Labwork 1 - Biomedical Signals

Le Linh Long - 22BI13262

I. Introduction

Electrocardiogram (ECG) is a diagnostic tool to record the electrical activity of the heart. With the advancement of machine learning techniques, automated ECG analysis has gained significant attention in medical research. In this labwork, I will implement the Random Forest Classifier to classify different heartbeat types.

II. Exploratory Data Analysis

A. MIT-BIH Arrhythmia Dataset

On one hand, the MIT-BIH Arrhythmia Database includes **109,446** heartbeat samples with **188** numerical features and recorded at a sampling frequency of **125** Hz. The dataset is categorized into five classes:

- 0: Normal Beats
- 1: Supraventricular Ectopy Beats
- 2: Ventricular Ectopy Beats
- 3: Fusion Beats
- 4: Unclassifiable Beats

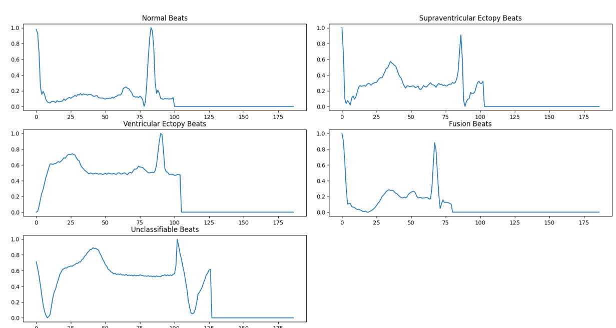

Fig. 1: Heartbeat Samples from Each Class

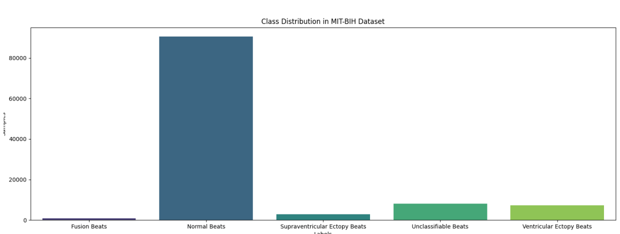

Fig. 2: Class Distribution in MIT-BIH Dataset

B. PTB Diagnostic ECG Dataset

On the other hand, the PTB Diagnostic ECG Dataset contains **14552** heartbeat samples with also **188** numerical features and a **125** Hz sampling frequency recorded. However, unlike the MIT-BIH dataset, this dataset is categorized only into two classes:

- 0: Normal
- 1: Abnormal

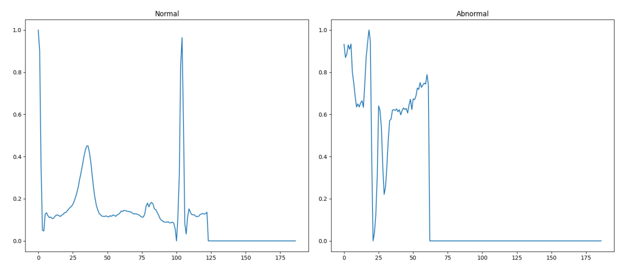

Fig. 3: Heartbeat Samples from Each Class

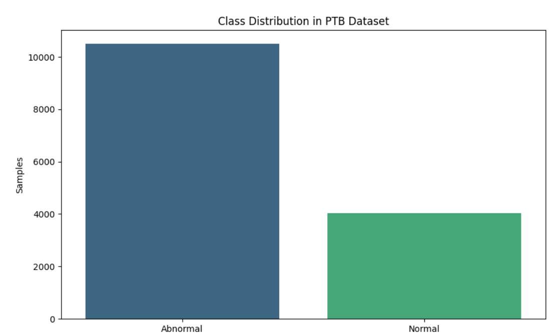

Fig. 4: Class Distribution in PTB Diagnostic ECG Dataset

As shown in Figures 2 and 4, there is a significant data imbalance in these datasets. In the MIT-BIH Arrhythmia Dataset, the Normal Beats class overwhelmingly dominates the others with more than **80,000** samples. In contrast, the PTB Diagnostic ECG Dataset witnessed the opposite trend, with the Abnormal class containing more than **10,000** samples — over twice the number of samples in the Normal class.

III. Methods

A. Principal Component Analysis

Both datasets contain a large number of features (**188**), which may lead to high computational costs. To avoid that, I will apply **Principal Component Analysis** (PCA) to reduce the dimensionality while preserving the most important patterns and relationships in the data. In this case, the threshold for the cumulative explained variance is set to **95%** to balance information preservation with reduced computational costs.

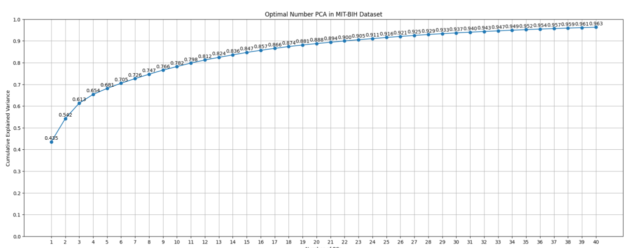

Fig. 5: Optimal Number of Principal Components in the MIT-BIH Dataset

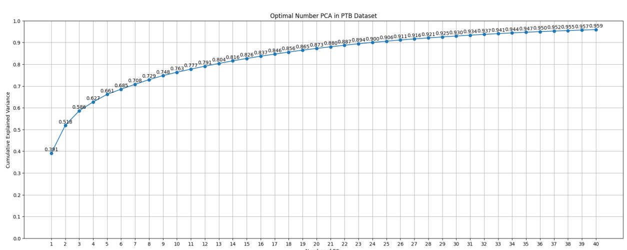

Fig. 6: Optimal Number of Principal Components in the PTB Diagnostic ECG Dataset

As shown in Figures 5 and 6, the number of principal components selected for the MIT-BIH and PTB Diagnostic datasets is 35 and 37, respectively. *(Note: In the PTB Diagnostic dataset, although the graph displays a value of 0.950 at PC 36, the actual value is 0.949.)*

B. Random Forest Classifier

In this labwork, I chose the Random Forest Classifier because it effectively handles large datasets and prevents overfitting by making predictions based on the majority vote or average of many decision trees. To further experiment, I experimented with the number of estimators set to 100, 200, and 300.

IV. Model Evaluation

A. Performance on MIT-BIH Dataset (PCA)

Class	Precision	Recall	F1-Score	Support
Fusion Beats	0.86	0.58	0.69	162
Normal Beats	0.97	1.00	0.98	18,118
Supraventricular Ectopy Beats	0.99	0.58	0.73	556
Unclassifiable Beats	1.00	0.92	0.96	1,608
Ventricular Ectopy Beats	0.97	0.86	0.91	1,448
Accuracy	-	-	0.97	21,892
Macro Avg	0.96	0.79	0.85	21,892
Weighted Avg	0.97	0.97	0.97	21,892

TABLE I: Classification Report at 100 Estimators

Class	Precision	Recall	F1-Score	Support
Fusion Beats	0.86	0.58	0.69	162
Normal Beats	0.97	1.00	0.98	18,118
Supraventricular Ectopy Beats	0.99	0.58	0.73	556
Unclassifiable Beats	1.00	0.92	0.96	1,608
Ventricular Ectopy Beats	0.97	0.86	0.91	1,448
Accuracy	-	-	0.97	21,892
Macro Avg	0.96	0.79	0.86	21,892
Weighted Avg	0.97	0.97	0.97	21,892

TABLE II: Classification Report at 200 Estimators

Class	Precision	Recall	F1-Score	Support
Fusion Beats	0.86	0.57	0.69	162
Normal Beats	0.97	1.00	0.98	18,118
Supraventricular Ectopy Beats	0.99	0.58	0.74	556
Unclassifiable Beats	1.00	0.92	0.96	1,608
Ventricular Ectopy Beats	0.97	0.86	0.91	1,448
Accuracy	-	-	0.97	21,892
Macro Avg	0.96	0.79	0.86	21,892
Weighted Avg	0.97	0.97	0.97	21,892

TABLE III: Classification Report at 300 Estimators

B. Performance on MIT-BIH Dataset (No PCA)

Class	Precision	Recall	F1-Score	Support
Fusion Beats	0.88	0.64	0.74	162
Normal Beats	0.97	1.00	0.99	18,118
Supraventricular Ectopy Beats	0.99	0.61	0.75	556
Unclassifiable Beats	0.99	0.94	0.97	1,608
Ventricular Ectopy Beats	0.98	0.88	0.93	1,448
Accuracy	-	-	0.97	21,892
Macro Avg	0.96	0.81	0.87	21,892
Weighted Avg	0.97	0.97	0.97	21,892

TABLE IV: Classification Report at 100 Estimators

Class	Precision	Recall	F1-Score	Support
Fusion Beats	0.88	0.64	0.74	162
Normal Beats	0.97	1.00	0.99	18,118
Supraventricular Ectopy Beats	0.99	0.61	0.75	556
Unclassifiable Beats	0.99	0.94	0.97	1,608
Ventricular Ectopy Beats	0.98	0.88	0.93	1,448
Accuracy	-	-	0.97	21,892
Macro Avg	0.96	0.81	0.88	21,892
Weighted Avg	0.97	0.97	0.97	21,892

TABLE V: Classification Report at 200 Estimators

Class	Precision	Recall	F1-Score	Support
Fusion Beats	0.88	0.63	0.73	162
Normal Beats	0.97	1.00	0.99	18,118
Supraventricular Ectopy Beats	0.99	0.60	0.75	556
Unclassifiable Beats	0.99	0.94	0.97	1,608
Ventricular Ectopy Beats	0.98	0.88	0.93	1,448
Accuracy	-	-	0.97	21,892
Macro Avg	0.96	0.81	0.87	21,892
Weighted Avg	0.97	0.97	0.97	21,892

TABLE VI: Classification Report at 300 Estimators

It is clear that increasing the number of estimators does not significantly enhance the performance of either model. However, applying PCA generally reduces performance. The best-performing model is the Random Forest Classifier with 200 estimators and no PCA.

C. Performance on PTB Diagnostic ECG Dataset (PCA)

Class	Precision	Recall	F1-Score	Support
Abnormal	0.95	0.97	0.96	2,112
Normal	0.92	0.85	0.89	799
Accuracy	-	-	0.94	2,911
Macro Avg	0.93	0.91	0.92	2,911
Weighted Avg	0.94	0.94	0.94	2,911

TABLE VII: Classification Report at 100 Estimators

Class	Precision	Recall	F1-Score	Support
Abnormal	0.95	0.97	0.96	2,112
Normal	0.93	0.86	0.89	799
Accuracy	-	-	0.94	2,911
Macro Avg	0.94	0.92	0.93	2,911
Weighted Avg	0.94	0.94	0.94	2,911

TABLE VIII: Classification Report at 200 Estimators

Class	Precision	Recall	F1-Score	Support
Abnormal	0.95	0.98	0.96	2,112
Normal	0.93	0.87	0.90	799
Accuracy	-	-	0.95	2,911
Macro Avg	0.94	0.92	0.93	2,911
Weighted Avg	0.95	0.95	0.95	2,911

TABLE IX: Classification Report at 300 Estimators

D. Performance on PTB Diagnostic ECG Dataset (No PCA)

Class	Precision	Recall	F1-Score	Support
Abnormal	0.98	0.99	0.98	2,149
Normal	0.98	0.94	0.96	762
Accuracy	-	-	0.98	2,911
Macro Avg	0.98	0.96	0.97	2,911
Weighted Avg	0.98	0.98	0.98	2,911

TABLE X: Classification Report at 100 Estimators

Class	Precision	Recall	F1-Score	Support
Abnormal	0.98	0.99	0.98	2,149
Normal	0.97	0.94	0.96	762
Accuracy	-	-	0.98	2,911
Macro Avg	0.98	0.96	0.97	2,911
Weighted Avg	0.98	0.98	0.98	2,911

TABLE XI: Classification Report at 200 Estimators

Class	Precision	Recall	F1-Score	Support
Abnormal	0.98	0.99	0.99	2,149
Normal	0.97	0.94	0.96	762
Accuracy	-	-	0.98	2,911
Macro Avg	0.98	0.97	0.97	2,911
Weighted Avg	0.98	0.98	0.98	2,911

TABLE XII: Classification Report at 300 Estimators

The model's performance remains almost unchanged across three different estimators, and applying PCA negatively impacts its results. The best-performing model is the Random Forest Classifier with 100 estimators and no PCA.

V. Results Comparison

In this labwork, I will compare my results with the paper **”ECG Heartbeat Classification: A Deep Transferable Representation”**, which utilizes a Deep Residual CNN to classify heartbeat signals in MIT-BIH dataset.

Predicted Label / True Label	F	N	S	Q	V
F (Fusion Beats)	**0.86**	0.08	0.00	0.00	0.05
N (Normal Beats)	0.00	**0.97**	0.01	0.00	0.00
S (Supraventricular Ectopy Beats)	0.00	0.08	**0.89**	0.00	0.00
Q (Unclassifiable Beats)	0.00	0.00	0.00	**0.98**	0.00
V (Ventricular Ectopy Beats)	0.00	0.02	0.00	0.00	**0.96**

TABLE XIII: Deep Residual CNN Approach

Predicted Label / True Label	F	N	S	Q	V
F (Fusion Beats)	**0.64**	0.28	0.00	0.00	0.07
N (Normal Beats)	0.00	**1.00**	0.00	0.00	0.00
S (Supraventricular Ectopy Beats)	0.00	0.39	**0.61**	0.00	0.00
Q (Unclassifiable Beats)	0.00	0.05	0.00	**0.94**	0.00
V (Ventricular Ectopy Beats)	0.01	0.11	0.00	0.00	**0.88**

TABLE XIV: My Best Random Forest Classifier Approach

Although my both Random Forest Classifiers returns a better results in Normal Beats class, the Deep Residual CNN consistently outperformed my model across the other beat types.

VI. CONCLUSION

To conclude, while the Random Forest Classifier did not outperform the Deep Residual CNN, it still achieved decent results, given its simplicity. Future work will focus on addressing imbalanced data and exploring deep learning models to further enhance classification performance.

REFERENCES

[1] Mayoclinic, ”Electrocardiogram (ECG or EKG)”
[2] Kaggle, ”ECG Heartbeat Categorization Dataset”
[3] Geeks for Geeks, ”How to Normalize a Confusion Matrix”
[4] Geeks for Geeks, ”Random Forest Algorithm in Machine Learning”
[5] Baeldung, ”How Many Principal Components to Take in PCA?”
[6] Mohammad Kachuee, Shayan Fazeli and Majid Sarrafzadeh, ”ECG Heartbeat Classification: A Deep Transferable Representation”