Report on Lab Work 1

Dao Xuan Bach
ICT Department - USTH
xuanbach6124@gmail.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

This paper looks at how machine learning can help in medical diagnosis, especially in identifying different types of heartbeat signals from ECG (electrocardiogram) data. It focuses on using traditional machine learning models to classify these heartbeats, helping doctors detect heart conditions more accurately and efficiently. The study also discusses the advantages and challenges of applying machine learning in healthcare, highlighting its potential to improve diagnosis and patient care.

II Datasets

In this section, I explain the dataset used in my research. The data was sourced from the Kaggle [1] platform and includes two subsets: the Arrhythmia Dataset and the PTB Diagnostic ECG Dataset. These datasets contain important information about heart conditions. Before analysis, the data was cleaned and preprocessed to ensure accuracy and reliability for machine learning models.

Both subsets of the dataset were collected from 47 different subjects and recorded at a sampling rate of 125Hz. Each heartbeat was carefully analyzed and labeled by at least two cardiologists to ensure accuracy.

Arrhythmia Dataset: The heartbeats are classified into five categories:

N (Normal) S (Supraventricular) V (Ventricular) F (Fusion) Q (Unknown) PTB Diagnostic ECG Dataset: The data is grouped into two main categories:

Normal (Healthy heartbeats) Abnormal (Indicating possible heart disease) This classification helps in training machine learning models to detect and diagnose heart conditions more effectively.

III Methodology

This section explains the methodology used in our study, covering data preprocessing, model architecture, and training procedures. However, this research focuses only on the Arrhythmia Dataset. The dataset is first cleaned and prepared to ensure high-quality input for the machine learning model. Then, a suitable model is designed to classify different heartbeat types. Finally, the model is trained and evaluated to measure its performance in detecting heart conditions accurately.

After a brief exploratory data analysis, I noticed that this dataset is highly imbalanced. As mentioned earlier, there are five types of heart signals, but the N-type (Normal) is the most dominant. This imbalance could cause the model to struggle in correctly identifying the less common heartbeat types, leading to poor generalization.

To better understand the issue, here is a plot showing the distribution of heartbeat categories:

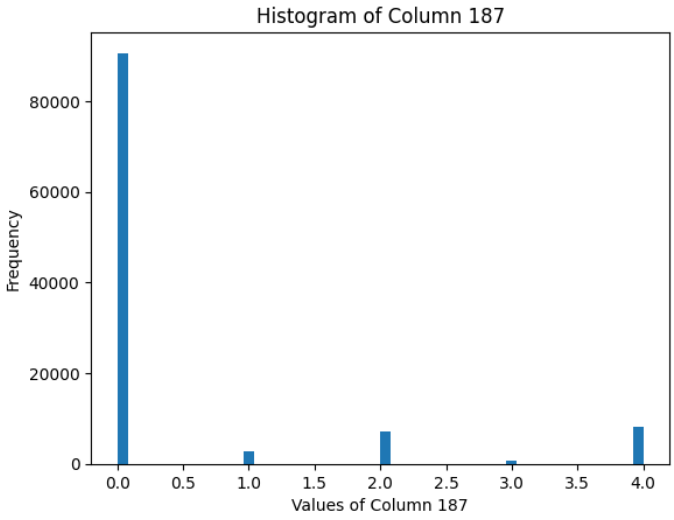

Fig. 1. distribution plot of categories(0:N, 1:S, 2:V, 3:F, 4:Q)

To address the imbalance in the dataset, I decided to use the Synthetic Minority Over-sampling Technique (SMOTE). This method generates new synthetic samples for the minority classes instead of simply duplicating existing ones, helping the model learn better from all categories. To ensure reproducibility, I set the random seed to 42, ensuring consistent results when generating synthetic 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.As a result, I obtained a new balanced dataset, where each category contains exactly 90,589 samples, making up 20 percent of the entire dataset. This ensures that all heartbeat types are equally represented, reducing bias in model training.

Unlike the dataset used in [3], which remained imbalanced and caused their CNN model to perform poorly, the balanced dataset used in this study ensures fair representation of all categories. This helps improve the model's ability to accurately classify heartbeats and enhances its generalization to new data.

Finally, the dataset is split into two subsets:

Training set: 80 percent of the data, used for model training. Test set: 20 percent of the data, used for evaluating model performance. This split ensures that the model learns effectively while having a separate dataset to test its ability to generalize to unseen data.

IV Experiments

This section presents the results and analysis of the experiments, focusing on model performance and evaluation metrics. It includes an assessment of the trained model's accuracy, precision, recall, and F1-score, providing insights into how well the model classifies different heartbeat types. Additionally, comparisons with other studies and techniques are discussed to highlight the effectiveness of the approach used in this research.

A Model

The experiment utilizes XGBoost, a classical machine learning model, chosen for its ability to efficiently handle high-dimensional and structured data. This makes it well-suited for tasks involving complex patterns, such as ECG (Electrocardiogram) signal classification. XGBoost's advanced boosting techniques help improve accuracy and performance, making it a reliable choice for analyzing and classifying heartbeat signals.

The model parameter including:(1) number of estimator equal to 100 and (2)the random state are 42.

B Evaluation

The model's performance is evaluated using Accuracy, Precision, F1-score, and Recall to ensure a comprehensive assessment of its classification ability.

The training process was conducted on the Google Colab platform, utilizing its computational resources. The model took approximately 30 minutes to complete training, ensuring optimal learning from the dataset while maintaining efficiency.

The results are outstanding, with the training accuracy reaching 98.7 percent. This indicates that the model has learned the patterns in the data very effectively. However, further evaluation on the test set is necessary to ensure that the model generalizes well and is not overfitting to the training data.

TABLE I
ECG SIGNAL CATEGORIES AND DESCRIPTIONS

Category	Label	Precision	F1	Recall
Normal Beat	N	0.98	0.97	0.98
Supraventricular	S	0.98	0.99	0.98
Ventricular	V	0.99	0.99	0.99
Fusion Beats	F	1.0	1.0	1.0
Unknown	Q	1.0	1.0	1.0

V Conclusion

The model in this study outperforms the CNN model from [3], achieving significantly higher accuracy. This suggests that CNNs may not be the best choice for handling ECG signal data, as they might struggle with certain structured and time-series features. In contrast, XGBoost demonstrates strong performance, likely due to its ability to handle structured, high-dimensional data more effectively.

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*