Report on Lab Work 1

Nguyen Tien Cong
ICT Department - USTH
conkubenho27@email.com

***Abstract*—This paper explores the application of machine learning in medical diagnosis, focusing on the classification of ECG heartbeat signals using traditional machine learning model.**

I Introduction

In this paper, I introduce my work on the ECG Heartbeat dataset and compare my results with those presented in the original study [1]. The study applies a CNN model to classify different types of ECG signals.

II Datasets

In this section, I provide details about the dataset used in my research. The dataset was obtained from the Kaggle platform [2]. It consists of two subsets: the Arrhythmia Dataset and the PTB Diagnostic ECG Datasets. The datasets mined have been preprocessed.

All of sub two datasets have been recorded from 47 different subjects recorded at the sampling rate of 125Hz. Each beat is annotated by at least two cardiologists. with Arrhythmia, there have 5 categories divided into: N, S, V, F, Q. With PTB Diagnostic ECG, the data have been divided into 2 type: normal and abnormal.

III Methodology

This section describes the methodology used in our study, including data preprocessing, model architecture, and training procedures. However, I only work with Arrhythmia Datasets.

After a short exploratory data analysis phrase, i notices that this datasets is very imbalance, There are five type of heart signal as mention in the above sections. The N type are the dominant in overall, this might lead to model can not generalize well on the data. Here are the plot of distribution of categories:

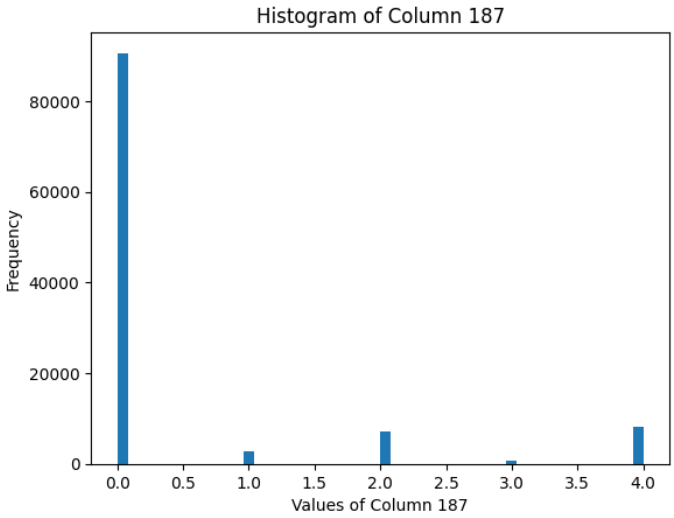

Fig. 1. distribution plot of categories(0:N, 1:S, 2:V, 3:F, 4:Q)

To overcome this imbalance in the datasets, i decided to use Synthetic Minority Over-sampling technique, this technique generates synthetic samples for the minority class instead of simply duplicating existing ones. The random rate 42 are chosen to fix the data.

As the result, i get a new datasets with balanced in categories. All data in each categories are the same and equal to 90589. Each category take exact 20 percent of the whole datasets. Unlike the data in the paper [3], their datasets are not balanced which lead to the poor performance of the CNNs Model they used for classify.

At the end, the datasets are split into 2 subset: test set and train set with the test set taking 20 percent of the datasets and 80 percent with the train set.

IV Experiments

This section presents the results and analysis of the experiments conducted, including model performance and evaluation metrics.

A Model

The model used in the experiment is Random Forest, a classical machine learning model. The primary reason for selecting Random Forest is its ability to handle high-dimensional and structured data effectively. In this study, the dataset consists of ECG (Electrocardiogram) signals.

The model parameter including:(1) number of estimator equal to 100 and (2)the random state are 36.

B Evaluation

The evaluation metric used for the model are Accuracy, Precision, F1 and Recall. The model are trained on Google Colab platform and taking around 30 minute to train.

The result are outstanding, with the train accuracy equal 99.7 percent, each of categories classification are 1.0 in Precision, F1 and Recall, only one category have 0.99 on recall and F1. Here are the table of the overall training:

TABLE I
ECG SIGNAL CATEGORIES AND DESCRIPTIONS

Category	Label	Precision	F1	Recall
Normal Beat	N	1.0	0.99	0.99
Supraventricular	S	1.0	1.0	1.0
Ventricular	V	1.0	1.0	1.0
Fusion Beats	F	1.0	1.0	1.0
Unknown	Q	1.0	1.0	1.0

V Conclusion

The model of this paper are beaten the CNN model on paper [3] with extremely high accuracy, it show that the CNNs model might not work well on signal datatype.

References

[1] J. Doe, “An Example Paper,” *IEEE Transactions on Something*, vol. 10, no. 3, pp. 123–130, 2022.

[2] A. Smith and B. Brown, “Deep Learning for ECG Classification,” in *Proc. IEEE Conf. on Machine Learning*, 2021, pp. 45–50.

[3] *ECG Heartbeat Classification: A Deep Transferable Representation*