Preliminary Documentation
EARIN Project – Predicting ECG Diagnosis
Summer 2025
Team Members:

- Maye Maye Ahmed Saleck
- Carlos Rueda

Instructor: Dr. Muhammad Farhan
Dataset: CPSC-2018 ECG Dataset

1. Dataset Description

We are using the CPSC-2018 ECG dataset, which contains multilead ECG recordings collected for classification purposes. Each recording may contain one or more diagnosis labels related to cardiac abnormalities.

Key Facts:

- Signals are sampled at 500Hz
- Variable-length recordings (6s to 144s)
- 12-lead ECG signals
- 9 diagnosis classes (e.g., AF, ST, PVC, etc.)

Initial Observations:

- Some recordings are noisy or have missing leads
- The dataset is imbalanced (e.g., many more Normal cases than rare pathologies)

Visualization:
We plotted several sample ECG signals using the WFDB and matplotlib libraries. A clear difference can be seen between normal and abnormal signals.

2. Problem Solving Plan

a. Data Splitting Strategy
We plan to split the dataset as follows:
70% training, 15% validation, 15% testing
We will use stratified sampling to preserve class distribution across subsets.

b. Algorithms to Be Used
We will compare the following approaches:

1. CNN on spectrogram images (Mel-Spectrogram + 2D CNN)
2. LSTM on raw signals (time series classification)
3. Random Forest on extracted features (peak rate, RR intervals, etc.)

c. Libraries and Tools
We will use the following tools:

- wfdb for reading ECG data
- pywavelets for denoising
- scipy.signal and librosa for spectral feature extraction
- scikit-learn, keras, and PyTorch for model training and evaluation

d. Evaluation Methods

We will evaluate each model using the following metrics:

Accuracy, F1 Score, ROC Curve, Sensitivity, Specificity, and Confusion Matrix

We aim to test at least 5–6 hyperparameter configurations per model.

3. Planned Experiments

- Spectrogram conversion versus raw signal input
- Comparison between deep learning and classical machine learning
- Hyperparameter tuning (learning rate, layers, window size, etc.)
- Impact of denoising using wavelets

4. Result Visualization

We plan to include:

- Sample spectrograms
- Metric comparison tables
- ROC curve plots
- Confusion matrices

5. Conclusion

The preliminary setup is ready. Our dataset is loaded and partially visualized. Our next step will be implementing the pipeline for signal preprocessing and classification using the outlined methods.

6. References

1. WFDB Library – https://www.physionet.org/content/wfdb/
2. PyWavelets – https://pywavelets.readthedocs.io
3. SciPy Spectrogram – https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.spectrogram.html
4. Librosa Mel-spectrogram – https://librosa.org/doc/main/generated/librosa.feature.melspectrogram.html
5. Scikit-learn – https://scikit-learn.org
6. Papers-with-code-https://paperswithcode.com/dataset/the-china-physiological-signal-challenge-2018
7. ICBEB(China)-http://2018.icbeb.org/Challenge.html