ECG Signal Classification Using Feature Engineering and ML Models

Rahul Gupta (202211069)

IIITV-ICD
Course: CS/IT 312 Data Analytics and Visualization
Instructor: Dr. Venkata Phanikrishna

21-04-2025

Dataset Description

- **Type of Data Considered:**
  - **Type:** 1D ECG Signal (Time-Series)
  - **Description:** The dataset consists of electrocardiogram (ECG) signals, where each sample is a sequence of 140 voltage measurements (time points) representing a single ECG waveform, along with a binary label (0 = normal, 1 = abnormal).
- **Why ECG Dataset:**
  - ECG signals are essential for the diagnosis of heart conditions.
  - Enables automated detection to aid medical professionals.

First 5 rows of the dataset:

	signal_0	signal_1	signal_2	signal_3	signal_4	signal_5	signal_6	signal_7	signal_8	signal_9	...	signal_131	signal_132	signal_133	signal_134	signal_135	signal_136	signal_137	signal_138	si
0	-0.112522	-2.827204	-3.773897	-4.349751	-4.376041	-3.474986	-2.181408	-1.818286	-1.250522	-0.477492	...	0.792168	0.933541	0.796958	0.578621	0.257740	0.228077	0.123431	0.925286	
1	-1.100878	-3.996840	-4.285843	-4.506579	-4.022377	-3.234368	-1.566126	-0.992258	-0.754680	0.042321	...	0.538356	0.656881	0.787490	0.724046	0.555784	0.476333	0.773820	1.119621	-
2	-0.567088	-2.593450	-3.874230	-4.584095	-4.187449	-3.151462	-1.742940	-1.490659	-1.183580	-0.394229	...	0.886073	0.531452	0.311377	-0.021919	-0.713683	-0.532197	0.321097	0.904227	-
3	0.490473	-1.914407	-3.616364	-4.318823	-4.268016	-3.881110	-2.993280	-1.671131	-1.333884	-0.965629	...	0.350816	0.499111	0.600345	0.842069	0.952074	0.990133	1.086798	1.403011	-
4	0.800232	-0.874252	-2.384761	-3.973292	-4.338224	-3.802422	-2.534510	-1.783423	-1.594450	-0.753199	...	1.148884	0.958434	1.059025	1.371682	1.277392	0.960304	0.971020	1.614392	

5 rows × 141 columns

Figure: Visualization of ECG Dataset

Dataset and Project Overview

- **Number of Observations / Subjects:**
  - **Total Number of Samples:** 4,998
  - **Categories (Samples per Class):**
    - Label 0 (Normal): 2,079 samples (41.6%)
    - Label 1 (Abnormal): 2,919 samples (58.4%)

- **Project Type:**
  - **Type:** Classification (Binary)
  - **Description:** The goal is to classify ECG signals as normal (0) or abnormal (1) based on extracted features, making this a supervised binary classification task.

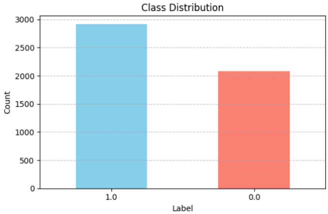

Figure: Class Labels Visualization

Data Source and Description

- **Data Source:**
  - **URL:** Kaggle ECG Dataset
- **Dataset Information:**
  - Contains ECG readings of patients.
  - Each row corresponds to a single complete ECG of a patient, composed of 140 data points (readings).
  - **Columns:**
    - Columns 0–140: ECG data points (floating-point numbers).
    - Label: Categorical variable indicating whether the ECG is normal (0) or abnormal (1).

Data Representation Before Feature Extraction

- Visualized as a heatmap to display signal patterns across all samples and time points.
- Total Graphs: 17 (1 heatmap + 16 Statistics Visualizations).

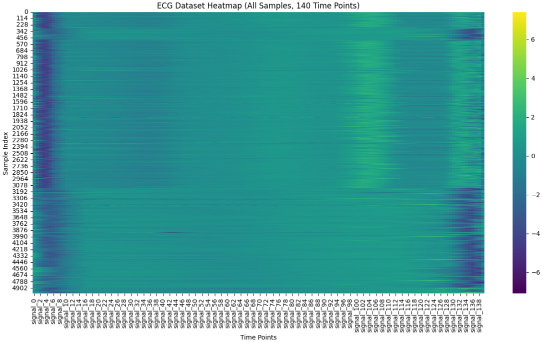

Figure: ECG Dataset Heatmap (All Samples, 140 Time Points)

Feature Extraction / Creation Details

- **Total Number of Features Extracted:** 15
- **Extracted Features:**
  - Mean, Std, Skewness, Kurtosis, Range
  - RMS, Zero-Crossing Rate (ZCR), Peak Count
  - PSD Mean, Dominant Frequency, PSD Total, FFT Max, Band Energy Ratio
  - Wavelet Energy, Wavelet Variance
- **Total Number of Features Created:** 9
- **Created Features:**
  - RR Mean, HRV (SDNN), RR Median
  - QRS Duration, QRS Amplitude
  - P-Wave Count, P-Wave Amplitude
  - T-Wave Count, T-Wave Amplitude

Data Representation After Feature Extraction

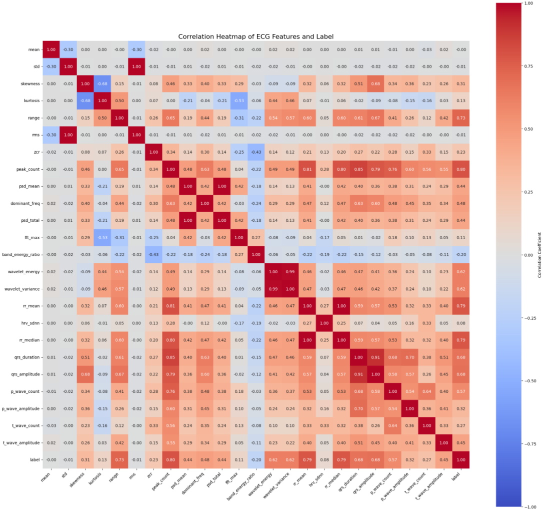

Figure: Correlation Heatmap of ECG Features and Label

Feature Selection Techniques Used

- **Filter Method:** Mutual Information (MI) for ranking features.
- **Correlation-Based Selection:**High MI and low correlation (< 0.9).

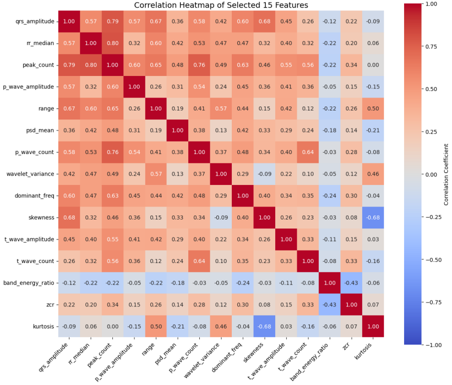

Figure: Correlation Heatmap of Selected 15 Features

Feature Transformation Techniques Used

- **Method:** Standardization (Z-score Normalization)
- **Description:**
  - Applied `StandardScaler` to the 15 selected features, transforming them to have zero mean and unit variance.
  - **Formula:** $z = \frac{x-\mu}{\sigma}$
    - x: Original feature value
    - μ: Mean of the feature
    - σ: Standard deviation of the feature
- **Purpose:**
  - Ensures all features are on the same scale, preventing features with larger ranges (e.g., wavelet_energy) from dominating distance-based algorithms (e.g., SVM).
  - Improves convergence and performance of models sensitive to feature scales (e.g., SVM, LDA).

Feature Reduction Techniques Used

- **Method:** Linear Discriminant Analysis (LDA)
- **Description:**
  - Reduces 15 standardized features to 1 component, maximizing class separability for binary classification.
- **Steps and Formulas:**
  - Compute within-class scatter matrix: $S_W = \sum_{c=0,1} \sum_{i \in c} (\mathbf{x}_i - \mathbf{m}_c)(\mathbf{x}_i - \mathbf{m}_c)^T$
  - Compute between-class scatter matrix: $S_B = (\mathbf{m}_0 - \mathbf{m}_1)(\mathbf{m}_0 - \mathbf{m}_1)^T$
  - Solve for projection vector **w** maximizing $\frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$.
  - Project 15D feature data onto 1D vector.
- **Terms:**
  - $\mathbf{w}$: Projection vector (1D direction for maximum class separation).
  - S_W: Within-class scatter matrix (variability within each class).
  - S_B: Between-class scatter matrix (variability between class means).
  - $\mathbf{x}_i$: Feature vector of sample i.
  - $\mathbf{m}_c$: Mean vector of class c (0 or 1).

Hypothesis Testing Methods Used

- **Method:** Independent Two-Sample t-test
- **Description:** T-tests compare feature distributions between normal (0) and abnormal (1) classes.
- **T-test Results for Selected 15 Features:**

Feature	t-statistic	p-value	Significant ($p < 0.05$)
qrs_amplitude	-65.95	0.00	True
rr_median	-89.64	0.00	True
peak_count	-93.87	0.00	True
p_wave_amplitude	-24.10	1.83e-121	True
range	-75.14	0.00	True
psd_mean	-34.67	3.11e-236	True
p_wave_count	-49.11	0.00	True
wavelet_variance	-56.14	0.00	True
dominant_freq	-38.25	6.00e-281	True
skewness	-23.29	6.03e-114	True
t_wave_amplitude	-35.56	4.12e-247	True
t_wave_count	-20.01	9.20e-86	True
band_energy_ratio	14.21	5.93e-45	True
zcr	-16.77	1.81e-61	True
kurtosis	-9.36	1.15e-20	True

- **Terms:**
  - **t-statistic:** Measures difference in means relative to variability.
  - **p-value:** Probability of results under null hypothesis.

Models Employed

- **Random Forest Classifier:**
  - Ensemble method using decision trees, robust to noise and non-linear relationships.
  - Utilizes bootstrap sampling to create diverse subsets of data for each tree.
  - Handles overfitting through averaging predictions across multiple trees.
- **Support Vector Machine (SVM):**
  - Linear kernel method, effective for linearly separable data with a single LDA component.
  - Uses soft margin to allow some misclassifications for better generalization.
- **Model Metrics:**

Metric	Random Forest	SVM
Training Accuracy	0.9997	0.9687
Test Set Accuracy	0.9510	0.9700
Prediction (First Row)	1.0 (True: 1.0)	1.0 (True: 1.0)

Best Model Selection Criteria (Beyond Accuracy)

- **Metrics Considered:**
  - F1-Score: Balances precision and recall, key for imbalanced data.
  - Recall: Detects abnormal ECGs (label 1), minimizing false negatives.
- **5-Fold Cross-Validation Results:**

Metric	Random Forest	SVM
Precision	0.9533	0.9669
Recall	0.9510	0.9801
F1-Score	0.9522	0.9735
AUC-ROC	0.9428	0.9665
False Negatives	143	58

- **Cross-Validation Accuracy:**
  - Random Forest: Mean = 0.9466, Std = 0.0091
  - SVM: Mean = 0.9688, Std = 0.0073
- **Paired T-Test Results:**
  - T-statistic: -12.6891
  - P-value: 0.0000
  - Reject the null hypothesis: Significant difference in performance between Random Forest and SVM ($p = 0.0000$).

Workflow Diagram

- **Description:** Represents the end-to-end process from ECG data input to model prediction.

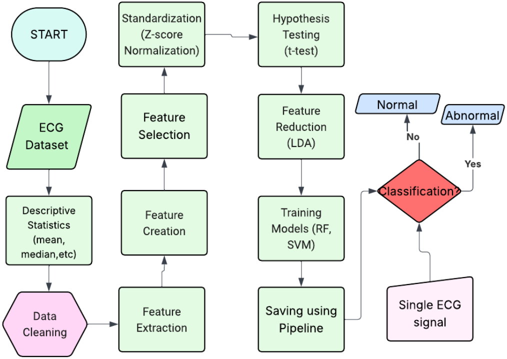

Figure: Workflow Diagram of ECG Classification Process