Study of ECG Heartbeat Categorization

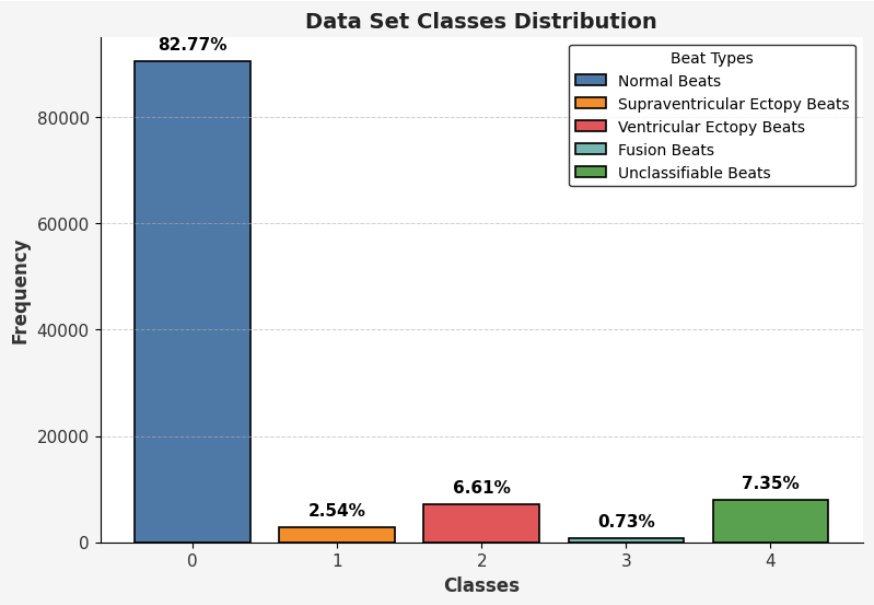
Name: Nguyen Hoang Ha
ID Number: BA12-066

1 Introduction

This report summarizes my study on ECG heartbeat classification using ECG Heartbeat Categorization Dataset from Kaggle, specifically the MIT-BIH Arrhythmia Database. In this study, I analyzed the dataset and implemented Machine Learning, Deep Learning models to classify five different heartbeat types.

2 Data Analysis

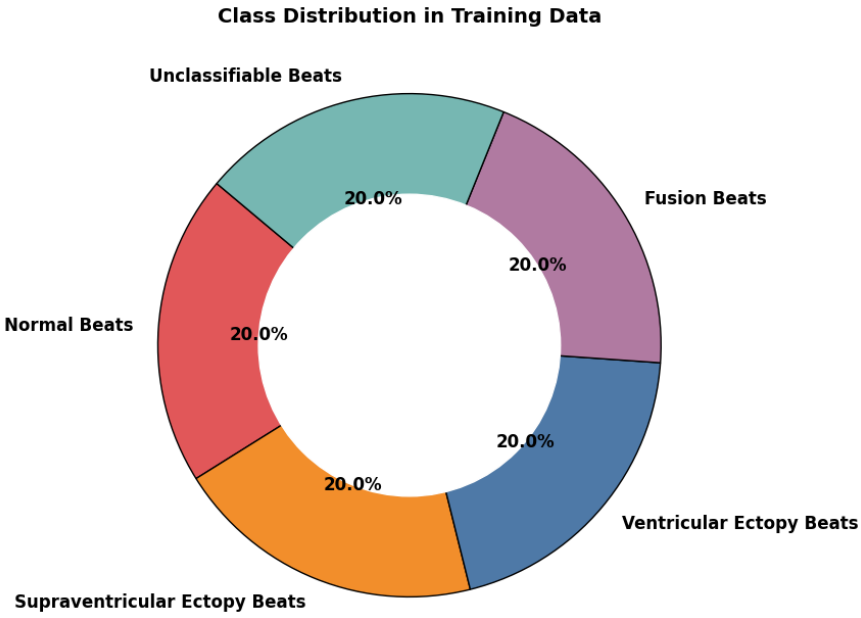
This study utilizes the MIT-BIH Arrhythmia Database, which includes five categories of heartbeats: Normal, Supraventricular, Ventricular, Fusion, and Unknown. These categories are represented by the labels 0, 1, 2, 3, and 4, respectively. Each record consists of 187 values corresponding to the heartbeat signal.



Figuur 1: Frequency of each class in the dataset

As shown in the histogram in Figure 1, the Kaggle training dataset is heavily skewed toward the Normal class, which is the most prevalent heartbeat type. This imbalance can cause the model to overfit to the Normal class.

To mitigate overfitting, I applied an upsampling technique to balance the dataset. By dividing the total number of records in the training set by five, I obtained approximately 17,510. Therefore, each class was resampled to contain 17,500 samples to ensure a more balanced distribution.



Figuur 2: Frequency of each class in the dataset after balancing

3 Model

For the classification task, one machine learning and two deep learning models took 187 values of heartbeat signals as input and predicted the corresponding heartbeat category.

3.1 Random Forest Classifier

The first model I employed was a Random Forest classifier, which is computationally efficient for structured data classification. This model was trained without utilizing a validation dataset. It consists of 100 estimators, and the random state is set to 42 for reproducibility.

3.2 Convolutional Neural Network

I implemented a convolutional neural network (CNN) that utilizes 1D convolutional layers to extract features from ECG signals, leveraging deep learning for effective signal processing. The network consists of two 1D convolutional layers with max pooling for feature extraction, followed by a fully connected layer and an output layer with softmax activation for classification. Batch normalization and dropout are incorporated to enhance stability and prevent overfitting.

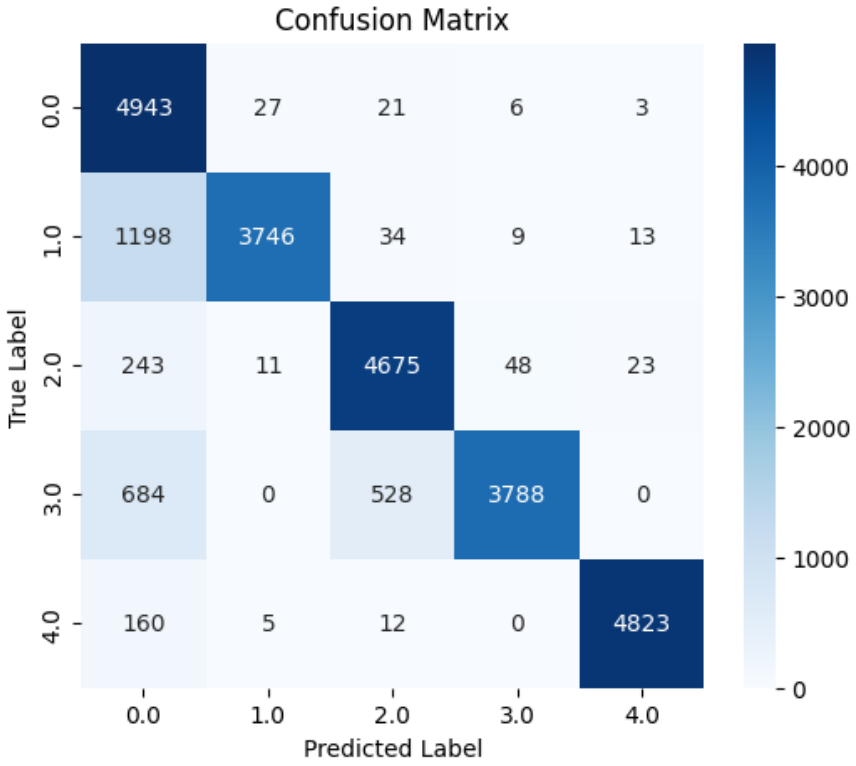
3.3 Transformer

I implemented a Transformer-based model for ECG classification, leveraging self-attention to extract meaningful features from heartbeat signals. The model consists of an embedding layer, followed by a multi-layer Transformer encoder to capture long-range dependencies. A global average pooling layer condenses the information, and a fully connected layer with softmax activation classifies the heartbeat into five categories. Dropout is used to enhance generalization.

4 Evaluation

4.1 Random Forest Classifier

The confusion matrix for the Random Forest classifier shows that the model performs well in classifying most heartbeat categories, with high accuracy in Normal Beats (Class 0) and Unclassifiable Beats (Class 4). However, there is noticeable misclassification in Supraventricular (Class 1) and Fusion Beats (Class 3), where a significant number of samples are incorrectly classified as Normal Beats (Class 0) or Ventricular Beats (Class 2). This suggests that while Random Forest is effective for structured data, it may struggle with distinguishing certain ECG patterns, indicating the need for feature engineering or alternative models like deep learning.

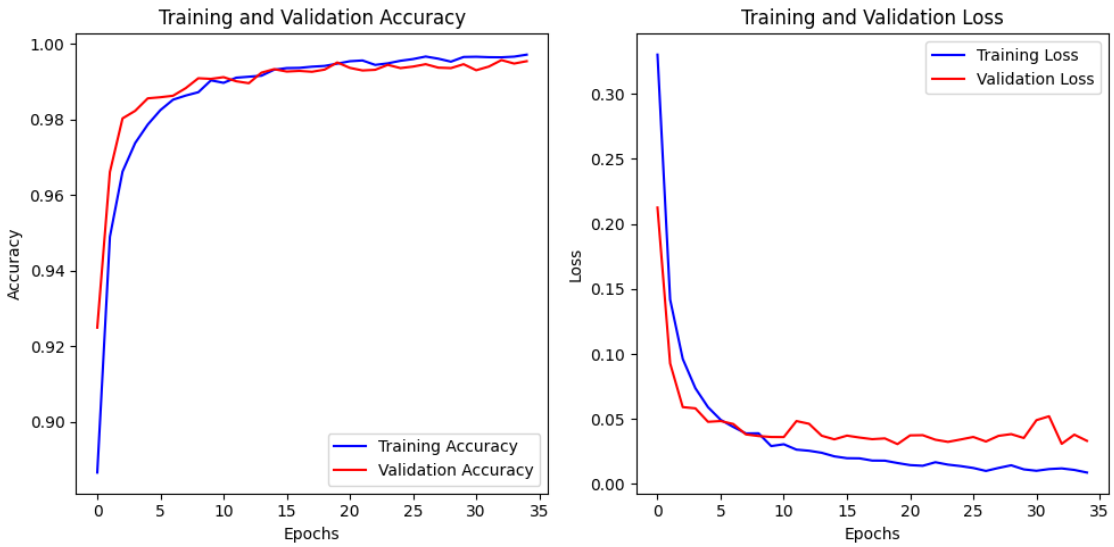


Figuur 3: Confusion Matrix of Random Forest Classifier

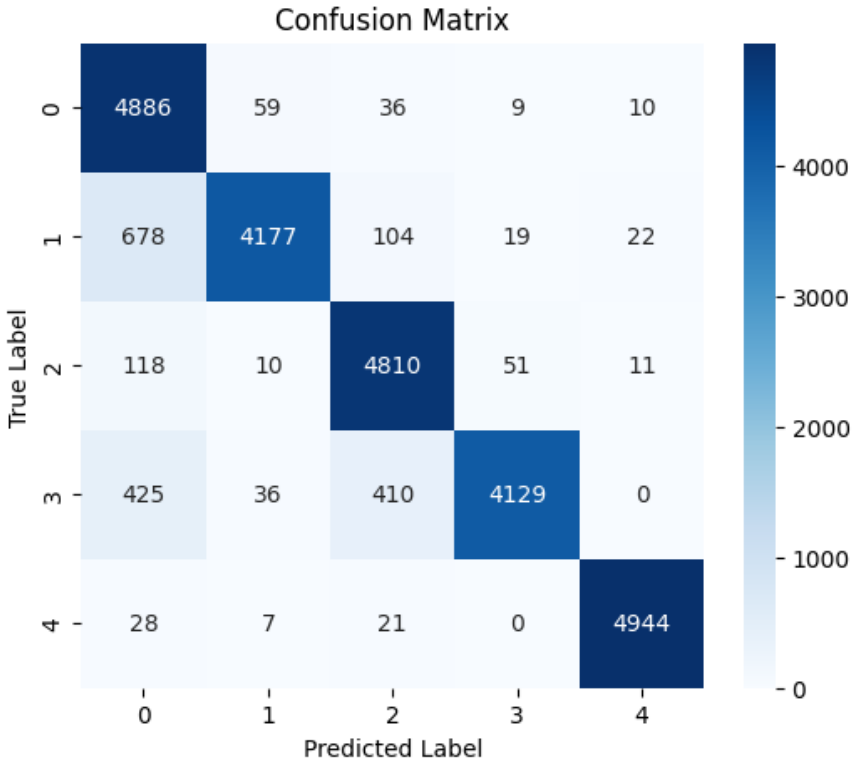
4.2 Convolutional Neural Network

The training and validation accuracy curves show that the Convolutional Neural Network achieves near-perfect accuracy after a few epochs, indicating that the model learns quickly and generalizes well. The training and validation loss curves decrease rapidly in the first few epochs and then stabilize at very low values, suggesting effective learning. However, slight fluctuations in validation loss after epoch 10 may indicate minor overfitting. Overall, the CNN performs well, but further evaluation on unseen data is necessary to confirm its robustness.

The confusion matrix for the CNN model shows strong performance, especially in Normal (Class 0), Ventricular (Class 2), and Unclassifiable Beats (Class 4). However, some misclassification occurs in Supraventricular (Class 1) and Fusion Beats (Class 3), often mistaken for Normal or Ventricular Beats. Further optimization, such as tuning or data augmentation, could improve accuracy.



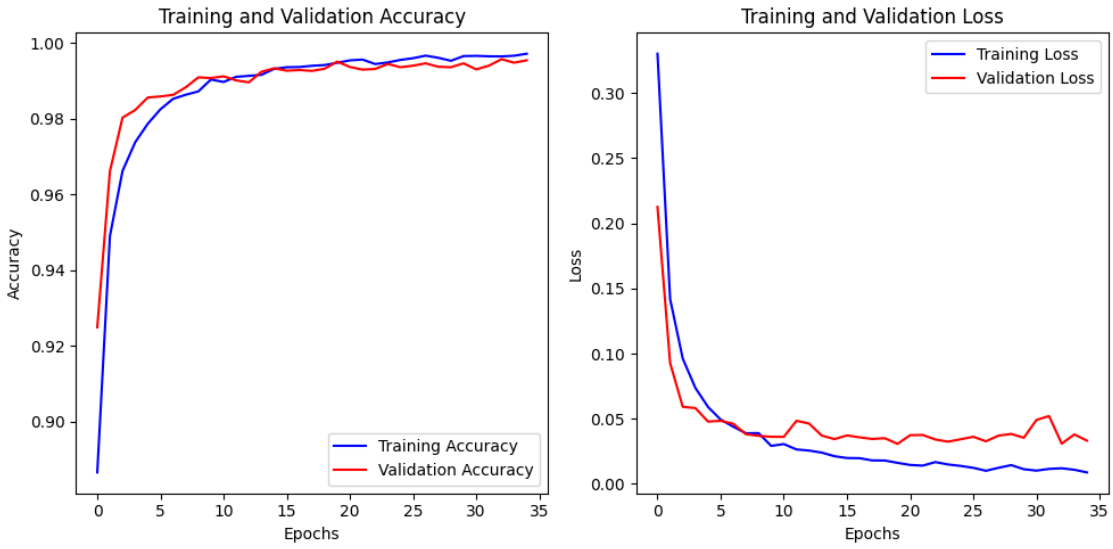
Figuur 4: Accuracy and Loss of Convolutional Neural Network



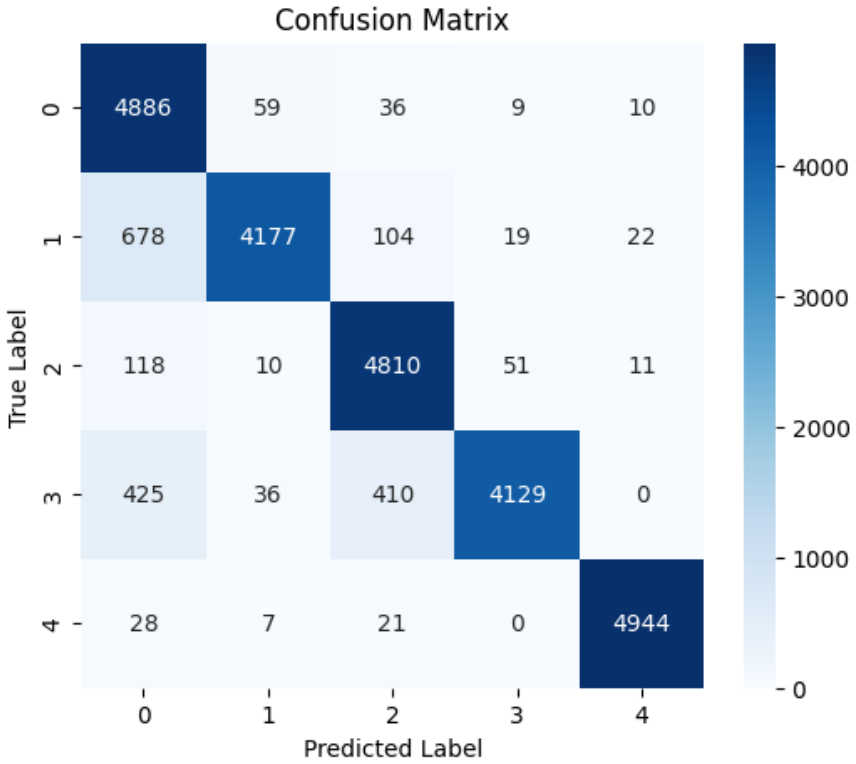
Figuur 5: Confusion Matrix of Convolutional Neural Network

4.3 Transformer

The Transformer model achieves high accuracy with a steady decline in training and validation loss, indicating effective learning. The confusion matrix shows strong classification performance across all heartbeat categories, with minimal misclassification. Compared to CNN and Random Forest, the Transformer demonstrates better generalization, particularly in distinguishing challenging classes like Supraventricular (Class 1) and Fusion Beats (Class 3).



Figuur 6: Accuracy and Loss of Transformer



Figuur 7: Confusion Matrix of Transformer