Machine Learning in Medicine
Practical Work 1

Ngô Anh Tâm - 23BI14392

January 2026

1 Introduction

This is a practical work for the first work in Machine Learning in Medicine. The dataset is provided by Kaggle.

In this section, the topic is the ECG Heartbeat Categorization Dataset, which is used to explore heartbeat classification using deep neural network architectures and observe some of the capabilities of transfer learning on it.

In this dataset, two Excel files are split into training and testing sets with a ratio of 4:1; the number of samples is 109,446, and there are 188 columns. Each row represents a sequence of ECG heartbeats over time. All data are numerical values, with the final column serving as the target to be classified. 5 classes required model to select which contain

- **0.0**: Normal
- **1.0**: Supraventricular Ectopic
- **2.0**: Ventricular Ectopic
- **3.0**: Fusion
- **4.0**: Unknown

2 Explore the dataset

(a) Missing values

There are no missing values in any rows after the dataset is checked, indicating that no values that could compromise the model's accuracy need to be filled in.

(b) Class distribution

This dataset consists of five classifications that we must distinguish. In the introduction, their names are mentioned. This part will look at the ratio of the five classes and count the number of each class in the training set.

The number of each classes are represented in this table:

Table 1: Class distribution of ECG heartbeat dataset

Class	Count	Ratio (%)
Normal	72,470	82.8
Supraventricular Ectopic	2,223	2.5
Ventricular Ectopic	5,788	6.6
Fusion	641	0.7
Unknown	6,431	7.3

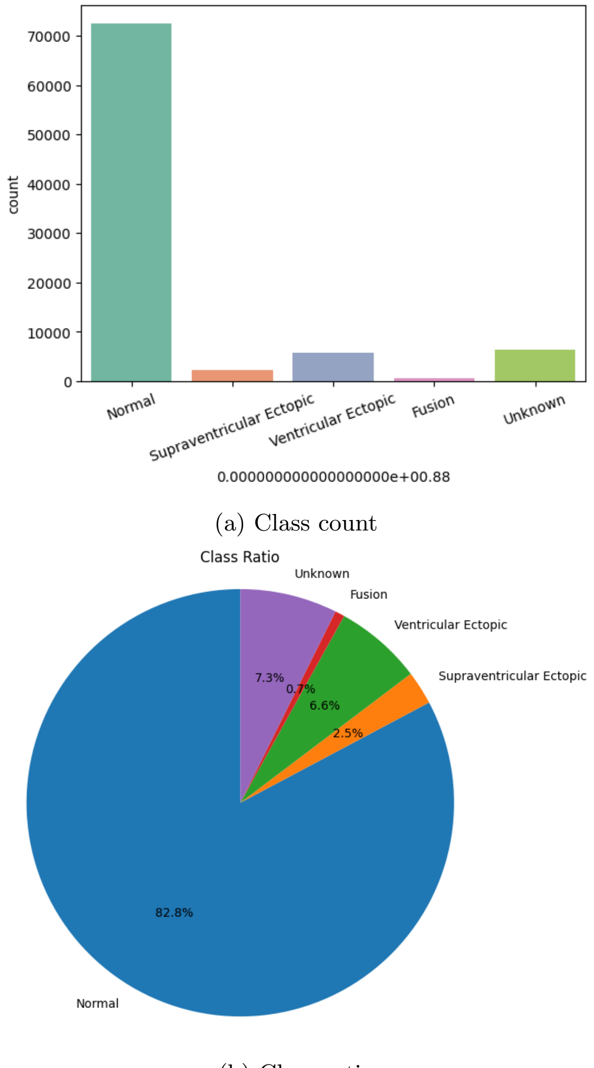
These charts display the number of each class as well as the ratio of each category.

As the charts illustrate, the count of normal classes dominates the dataset, approximately 82.8%, while the Fusion class represents only 0.7%. The remaining classes account for a small percentage, less than 8%.

3 Training model

1. Data processing

Since all data values in the dataset are numerical values in the range [0, 1], which indicates



(a) Class count

(b) Class ratio

Figure 1: ECG heartbeat classes

that the normalized value, the train dataset is prepared to train the classification model.

2. **Training**

For the machine learning algorithm used for the model, the Random Forest Classifier is applied because the dataset contains 109446 samples, which is large number, so the decision tree using the random part of the data, so that every tree is a bit different. Therefore, it helps reduce overfitting and makes the overall prediction more accurate and trustworthy.

The hyperparameter used for the algorithm includes the number of tree is 300, with a maximum depth and minimal leaf.

3. **Results**

After training the model, the accuracy is approximately 97.3%, which demonstrates the high efficiency in classifying the people who have abnormal heartbeats.

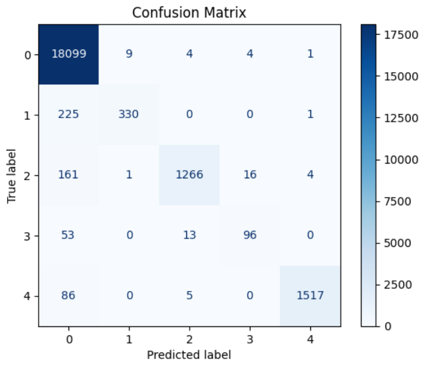


Figure 2: Confusion matrix of the model

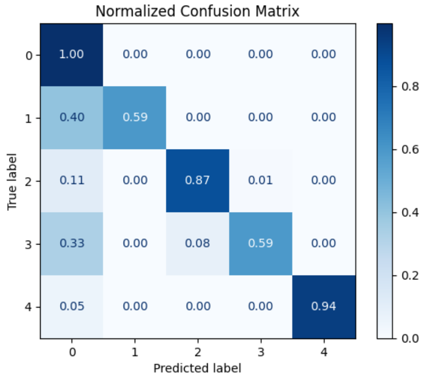


Figure 3: Normalized confusion matrix of the model

The accuracy of classifying the person's heartbeat as normal or unknown using the confusion matrices is excellent, at roughly 100% and 94%, respectively. People with ventricular ectopic heartbeat have an accuracy of around 87%, but the person with a supraventricular heartbeat and fusion has a low accuracy of about 59%.