FINAL PROJECT

UTILIZING MULTI-DATASET ECG ANALYSIS APPLYING DEEP NEURAL NETWORKS (DNNS), SUPPORT VECTOR MACHINES (SVMS), AND RANDOM FOREST (RF) FOR HEARTBEAT CLASSIFICATION AND CARDIAC ANOMALY DETECTION

August 2024

COMP3035

Table Of Content

Ngo Thi Thu Suong

22052908

To Mai Huong

22050558

Background - Introduction

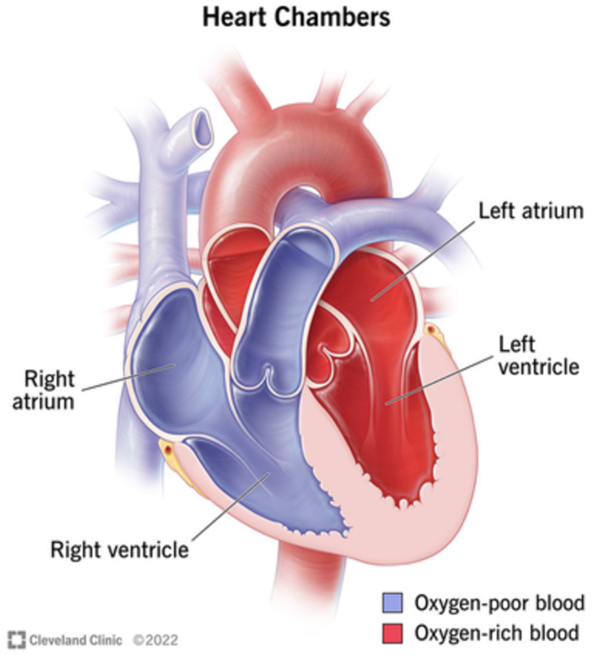
Importance of ECGs in Cardiovascular Health

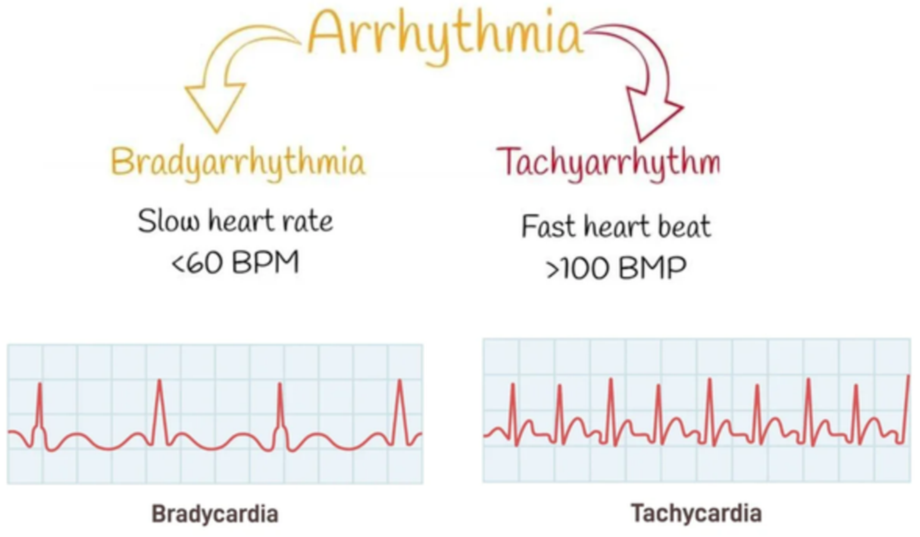
- Heart arrhythmia is a leading cause of mortality globally. Electrocardiography (ECG) is crucial for detecting arrhythmias (WHO, 2024).

It is desirable to diagnose arrhythmic heartbeats accurately and affordably

ECG Analysis Process:

- Electrocardiograms (ECGs) signals represent heart muscle contractions.
- The heart is a four-chambered pump with two ventricles for blood outflow and two atria for blood collection, with systole being the contracting phase and diastole being the resting phase.
- Tachycardia is a heartbeat exceeding 100 beats per minute (bpm), while bradycardia is a pulse below 60 bpm.





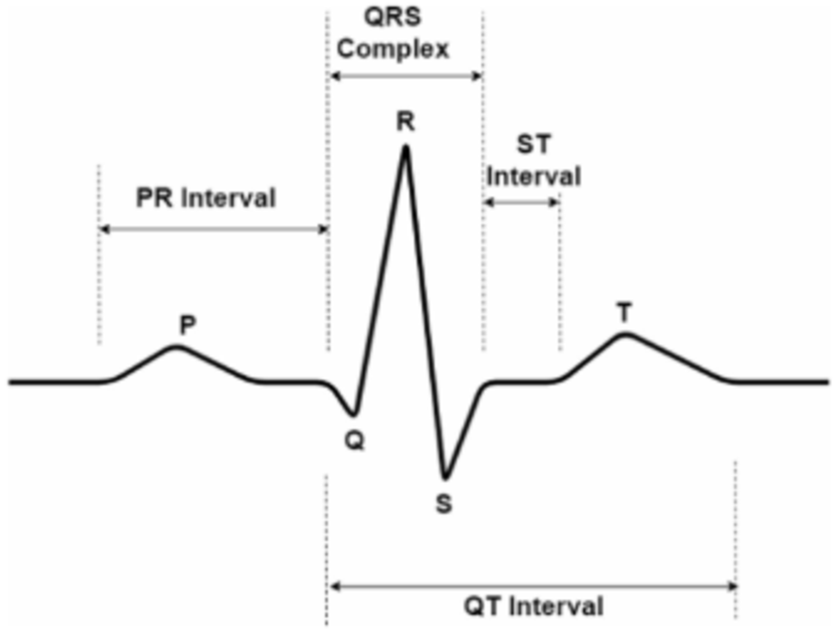
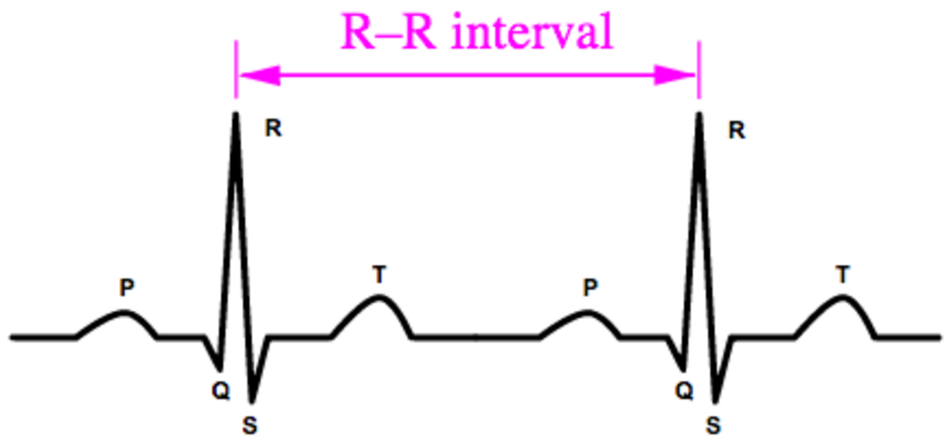


Figure 2: A typical heartbeat ECG signal contains P, Q, R, S, and T waves. A QRS complex is a combination of Q, R, S waves.

- Three waves make up an ECG signal (Acharya et al., 2017):
  - The P wave, which indicates atrial depolarization
  - The QRS complex wave, which indicates ventral depolarization.
  - T wave, which indicates repolarization.

- Heart signals are converted to heart rate through ECG data collection, preprocessing, segmentation, and R-peaks identification, in which two R-peaks are called RR intervals according to Liu & Wang (2021).



Background - Literature Review

Support Vector Machine (SVM)

2004

Osows et al.'s 2004 study:

• SVM for arrhythmia classification, achieving high accuracy through feature extraction techniques on ECG signals.

2009

Ince et al.'s 2009 demonstrated that:

• Using SVM combined with discrete wavelet transform (DWT) for feature extraction has shown promising results in detecting arrhythmias.

Random Forest (RF)

2006

Rodriguez et al., 2006 demonstrated that:

• Combining RF with feature selection and data augmentation significantly enhances ECG analysis effectiveness.

2017

Islam et al., 2017 demonstrated that:

• RF can accurately classify ECG signals into normal and abnormal categories with minimal feature engineering.

Deep Neural Network (DNNs)

- By employing advanced techniques such as DNNs, the accuracy of ECG heartbeat classification can be significantly enhanced (Strodthoff et al., 2021)

2017

Rajpurkar et al.'s 2017 study:

• Utilized CNNs for high arrhythmia classification accuracy.

• Demonstrated deep learning's effectiveness in ECG data processing.

2019

Hannun et al., 2019 demonstrated that:

• Deep neural network outperforms cardiologists.

• Highlights AI's potential in cardiac diagnostics.

2020

The study by Xie et al. (2020) highlights the importance of timely and accurate diagnoses in reducing stroke incidence and improving patient outcomes.

OBJECTIVE

HEART CARE AI

TO SAVE AND PROTECT YOUR HEALTH

- Utilize classification models and deep neural networks to enhance the accuracy and reliability of ECG data analysis.
- Aim to effectively differentiate between healthy heart signals and those that may indicate cardiac abnormalities.

➡ **Identify normal vs abnormal heart signals and arrhythmias by comparing traditional techniques such as SVM and RF with modern deep learning DNN.**
Proposing Heart Care AI platform: Enhancing Cardiac Diagnosis with Deep Learning.

HYPOTHESIS

"Given the heart signals (property: MLII), can ECG graphs accurately identify and determine the heart rate level normal or abnormal for each patient?"

- By providing empirical evidence supporting or refuting the hypothesis that ECG graphs, based on MLII signals, can be a reliable tool for assessing heart rate abnormalities, contributing to the advancement of diagnostic capacities in cardiovascular healthcare.

STUDY'S METHODOLOGY

Dataset

Dataset A: MIT-BIH Arrhythmia Database

- **Dataset Overview**: including 650,000 rows and 3 columns recordings (48 half-hour excerpts from two-channel ECGs).
- **Data Collection**: collected from 47 patients at 125 Hz between 1975 and 1979
- **Beat Categories**: The beat classification system used, referencing the AAMI EC57 standard and the five beat categories

Category	Annotations
N	• Normal • Left/Right bundle branch block • Atrial escape • Nodal escape
S	• Supraventricular premature • Aberrant atrial premature • Nodal premature
V	• Premature ventricular contraction • Ventricular escape
F	• Fusion of ventricular and normal
Q	• Paced • Fusion of paced and normal • Unclassified

Sample for Dataset A

```
   sample #  MLII   V1
0         0   955  992
1         1   955  992
2         2   955  992
3         3   955  992
4         4   955  992
<class 'str'>
```

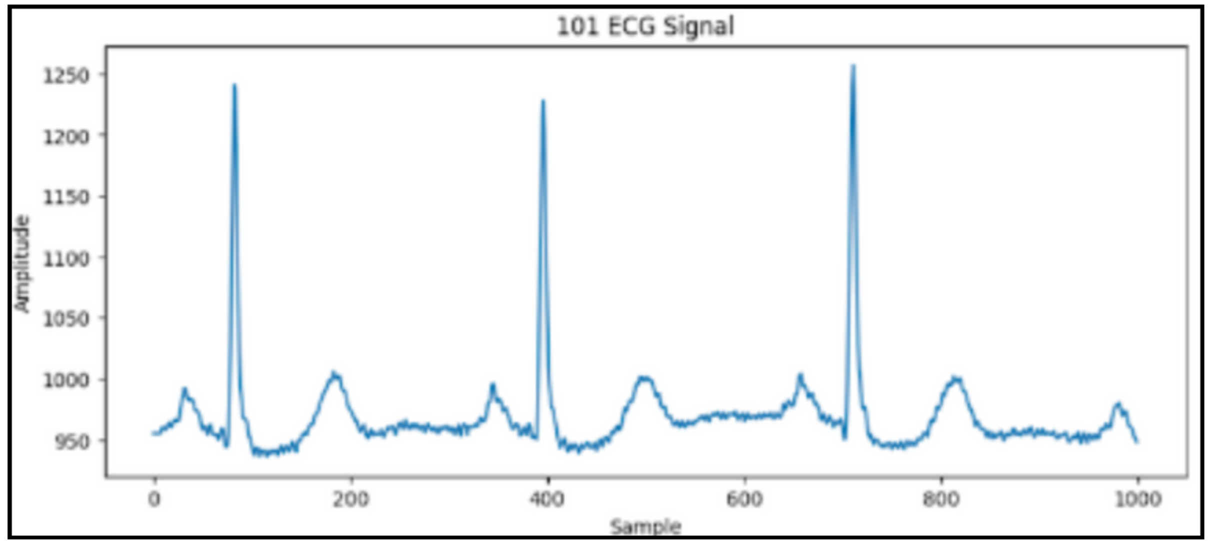
Data structure for Dataset A

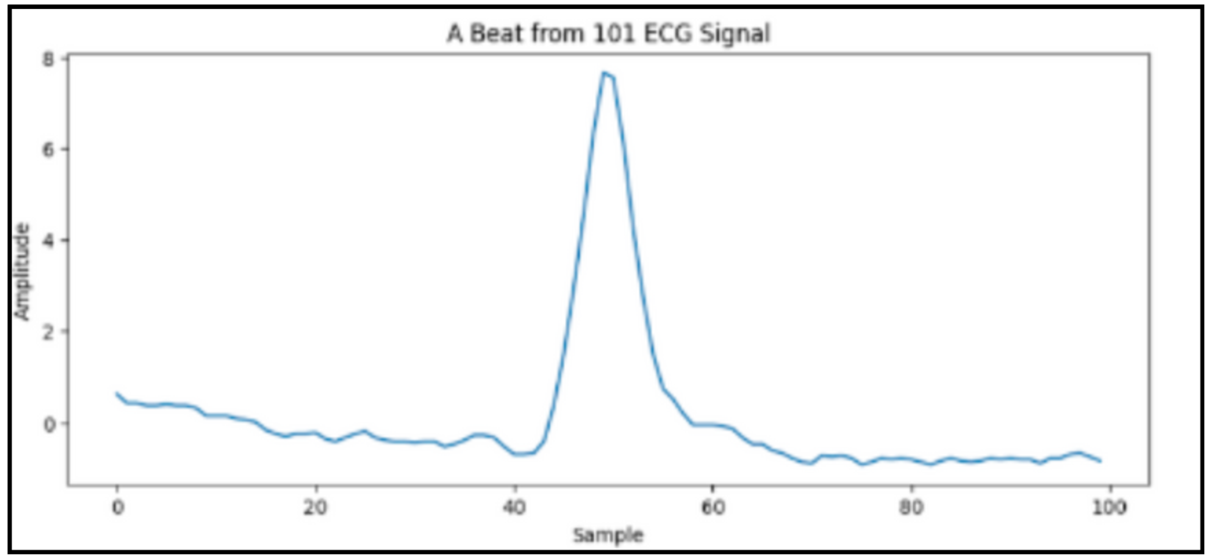
	sample #	MLII	V1
count	650000.000000	650000.000000	650000.000000
mean	324999.500000	967.155606	1008.377115
std	187638.981824	52.231329	9.882962
min	0.000000	389.000000	911.000000
25%	162499.750000	944.000000	1002.000000
50%	324999.500000	958.000000	1008.000000
75%	487499.250000	977.000000	1014.000000
max	649999.000000	1508.000000	1171.000000

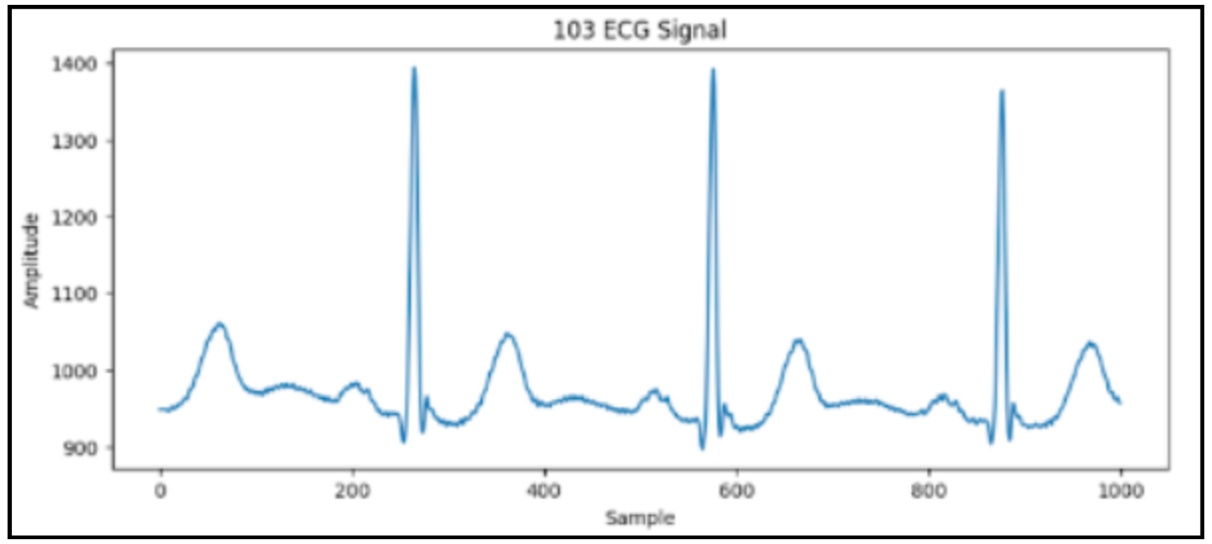
VISUALIZATION

Dataset A: MIT-BIH Arrhythmia Database

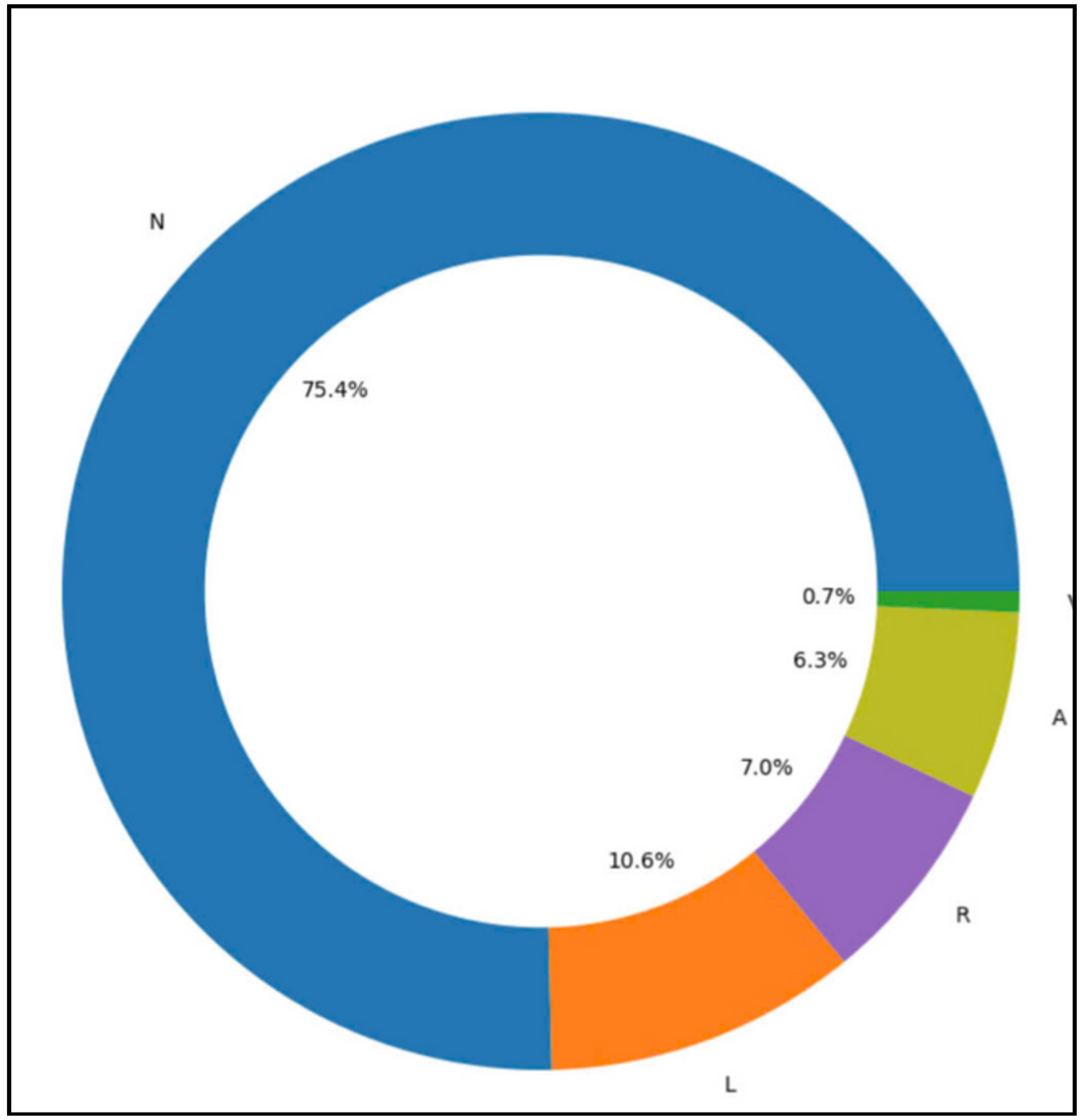
A typical heartbeat ECG signal



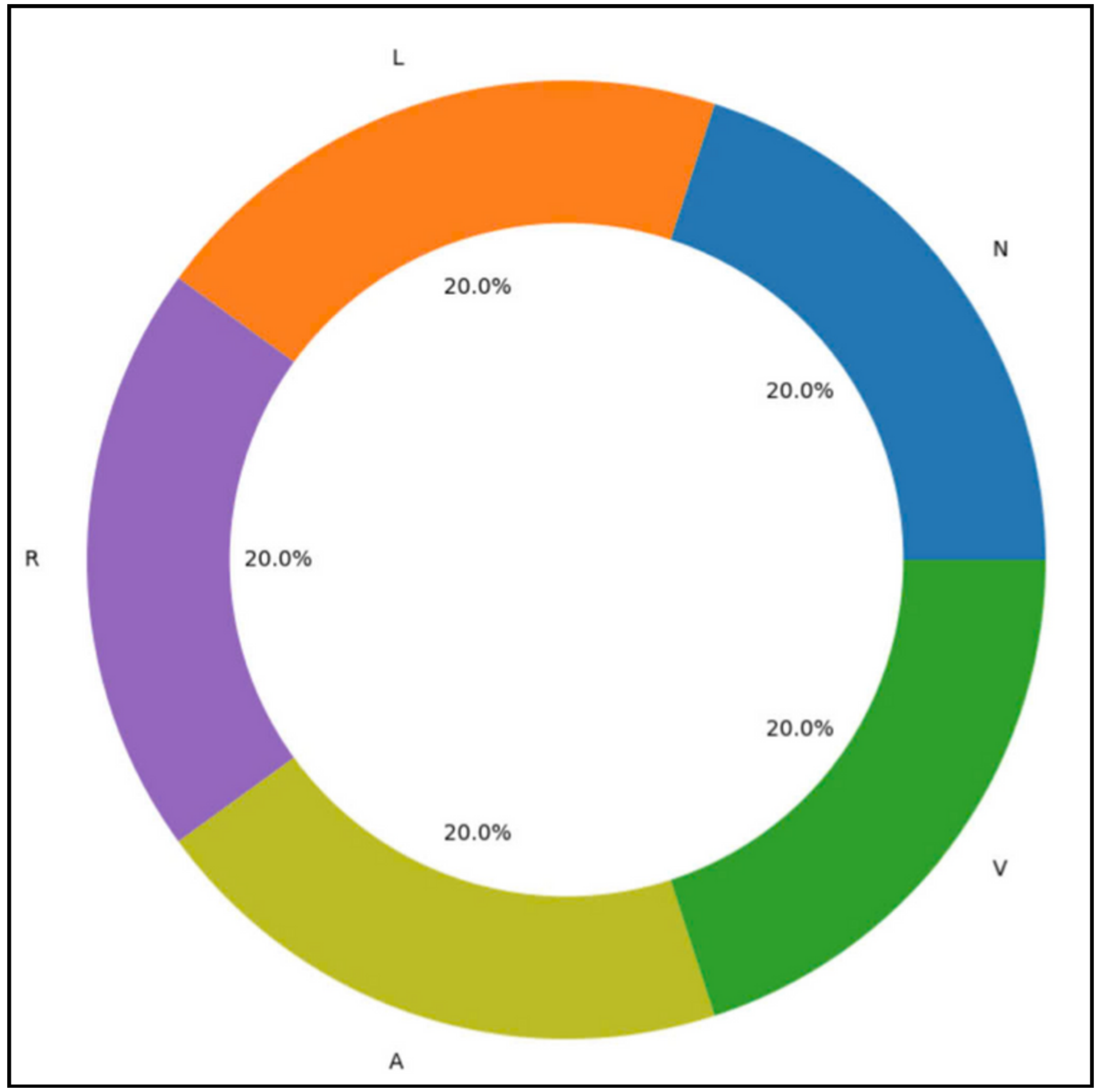




Class Distribution in ECG Data



Class Distribution After Upsampling



Unbalanced data

75.4% of the dataset is normal beats, with other classes underrepresented

Upsampling data

Five different heartbeat classes, each occupying **20%** of the total data.

label	
0	5000
1	5000
2	5000
3	5000
4	5000

STUDY'S METHODOLOGY

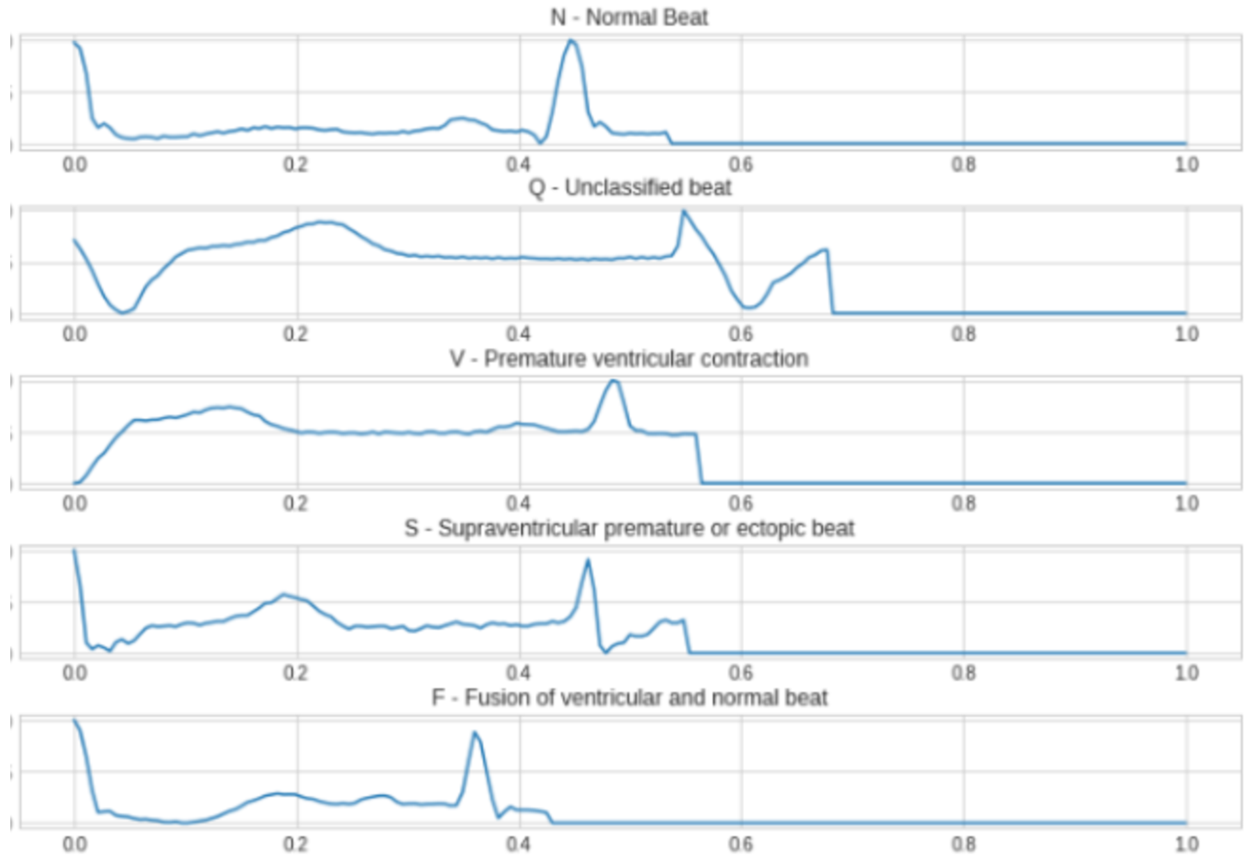
Dataset

Dataset B: Shayan Fazeli's Heartbeat Dataset

- **Dataset Overview:** combining heartbeat signals from the MIT-BIH Arrhythmia Database
- **Classification Categories**: Five categories used for heartbeat classification
- **Preprocessing and Usability**: the data is preprocessed and pre-split into training and testing sets

Number of Samples:	109,446
Number of Categories:	5
Sampling Frequency:	125 HZ

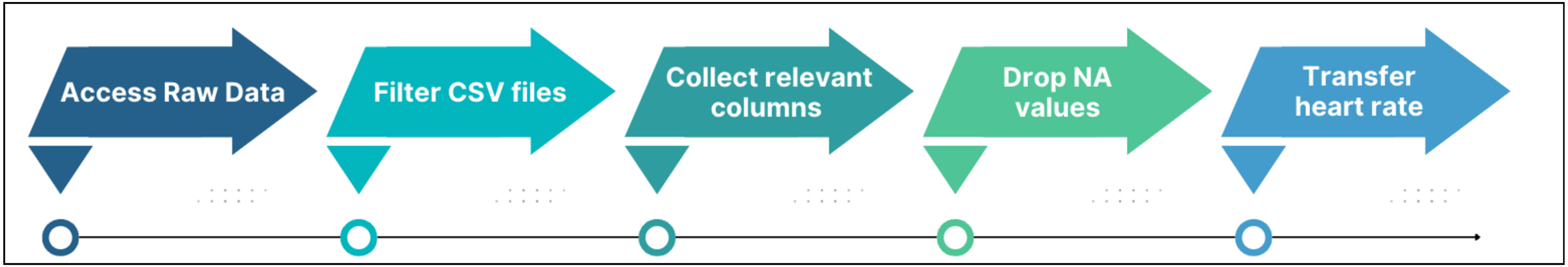
Category	Annotations
N	• Normal
S	• Supraventricular premature
V	• Ventricular escape
F	• Fusion beat
Q	• Unclassified



Sample for Dataset B

5752	1	0.982659	0.763006	0.526012	0.277457	0.144509	0.132948	0.132948	0.115607	0.150289	0.16763	0.16763	0.150289	0.17341
5753	0.929825	0.584795	0.403509	0.157895	0	0.00584795	0.134503	0.175439	0.19883	0.175439	0.233918	0.210526	0.251462	0.245614
5754	0.968421	0.842105	0.463158	0.196491	0.0877193	0.0736842	0.0491228	0.0561404	0.0245614	0.045614	0.0175439	0.0385965	0.0245614	0.045614
5755	0.941275	0.813758	0.276846	0	0.139262	0.234899	0.260067	0.283557	0.276846	0.27349	0.290268	0.285235	0.280201	0.286913
5756	1	0.931472	0.520305	0.111675	0.137056	0.185279	0.190355	0.172589	0.175127	0.164975	0.175127	0.14467	0.152284	0.15736
5757	0.966507	0.411483	0	0.148325	0.138756	0.110048	0.110048	0.114833	0.119617	0.124402	0.124402	0.133971	0.124402	0.143541
5758	0.116402	0.128748	0.146384	0.201058	0.229277	0.250441	0.276896	0.29806	0.292769	0.29806	0.282187	0.27866	0.262787	0.262787
5759	0.67027	0.362162	0.0324324	0.0864865	0.0486486	0.0432432	0	0.0324324	0.0162162	0.0486486	0.0486486	0.0486486	0.0216216	0.0378378
5760	1	0.802632	0.335526	0.111842	0.0394737	0.0197368	0	0.0197368	0.00657895	0.0197368	0.0460526	0.0657895	0.0394737	0.0789474
5761	0.855908	0.829971	0.538905	0.037464	0	0.074928	0.112392	0.135447	0.126801	0.132565	0.15562	0.152738	0.152738	0.161383

EXPLORATORY DATA ANALYSIS



1. Access Raw Data And Choose Relevant Collumns

Loading and preprocessing ECG data from **CSV files,** emphasizing the extraction of the 'MLII' column for accurate signal measurement.

- Remove all columns named "V1", "V2", "V3", or "V5".

Reason

- **"**MLII" column provides more accurate measurements
- enhance the depth and usefulness

2. Heart Rate Conversion and Label Data

Calculate heart rate with formula:

- Heart Rate (BPM) = 60 / (average RR interval).

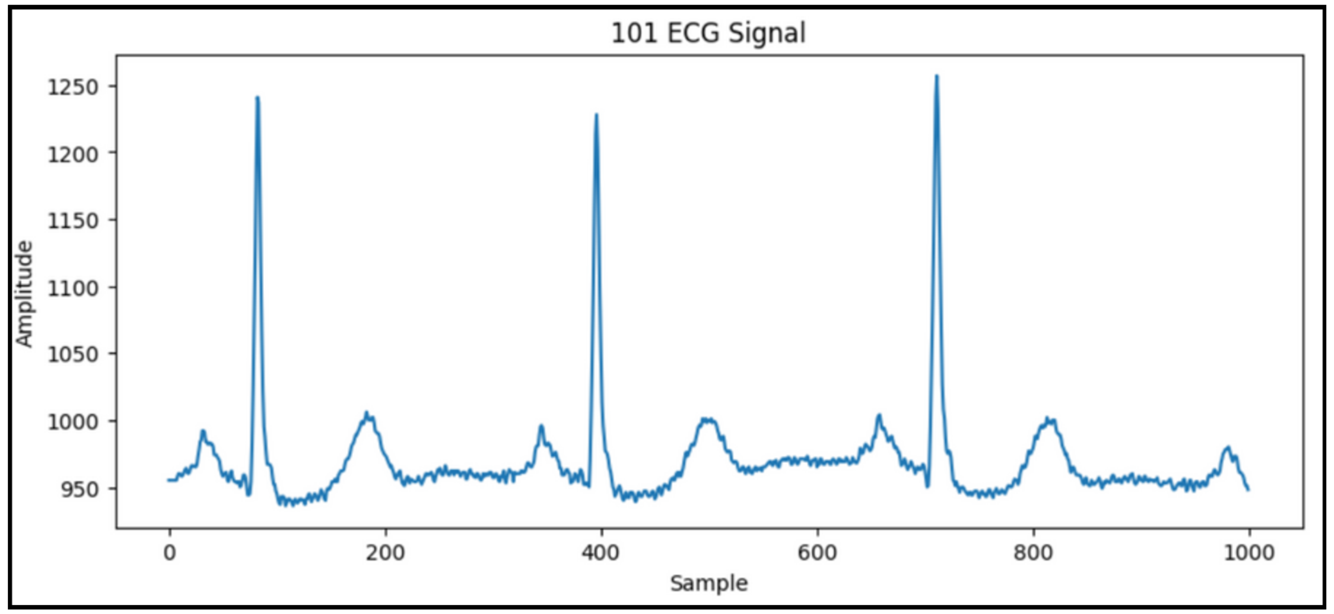
Convert data into heart rate and label data with criteria:

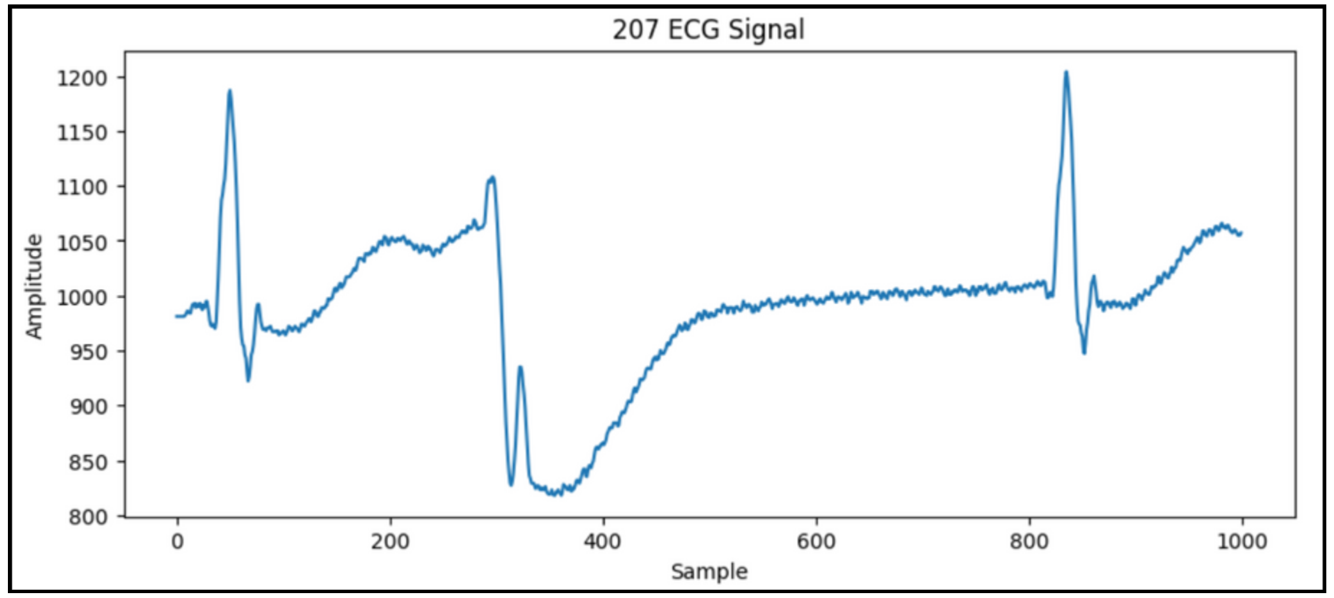
- Low: under 60;
- Normal: from 60 to 80
- Fast: above 80

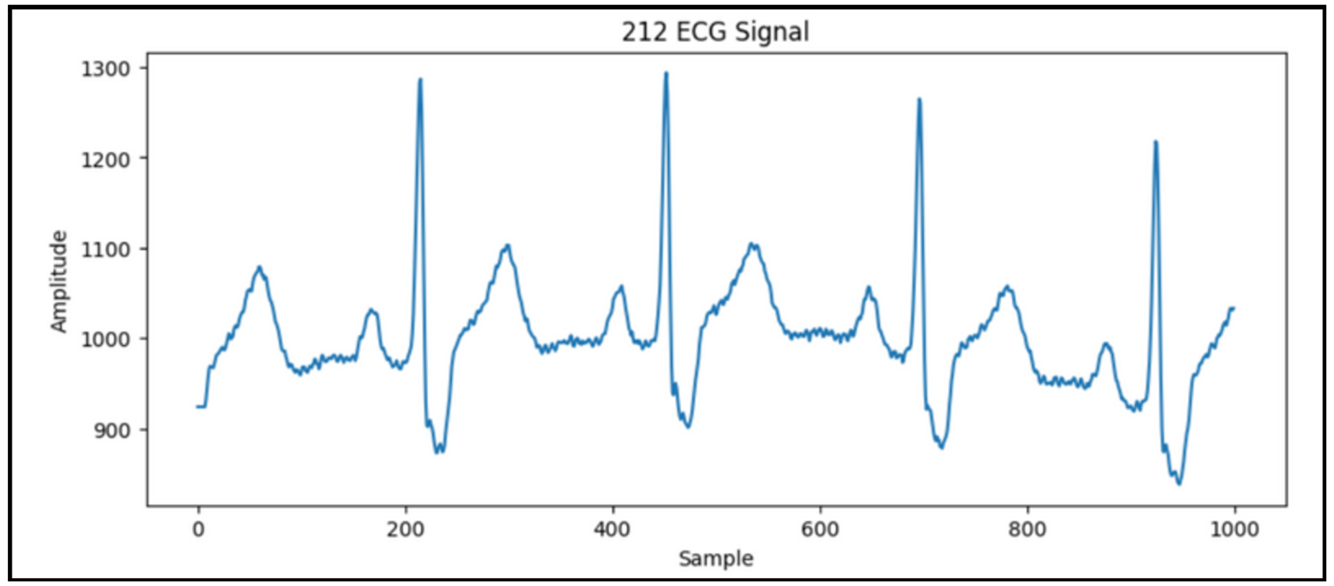
Load and Preprocess Data

```python
def load_and_preprocess_data(directory):
    # Get lists of CSV and TXT files
    csv_files = glob.glob(os.path.join(directory, "*.csv"))
    txt_files = glob.glob(os.path.join(directory, "*annotations.txt"))

    # Create dictionaries to map filenames without extensions
    csv_dict = {os.path.basename(f).replace('.csv', ''): f for f in csv_files}
    txt_dict = {os.path.basename(f).replace('annotations.txt', ''): f for f in txt_files}

    # Initialize lists for ECG signals and annotations
    ecg_signals = []
    annotations = []
    sampling_rate = 125  # Default sampling rate, adjust as needed

    # Iterate over CSV files
    for csv_filename in csv_files:
        # Extract the base name of the file (without extension)
        base_name = os.path.basename(csv_filename).replace('.csv', '')
        txt_filename = txt_dict.get(base_name)  # Find the corresponding TXT file

        if not txt_filename:
            print(f"No matching annotation file for {csv_filename}")
            continue  # Skip if no matching TXT file is found

        # Load the CSV data
        df = pd.read_csv(csv_filename)
        df.columns = [col.replace("'", "") for col in df.columns]

        if 'MLII' in df.columns:
            mlii_data = df['MLII'].values
```

```python
            # Plot the original ECG signal
            plt.figure(figsize=(10, 4))
            plt.plot(limited_mlii_data)
            plt.title(f"{base_name} ECG Signal")
            plt.xlabel("Sample")
            plt.ylabel("Amplitude")
            plt.show()
```







HEART RATE CONVERSION

Labeled and Heart Converted Data

```
Data after dropping NaN values:
Minimum heart rate value: 51.19690264124928
Maximum heart rate value: 340.3231192285024
    heart_rate   label
0   140.942990    fast
1   114.902427    fast
2    91.010141    fast
3   130.767526    fast
4   116.926195    fast
5    52.275826     low
6   235.019723    fast
7   284.809980    fast
8    71.989136  normal
9   102.570238    fast
10  180.265100    fast
11  211.179011    fast
12   97.104819    fast
13   97.665025    fast
14  214.244172    fast
15  171.446907    fast
16   60.935170  normal
17  104.592159    fast
18  125.308963    fast
19  125.317286    fast
20   97.828496    fast
21  101.409524    fast
22  154.331296    fast
23   96.551262    fast
24  340.323119    fast
25  126.263675    fast
```

```
16   60.935170  normal
17  104.592159    fast
18  125.308963    fast
19  125.317286    fast
20   97.828496    fast
21  101.409524    fast
22  154.331296    fast
23   96.551262    fast
24  340.323119    fast
25  126.263675    fast
26  103.468969    fast
27  236.523966    fast
28  233.922043    fast
29   51.196903     low
30  120.434469    fast
31  208.003170    fast
32  154.038740    fast
33  174.543758    fast
34   74.608599  normal
35   91.411208    fast
36  335.755787    fast
37   63.946463  normal
38  222.491231    fast
39  164.493224    fast
40  115.778506    fast
41  257.152827    fast
42  107.936757    fast
43  107.171047    fast
44  127.734110    fast
45  232.747872    fast
```

47 patients

=> 45 patients got MLII colunms

=> convert into 45 heart rate values

CLASSIFICATION

Dataset: training and testing set. The ratio is 80:20.

Support Vector Machine

- **Application to Heart Rate Classification:**
  - SVM effectively categorizes heart rate data into low, normal, and rapid categories by learning from preprocessed heart rate values. (Cortes & Vapnik, 1995)

```
# Split data into training and testing sets
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y_encoded, test_size=0.2, random_state=42)

# Train and evaluate the SVM model
svm_model = train_and_evaluate_svm(X_train, y_train)
```

Random Forest: trained on bootstrap samples and random feature subsets (Breiman, 2001)

- **Heart Rate Classification:**
  - The model uses heart rate values to classify heartbeats into categories such as low, normal, and fast, with a hierarchical structure that improves classification of varying heart rates

```
def train_and_evaluate_random_forest(X, y):
    # Split the data into training and testing sets
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
train_and_evaluate_random_forest(X, y)
```

CLASSIFICATION

Deep Neural Networks (DNNs)

- **Training and Testing:**
  - The model was trained on precisely calibrated heart rate data from Database B and tested on Database A, demonstrating its ability to accurately classify new ECG data and its potential for practical medical diagnostics.

PROCESS

IMPORT DATASET B	**CLASSIFY ECG Signal** LOW, FAST, NORMAL	**GOLD MODEL** TRAIN DATA	**DATASET A** TEST DATA	**CLASSIFY ACCURACY**

RESULT

- **Accuracy:**
  - Reflect the correctness of the model's predict
- **Precision:**
  - In heartbeat classification, precision indicates the model’s ability to avoid incorrectly classifying normal rhythms as abnormal, thus minimizing false positives.
- **Recall**:
  - **Emphasis**: Improve diagnostic reliability by ensuring comprehensive identification of abnormal heart rhythms.
  - **Importance**: High recall is crucial for identifying abnormal rhythms (e.g., tachycardia and bradycardia) and ensures that all potential abnormal cases are detected.

Classifier	Accuracy (%)	Precision (%)	Recall (%)
SVMs	89.7%	86%	90%
Random Forest	86%	77%	86%
DNNs	90%	81%	90%

Table 1: Performance metrics of the classifiers

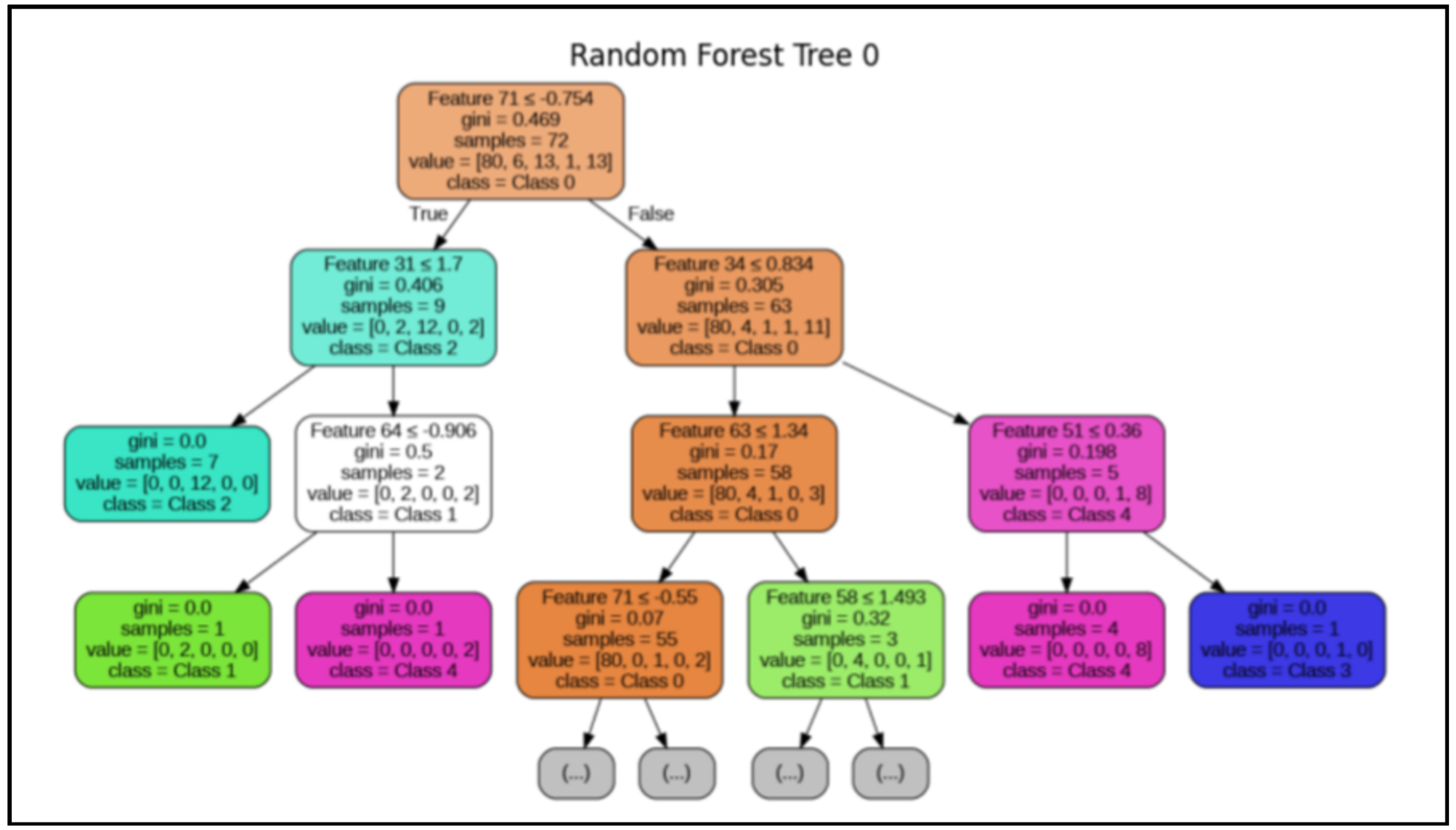
Evaluation Metrics

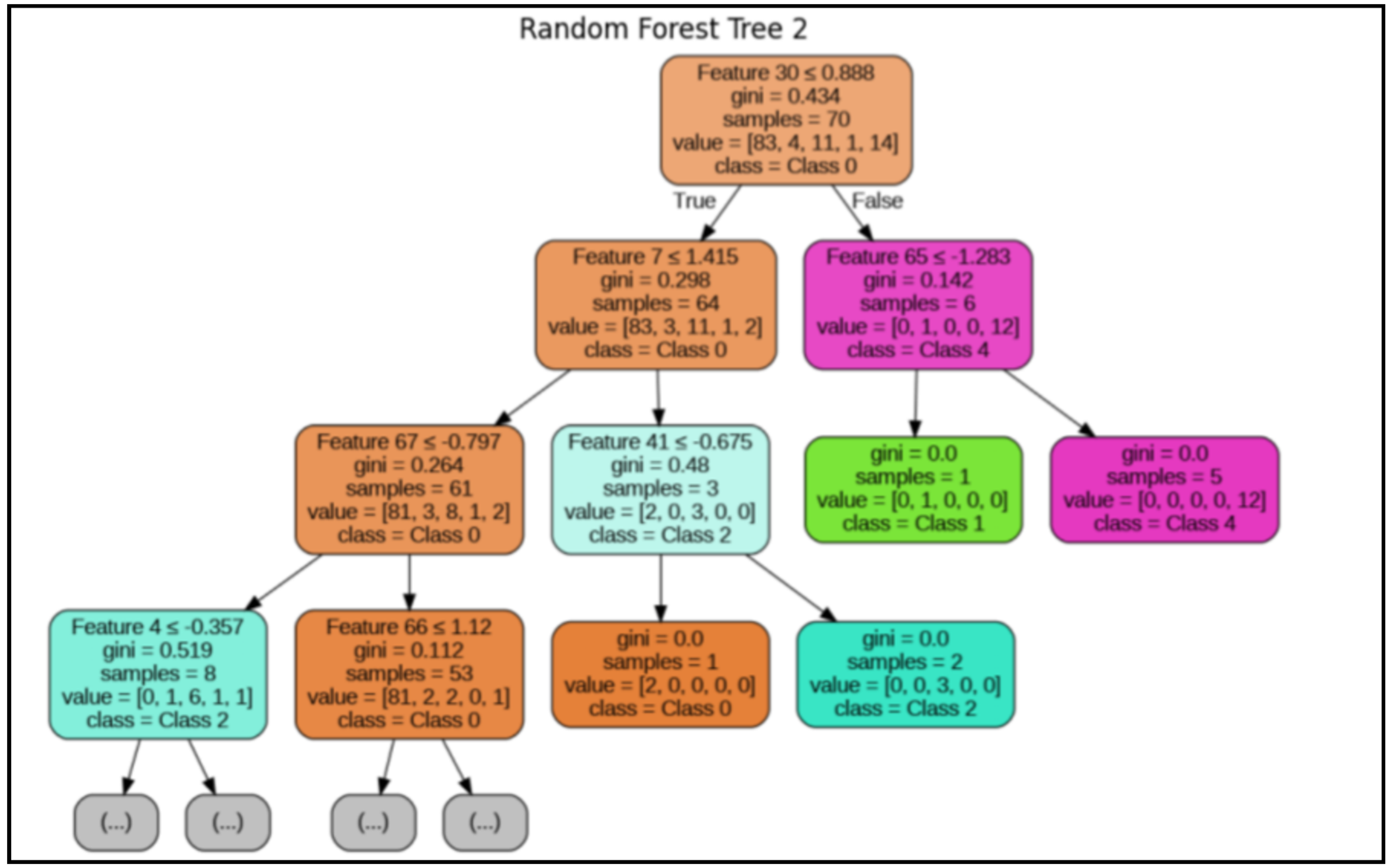
Support Vector Machine

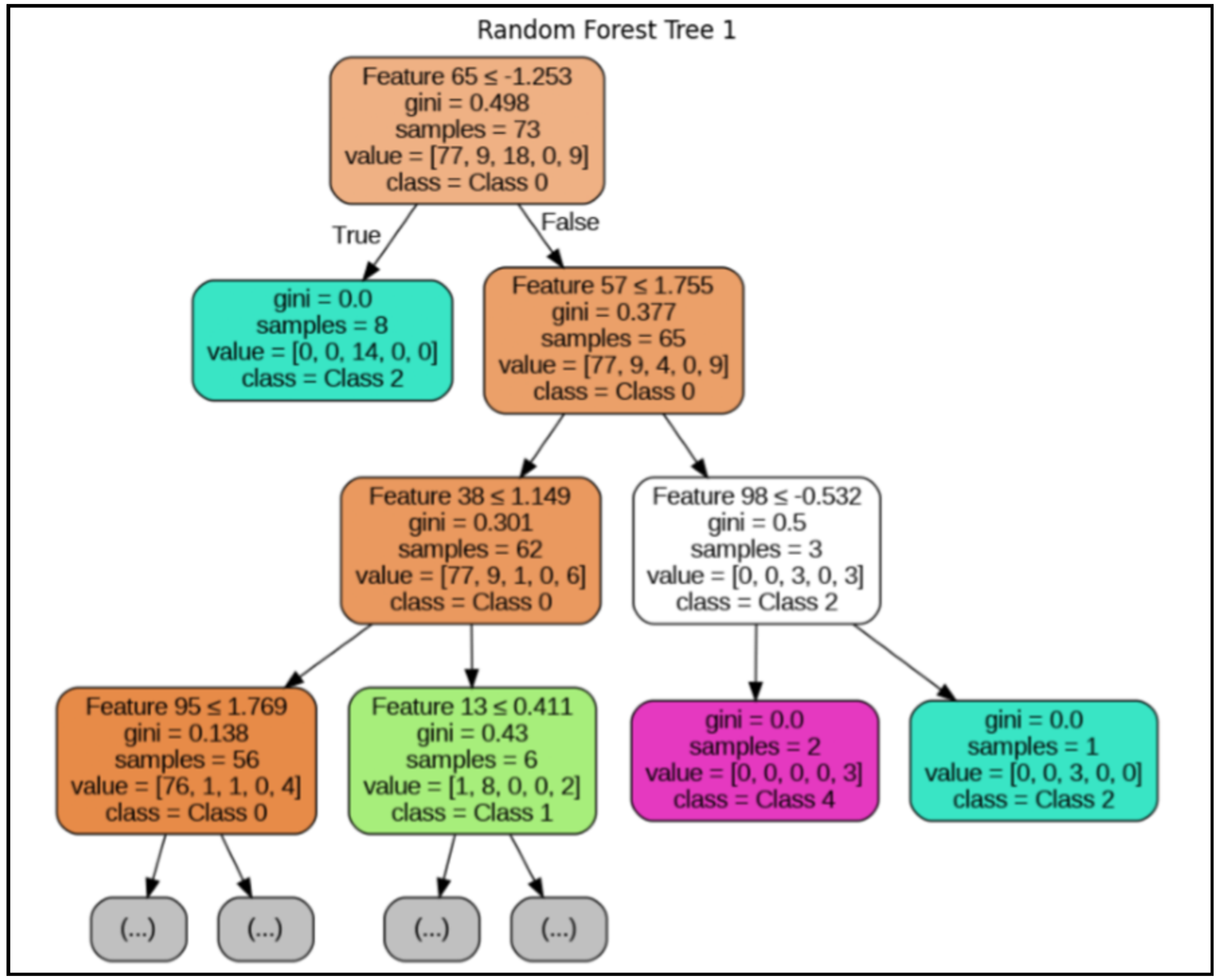
```
SVM Accuracy: 0.896551724137931
SVM Classification Report:
              precision    recall  f1-score   support

    accuracy                           0.90        29
   macro avg       0.49      0.65      0.54        29
weighted avg       0.86      0.90      0.87        29
```

- **Overall Accuracy:** High at approximately 89.65%.
- **Macro Averages:**
  - Precision: Moderate at 0.49, indicating average accuracy across all classes.
  - Recall: Fairly good at 0.65, suggesting the model reasonably identifies actual positives.
  - F1-Score: 0.54, reflecting moderate effectiveness in balancing precision and recall across classes.
- **Weighted Averages:**
  - Precision: High at 0.86, showing better prediction accuracy when considering class size.
  - Recall: Very good at 0.90, indicating effective identification of true positives in larger classes.
  - F1-Score: Strong at 0.87, demonstrating a good balance between precision and recall when adjusted for class distribution.

Random Forest Tree







Class Distribution and Imbalance:

- The model predominantly classifies instances into Class 0 (normal heartbeats), reflecting the higher occurrence of this class in the dataset.
- Minority classes (e.g., Class 1, 2, and 4) are less represented and are classified based on fewer samples.

Evaluation Metrics

Random Forest

```
Random Forest Accuracy: 0.8620689655172413
Random Forest Classification Report:
              precision    recall  f1-score   support

    accuracy                           0.86        29
   macro avg       0.22      0.25      0.24        29
weighted avg       0.77      0.86      0.81        29
```

- **Overall Accuracy:** Achieved 86.21%.
- **Macro Averages:**
  - Precision: Very low at 0.22, indicating poor accuracy in predicting true positives across classes.
  - Recall: Also low at 0.25, showing that the model identifies only 25% of all actual positives correctly.
  - F1-Score: 0.24, reflecting a weak balance between precision and recall.
- **Weighted Averages:**
  - Precision: Higher at 0.77, suggesting better performance weighted by class size.
  - Recall: 0.86, indicating more effective identification of true positives in larger classes.
  - F1-Score: 0.81, showing a stronger balance between precision and recall when accounting for class distribution.

Evaluation Metrics

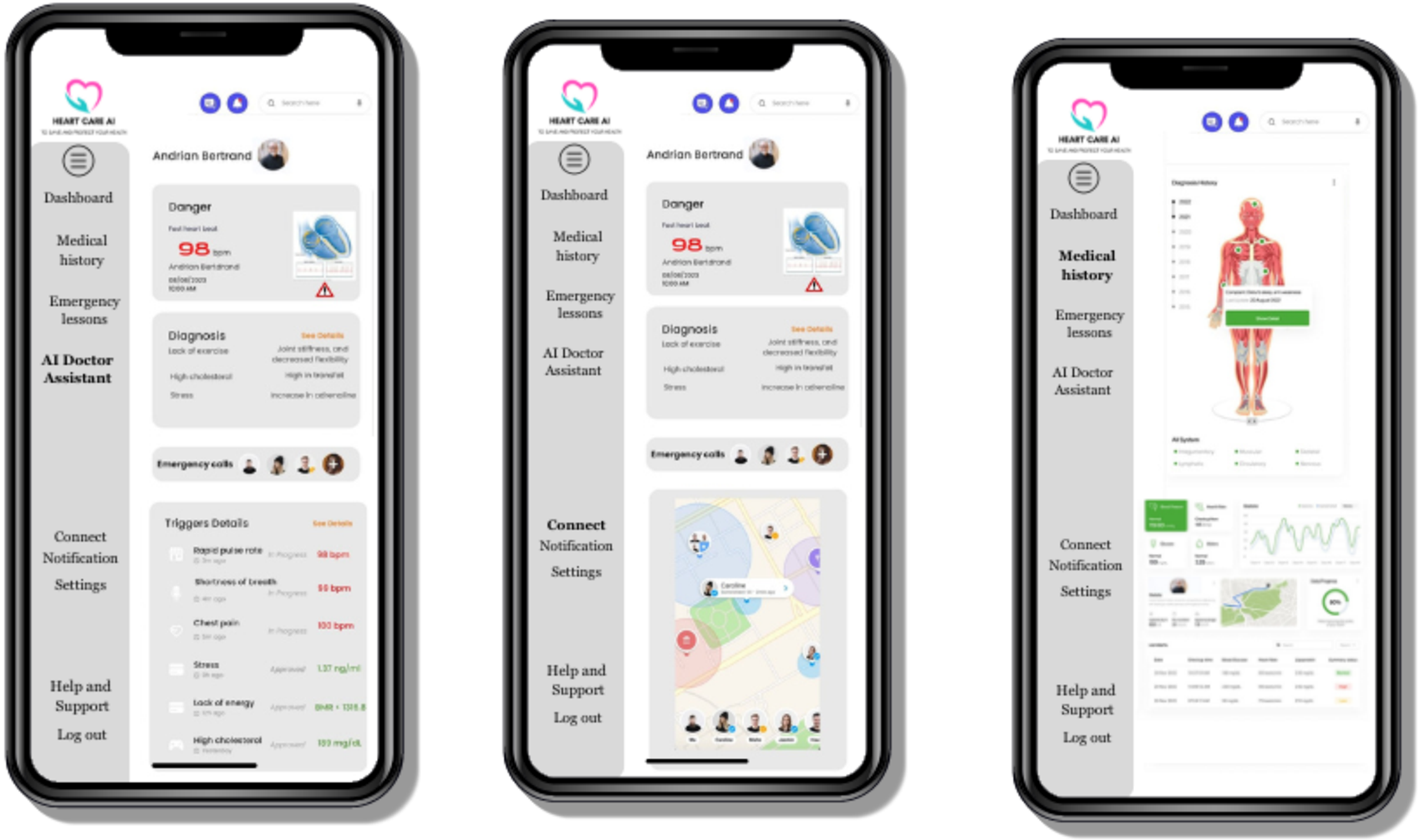
Deep Neural Networks

```
Accuracy: 0.9
Classification Report:
              precision    recall  f1-score   support

    accuracy                           0.90        10
   macro avg       0.45      0.50      0.47        10
weighted avg       0.81      0.90      0.85        10
```

- **Accuracy:** High at 90%, indicating effective class label prediction for most cases.
- **Macro Averages:**
  - Precision: Low at 0.45, suggesting limited accuracy across all classes.
  - Recall: Moderate at 0.50, reflecting average identification of true positives.
  - F1-Score: Also moderate at 0.47, pointing to performance issues with minority classes.
- **Weighted Averages:**
  - Precision: Improved at 0.81, better accounting for class frequency.
  - Recall: High at 0.90, effectively identifying true positives.
  - F1-Score: Strong at 0.85, indicating a good balance between precision and recall.

HEARTCARE AI PLATFORM

The proposal for launching of Heart Care AI, an AI heartbeat detecting platform based on Deep Neural Networks for timely intervention and improve arrhythmia detection accuracy and reliability



The proposed model could be integrated into a wearable device for real-time cardiac surveillance, requiring rigorous clinical trials to validate its efficacy and safety in real-world settings.

LIMITATIONS

Limited Dataset Diversity:

- The datasets used (MIT-BIH Arrhythmia Database and Shayan Fazeli’s Heartbeat dataset) **lack diversity,** as they mainly consist of data from specific demographics and controlled conditions.

==> Affect **the model's generalizability** to broader populations and real world .

Class Imbalance

- The datasets have an imbalance in class distribution, particularly with rare arrhythmias.
- Even though techniques like SMOTE were employed, the model's performance on minority classes may still be suboptimal.

FUTURE WORKS

Diversify Datasets

Include a broader range of ECG recordings from various populations and environments to enhance model generalizability and reliability across different demographics.

Advanced-Data Augmentation:

Implement advanced data augmentation techniques to address class imbalance and improve the detection of rare arrhythmias, thereby enhancing model performance.

Continued Development of Heart Care AI

Focus on iterative improvements and updates to the Heart Care AI platform based on ongoing research findings, user feedback, and real-world application data to continuously refine its diagnostic algorithms.

REFERENCES

- Acharya, U. R., Fujita, H., Lih, O. S., Adam, M., Tan, J. H., & Chua, C. K. (2017). Automated detection of arrhythmias using different intervals of tachycardia ecg segments with convolutional neural network. Information Sciences, 405, 81–90.
- American Heart Association. (2021). Arrhythmia [Accessed: 2024-08-09].
- Breiman, L. (2001). Random forests. Machine Learning, 45 (1), 5–32.
- Brown, T., & Roberts, J. (2020). Deep learning for ecg classification: A review. International Journal of Medical Informatics, 141, 104198. https://doi.org/10.1016/j.ijmedinf.2020.104198
- Chen, Y., & Zhang, H. (2021). Automated ecg analysis using machine learning techniques. Computers in Biology and Medicine, 135, 104554.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20 (3), 273–297.
- Fazeli, S. (2018). Heartbeat sounds [Accessed: 2024-08-09]. https : / / www . kaggle .com/datasets/shayanfazeli/heartbeat
- Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark,
- R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet: Components of a new research
- resource for complex physiologic signals [Accessed: 2024-08-09]. https://doi.org/10.1161/01.CIR.101.23.e215
- Gusev, M., & Boshkovska, M. (2019). Performance evaluation of atrial fibrillationdetection. 2019 42nd International Convention on Information and Commu-nication Technology, Electronics and Microelectronics (MIPRO), 342–347.https://doi.org/10.23919/MIPRO.2019.8757193
- Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and clas-sification in ambulatory electrocardiograms using a deep neural network.
- Nature Medicine, 25 (1), 65–69.Ince, T., Kiranyaz, S., & Gabbouj, M. (2009). A generic and robust system for automated patient-specific classification of ecg signals. IEEE Transactions on Biomedical Engineering.
- Islam, M. S., Khan, M. M., & Hossain, M. F. (2017). An efficient random forestapproach for ecg signal classification. Journal of Biomedical Research, 31 (2),123–130. https://doi.org/10.1016/j.jbiome.2017.01.123
- Liu, C., & Wang, H. (2021). A novel deep learning approach for arrhythmia classifica- tion using convolutional neural networks. IEEE Transactions on Biomedical
- Engineering, 68 (1), 349–360.
- Luz, E. J. S., Schwartz, W. R., C´amara-Ch´avez, G., & Menotti, D. (2016). Ecg-based heartbeat classification for arrhythmia detection: A survey. Computer Methods and Programs in Biomedicine, 127, 144–164.
- Machado, J. M., Peixoto, H., & Sousa, R. (Eds.). (2022). Big data analytics and artificial intelligence in the healthcare industry [Includes bibliographical ref- erences and index]. Medical Information Science Reference.
- Mohamad, R., Zhang, T., & Lee, K. (2018). Comparative analysis of ecg signal classification using machine learning techniques. IEEE Transactions on Biomedical Engineering, 65 (7), 1221–1230. https://doi.org/10.1109/TBME.2018.2808302

REFERENCES

- Moody, B., & Mark, R. (2010). The impact of using the mlii lead in ecg signal classification. Journal of Cardiovascular Technology, 27 (4), 123–130.
- Nguyen, H. H., & Sun, K. (2019). Transfer learning in ecg classification using deep neural networks. IEEE Access, 7, 122725–122735.
- Osowski, S., Hoai, L. T., & Markiewicz, T. (2004). Support vector machine-based ex-pert system for reliable heartbeat recognition. IEEE Transactions on Biomedical Engineering, 51 (4), 582–589. https:/ /doi .org /10 .1109/ TBME. 2004.824139
- Rahul, K. K. (2016). Mobile and e-health care; recent trends and future directions. Journal of Health & Medical Economics, 10, 1–10.
- Rajpurkar, P., Hannun, A. Y., Haghpanahi, M., Bourn, C., & Ng, A. Y. (2017). Cardiologist-level arrhythmia detection with convolutional neural networks. arXiv preprint arXiv:1707.01836.
- Rangayyan, R. M. (2015). Biomedical signal analysis. John Wiley & Sons. Rodŕ ıguez, J. J., Alonso, C. J., & de la Cruz, G. A. (2006). Using random forests and data augmentation to improve ecg signal classification. IEEE Transactions on Information Technology in Biomedicine, 10 (3), 532–540. https://doi.org/10.1109/TITB.2006.862140
- Smith, D., & Johnson, P. (2019). An overview of ecg interpretation and diagnosis. Journal of Cardiovascular Nursing, 34 (2), 121–127.
- Smith, R., Jones, L., & Green, P. (2019). Evaluation of machine learning models for ecg heartbeat classification. IEEE Transactions on Biomedical Engineering, 66 (8), 2475–2485. https://doi.org/10.1109/TBME.2019.2903011
- World Health Organization. (2024). Cardiovascular diseases (cvds) [Accessed: 2024-08-09].
- Yu, S., & Hu, S. (2015). Detection of qrs complexes and p waves in ecg signals using wavelet transform. Journal of Biomedical Research and Applications, 3 (2),123–130.
- Zheng, Y., Liu, Q., Chen, S., Ge, T., Li, J., Zhang, X., & Wang, D. (2014). A method for automatic identification of arrhythmia using a machine learning approach. IEEE Transactions on Biomedical Engineering, 61 (3), 1080–1086. https://doi.org/10.1109/TBME.2013.2290774 24

THANK YOU