifically, the training set now contains 20,000 samples per class, and the test set contains 5,000 samples per class, resulting in a total of exactly 125,000 samples equally distributed among the classes. So that the rate between traning and testing is 80% and 20% respectively

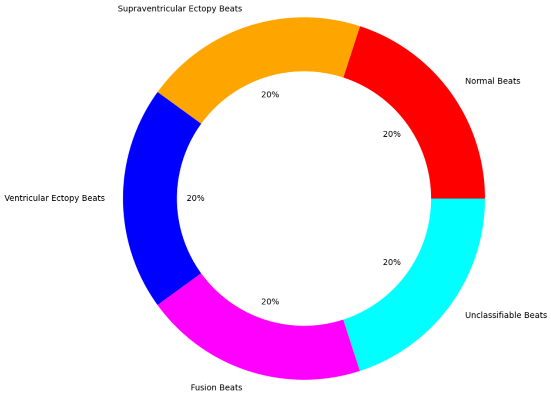


Figure 2: Data post upscale ratio

3 Experiments

For the classification task, we have many ways to do it, but I choose the 3 most common in Machine Learning and Deep Learning: SVM, FNN, and LSTM.

3.1 SVM

Support Vector Machine (SVM) is a supervised machine learning algorithm primarily used for classification tasks, although it can also be adapted for regression. The main goal of SVM is to find the optimal hyperplane that best separates different classes in a high-dimensional space. It achieves this by identifying support vectors—the data points that are closest to the decision boundary—and maximizing the margin between these points and the hyperplane. This maximization of the margin helps improve the model's generalization ability.
These results indicate decent performance, but there are some key observations from the confusion matrix:

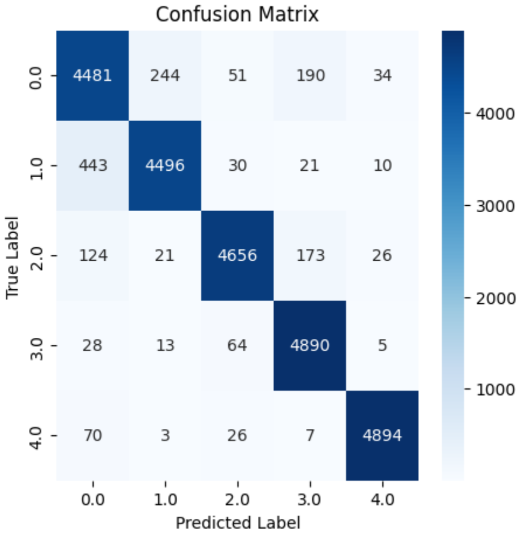


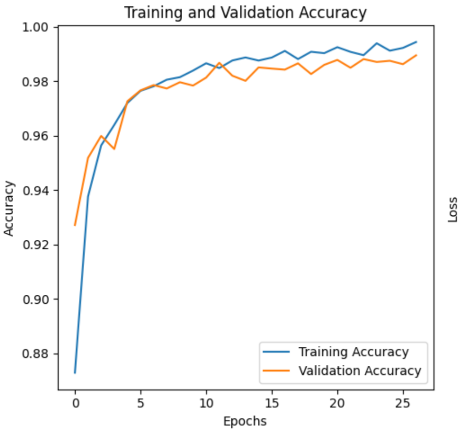
Figure 3: SVM confusion matrix

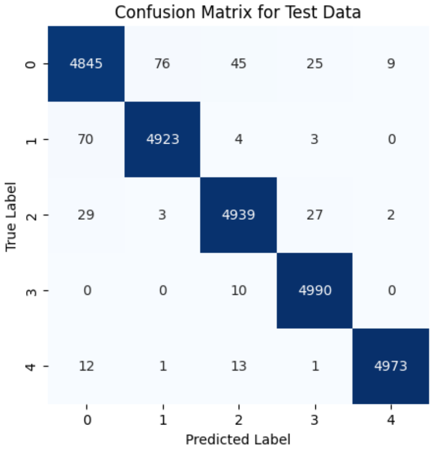
Misclassification of Classes:
Some misclassifications occur, especially in class 0 (Normal Beats) and class 1 (Supraventricular Ectopy Beats).
For example, 244 samples from class 0 were predicted as class 1. Similarly, 443 samples from class 1

were misclassified as class 0.
Class 2 (Ventricular Ectopy Beats) is relatively well classified, but 173 samples were misclassified as class 3 (Fusion Beats).
Class 4 (Unclassifiable Beats) performed best, with minimal misclassification.

3.2 FNN

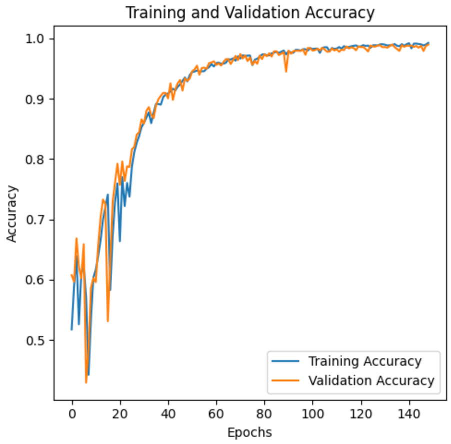
FNN (Feedforward Neural Network) is a type of artificial neural network used for both classification and regression tasks. It consists of an input layer, one or more hidden layers, and an output layer, with information flowing in one direction only—from input to output. Each neuron applies a non-linear activation function, allowing the network to learn complex patterns in the data. This straightforward structure makes FNNs widely used in applications such as image recognition, natural language processing, and time series forecasting.

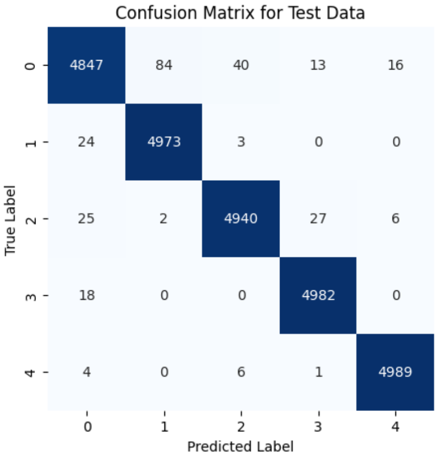




3.3 LSTM

LSTM (Long Short-Term Memory) is a specialized type of recurrent neural network designed to capture long-term dependencies in sequential data. It uses memory cells and gating mechanisms (input, forget, and output gates) to selectively retain or discard information, effectively addressing issues like the vanishing gradient problem. This makes LSTMs well-suited for tasks such as language modeling, speech recognition, and time series forecasting.


Training and Validation Accuracy
Accuracy
Epochs
Training Accuracy
Validation Accuracy


Confusion Matrix for Test Data
True Label
Predicted Label