Report Practice 1: EGC Heartbeat Categorization using Random Forest and CNN

Nguyen Lam Tung - 23BI14446

***Abstract*—This study investigates the performance of Random Forest and Convolutional Neural Network (CNN) models for ECG heartbeat categorization using the MIT-BIH Arrhythmia Database. The dataset contains 109,446 samples across 5 classes (Normal, Supraventricular, Ventricular, Fusion, and Unknown) with significant class imbalance. We compare model performance on both original and balanced datasets, where the balanced version upsamples minority classes and downsamples the majority class. Results show that while overall accuracy remains at 0.97 for both datasets, the balanced approach significantly improves minority class detection. The Random Forest model demonstrates high performance with improved macro average recall from 0.81 to 0.90 on balanced data. The CNN model shows similar patterns, this proves that with proper data process and class imbalance handling, the model can achieve better performance.**

I. Introduction

This report presents the results of my practice on the ECG Heartbeat Categorization Dataset from Kaggle. The dataset is the combination of 2 datasets, the MIT-BIH Arrhythmia Dataset and the PTB Diagnostic ECG Database. In this study, I will focus only on the MIT-BIH Arrythmia Dataset which from now on called Arrhythmia Dataset

II. Dataset

The data used in this study is the MIT-BIH Arrhythmia Database. The dataset contains signals from 5 types of hearbeat problem: Normal, Supraventricular, Ventricular, Fusion and Unknown, which are encoded to 0, 1, 2, 3 and 4 respectively. Each observation contains 187 values of hearbeat signals

The dataset contains in total 109446 samples and is splitted into 2 files. The training file consists of 87554 samples, and the test file consists of 21892 samples

The class distribution is highly imbalanced, with class 0 having the most samples (72471) and class 3 having the least samples (641). This imbalance can cause problems for the models, as they may be biased towards the majority class and perform poorly on the minority classes. To address this issue, I applied upsampling on the minority classes (Class 1, 2, 3, 4) and downsampling on the majority classes (Class 0). The new class distribution is shown in Figure 2.

For the resampling size of each class, I choose an arbitrary size of 20000 for each class, this number is chosen because the resulting total size of the training dataset is 100000, which is close to the original size of the dataset (87554). This number makes sure the model can be trained fast enough on a Mac Mini M4 with 16GB of RAM.

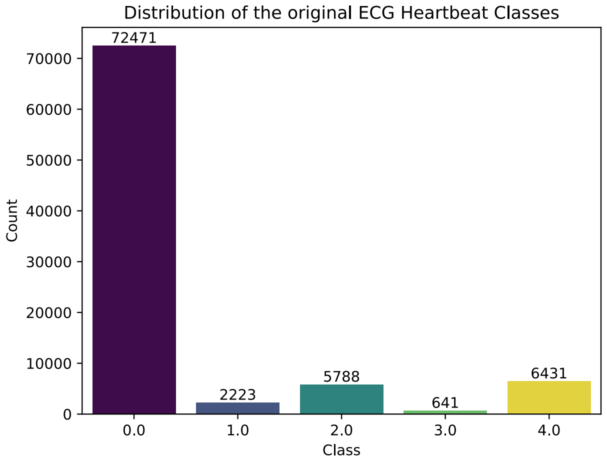

Fig. 1. Original class distribution in the training dataset

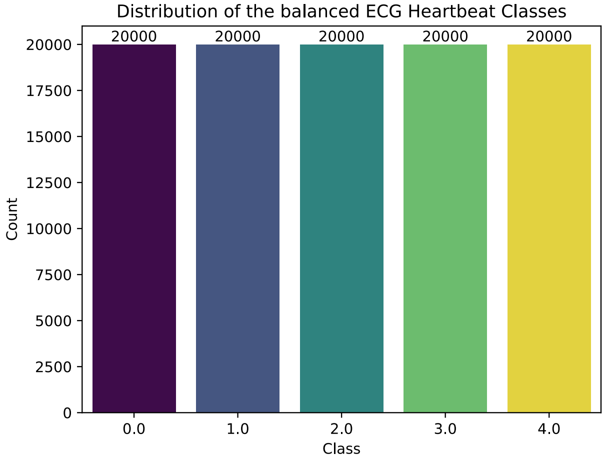

Fig. 2. Balanced class distribution in the training dataset

III. Methodology

In this study, I use 2 models to apply on the dataset. The first model is the Random Forest model, and the second model is a simple CNN model, I will also train the model on both the original and balanced datasets to compare the effect of the imbalance on the performance of the models.

IV. Results

A. Random Forest

TABLE I
CLASSIFICATION REPORT - RANDOM FOREST ON ORIGINAL DATASET

Class	Precision	Recall	F1-Score	Support
0	0.97	1.00	0.99	18118
1	0.98	0.60	0.75	556
2	0.98	0.88	0.93	1448
3	0.89	0.62	0.73	162
4	0.99	0.94	0.97	1608
Accuracy			0.97	21892
Macro avg	0.96	0.81	0.87	21892
Weighted avg	0.97	0.97	0.97	21892

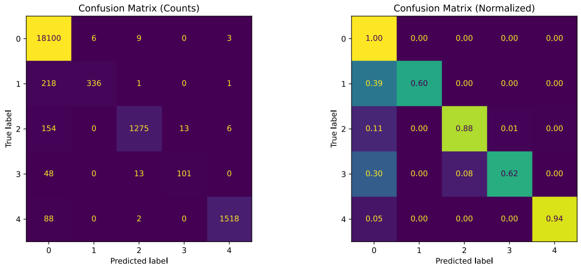

Fig. 3. Confusion matrix of the Random Forest model on the original dataset

1) Original dataset:

TABLE II
CLASSIFICATION REPORT - RANDOM FOREST ON BALANCED DATASET

Class	Precision	Recall	F1-Score	Support
0	0.99	0.98	0.99	18118
1	0.72	0.80	0.76	556
2	0.93	0.94	0.94	1448
3	0.69	0.81	0.74	162
4	0.98	0.97	0.98	1608
Accuracy			0.97	21892
Macro avg	0.86	0.90	0.88	21892
Weighted avg	0.97	0.97	0.97	21892

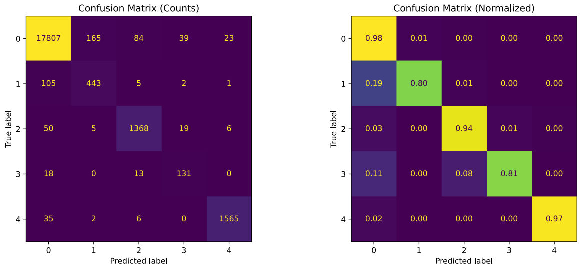

Fig. 4. Confusion matrix of the Random Forest model on the balanced dataset

2) Balanced dataset: As we can see from the results, while the overall accuracy of the model trained on the original dataset is the same as the model trained on the balanced dataset (0.97), the balanced model shows significant improvements in detecting minority classes. The recall for class 1 improved from 0.60 to 0.80, and class 3 improved from 0.62 to 0.81.

The improvement for the model trained on the balanced dataset comes at the cost of slightly lower accuracy for the model, but the trade-off is worth it as the model now can detect the minority classes much better.

B. CNN

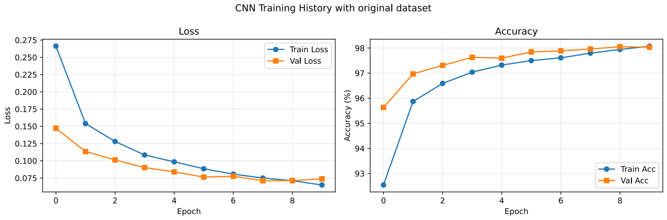

Fig. 5. Training history of the CNN model on the original dataset

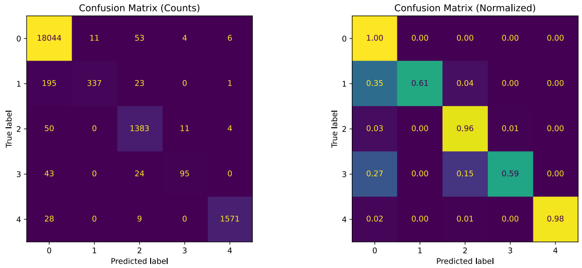

Fig. 6. Confusion matrix of the CNN model on the original dataset

1) Original dataset:

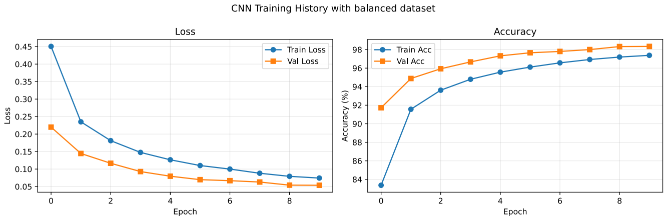

Fig. 7. Training history of the CNN model on the balanced dataset

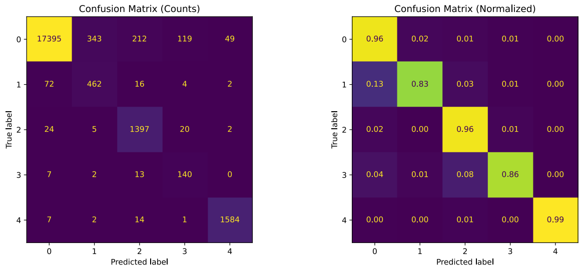

Fig. 8. Confusion matrix of the CNN model on the balanced dataset

2) Balanced dataset: We observe a similar pattern to the Random Forest model, the CNN model trained on the balanced dataset shows better performance in detecting minority classes. The CNN model achieves higher recall for class 1 (0.83) and class 3 (0.86) compared to the original dataset.

V. Conclusion

In this study, we have compared the performance of Random Forest and CNN models for ECG heartbeat categorization using the MIT-BIH Arrhythmia Database. We have shown that the balanced dataset can significantly improve the performance of the models in detecting minority classes. The Random Forest model shows better performance in detecting minority classes, while the CNN model shows signs of overfitting in the training history.

VI. Discussions

When comparing the two models trained on the balanced dataset, both models achieve the same overall accuracy of 0.97. However, the Random Forest model shows better performance in detecting minority classes. The Random Forest model achieves higher recall for class 1 (0.80) and class 3 (0.81) compared to the CNN model.