ECG Heartbeat Classification

Nguyen Ha Phuong BI12-359

I. Introduction

The heartbeat, a key physiological action, moves blood all through the circulatory framework. A basic cycle guarantees the vehicle of oxygen, supplements, and waste end. Understanding the intricacy of the heartbeat is basic to grasping by and large cardiovascular well-being.

In clinical diagnostics, assessing the heartbeat is a significant indication of cardiovascular well-being. Clinical experts find out about heart capability utilizing hardware like electrocardiograms, stethoscopes, and heartbeat checks. Heartbeat examination assists with distinguishing abnormalities, analyzing heart issues, and screening general cardiovascular well-being.

A kind of artificial intelligence called machine learning (ML) predicts results, finds patterns, automates procedures, and enhances performance over time—especially in the medical field.

In the context of heartbeat classification, machine learning algorithms examine data to classify heart rhythms into different categories, such as normal or irregular. The incorporation of machine learning in this sector intends to increase the early identification of problems, diagnostic accuracy, and treatment plan personalization.

Research on AI for heartbeat characterization utilizes different procedures, calculations, and model structures to upgrade precision and productivity, directing the advancement of new arrangements and recognizing improvement regions.

Despite advances in ML for heartbeat classification, problems remain. Data privacy, model interpretability, and the necessity for rigorous clinical validation all present important challenges. Addressing these problems is critical to ensuring the reliability and effectiveness of ML applications in cardiovascular health. A comprehension of the present status of the issue makes way for imaginative arrangements and progressions in the field.

II. Background

An electrocardiogram (ECG) is a harmless test that records heart electrical signs to distinguish heart issues and screen heart well-being. It can assist with diagnosing unpredictable heart rhythms, coronary vein illness, past respiratory failures, and the adequacy of coronary illness medicines. Indications of ECG need to incorporate chest torment, dazedness, heart palpitations, quick heartbeat, windedness, and shortcomings.

III. Method

The random forest algorithm is a common and reliable method for handling classification tasks in machine learning. To improve prediction accuracy and get beyond restrictions imposed by individual decision trees, it uses the combined knowledge of several decision trees.

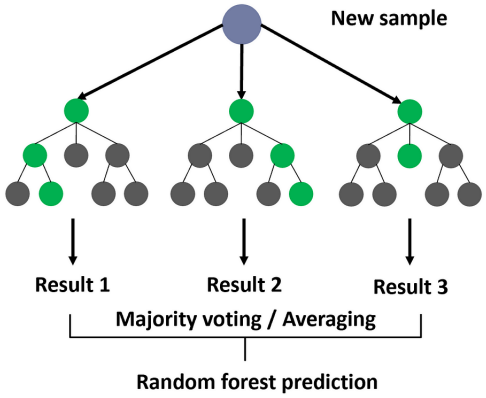

Fig. 1. Random forest architecture

Ensemble of Decision Trees: The random forest comprises a collection of individually trained decision trees that work together to classify data points. Each tree predicts based on the data it has been trained on.

Building Decision Trees: The structure concludes with 4 parts: Bootstrapping, random feature selection, node splitting, and leaf nodes. Bootstrapping introduces diversity by training trees using a random sample from the original data. Next is random feature selection at nodes, which enhances diversity. Then node splitting recursively separates data points using splitting criteria, creating branches. Finally, terminal nodes represent final class predictions for data points reaching that node.

Prediction Aggregation: Trees collectively predict class labels for new data points, with the most frequent class predicted by individual trees becoming the final classification.

The key points of the classification process involve dividing data points into distinct classes, maximizing purity at each node, and determining the final prediction through majority voting from individual trees.

Random forests provide enhanced generalizability to unknown data, robustness against overfitting, and better classification accuracy by utilizing the combined power of several distinct decision trees.

IV. Evaluation

A. Dataset

I use the ECG Heartbeat Categorization Dataset on the Kaggle. The dataset consists of two collections of heartbeat signals derived from two famous datasets in heartbeat classification, the MIT-BIH Arrhythmia Dataset and the PTB Diagnosis ECG Database. In this report, I only use the first collection mentioned, the MIT-BIH Arrhythmia Dataset.

The Arrhythmia Dataset has 109446 samples in total, which are divided into five categories 0 (N), 1 (S), 2 (V), 3 (F), and 4 (Q). The sampling frequency is 125Hz. The dataset is from Physionet's MIT-BIH Arrhythmia Dataset.

	1.0000000000000000e+00	7.582644820213317871e-01	1.1157024651765823 36e-01	0.0000000000000000e+00
0	0.908425	0.783883	0.531136	0.362637
1	0.730088	0.212389	0.000000	0.119469
2	1.000000	0.910417	0.681250	0.472917
3	0.570470	0.399329	0.238255	0.147651
4	1.000000	0.923664	0.656489	0.195929

Fig. 2. The dataset

There is an imbalance in the dataset visualization. This might result in overfitting in labels with a large number of images and insufficient sample sizes for learning, which could lead to inaccurate results. Overfitting and other potential issues can be avoided by upsampling the minority categories to balance the dataset and correct the imbalanced data distribution.

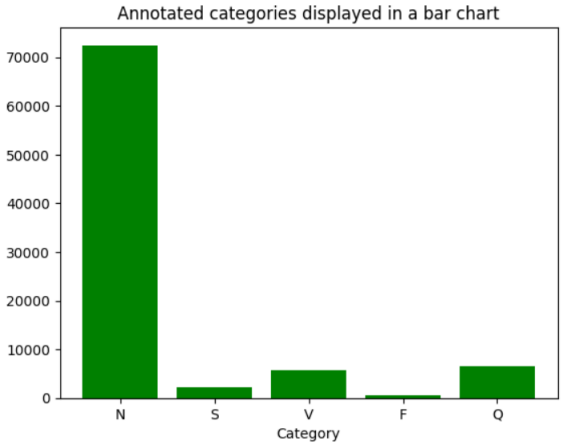

Fig. 3. The imbalanced dataset

B. Metric

Recall measures the capacity of the model by calculating the fraction of true positives among all actual positives. High recall means that the algorithm returned most of the relevant results.

$$Recall = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative(FN)}$$

Fig. 4. Recall metric

Precision quantifies the proportion of true positives among all positive predictions, assessing the model's capability to avoid false positives. High precision means the algorithm returned substantially more relevant results than irrelevant ones.

$$Precision = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Positive(FP)}$$

Fig. 5. Precision metric

```
Random Forest Classifier:
Accuracy: 0.9743273491389155
Classification Report:
              precision    recall  f1-score   support

         0.0       0.97      1.00      0.99     18117
         1.0       0.99      0.60      0.74       556
         2.0       0.98      0.88      0.93      1448
         3.0       0.88      0.62      0.73       162
         4.0       1.00      0.94      0.97      1608

    accuracy                           0.97     21891
   macro avg       0.96      0.81      0.87     21891
weighted avg       0.97      0.97      0.97     21891
```

Fig. 6. Result

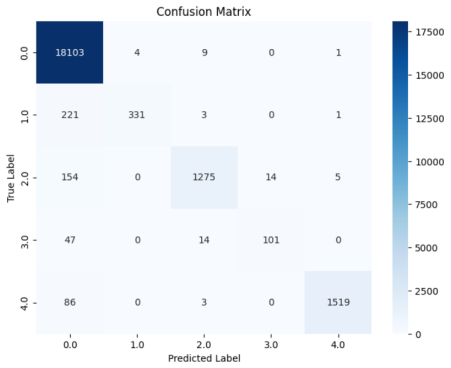

Fig. 7. Confusion matrix

C. Result

The result is not bad, the total accuracy is about 97 percent and class 1.0 has the most accuracy with 98 percent. However, the accuracy of class 1.0 is about 60 percent. It is mostly mistaken by class 0.0. Others have good accuracy but also usually being mistaken with class 0.0.

V. Conclusion

We explored the use of the random forest approach for ECG classification and gave an overview of the significance of the heartbeat in clinical diagnosis. However, the lack of particular examples and information is the difficulty of Machine Learning today. Combining deep learning approaches, developing real-time monitoring systems, addressing data privacy, model interpretability, and clinical validation should be the main areas of attention for future actions.