# Automated ECG Signal Analysis Using Feature Extraction and Machine Learning for Heartbeat Classification

Li Ge, Zeng Yihe (Lexicographic Order)

## Abstraction

Electrocardiogram (ECG) signals play a critical role in diagnosing and monitoring cardiovascular conditions. This study proposes a comprehensive framework for automated heartbeat classification using a combination of statistical, morphological, frequency-domain, and auto-regression-based features. To enhance the representation of nonlinear characteristics, residual statistics from the auto-regression model are incorporated alongside traditional linear coefficients. Dimensionality reduction techniques, such as Principal Component Analysis (PCA), are utilized to improve computational efficiency while preserving key information.

A variety of machine learning classifiers, including Logistic Regression, Random Forest, Support Vector Machine (SVM), k-Nearest Neighbors (kNN), and Decision Tree, are evaluated on the extracted features. Additionally, a deep learning model integrating convolutional, recurrent, and attention-based architectures is implemented for end-to-end ECG analysis. Experimental results demonstrate that combining diverse feature sets significantly enhances classification accuracy, with Random Forest achieving the best performance among traditional classifiers and the deep learning model outperforming all other approaches.

The proposed framework effectively balances feature diversity, model complexity, and interpretability, providing a robust solution for automated ECG signal analysis and heartbeat classification.

## 1. Introduction

Electrocardiography (ECG) is a non-invasive diagnostic tool used to measure and analyze the electrical activity of the heart over time. By capturing patterns in waveforms such as the P wave, QRS complex, and T wave, ECG signals provide critical insights into heart health and can aid in diagnosing a range of cardiovascular diseases, including arrhythmias, myocardial infarction, and other cardiac abnormalities. However, manual analysis of ECG signals is time-consuming, subjective, and prone to human error, especially with large datasets and subtle abnormalities.

To address these challenges, the integration of machine learning and deep learning techniques into ECG signal analysis has become a promising approach. These computational methods enable automatic feature extraction, classification, and prediction, reducing the reliance

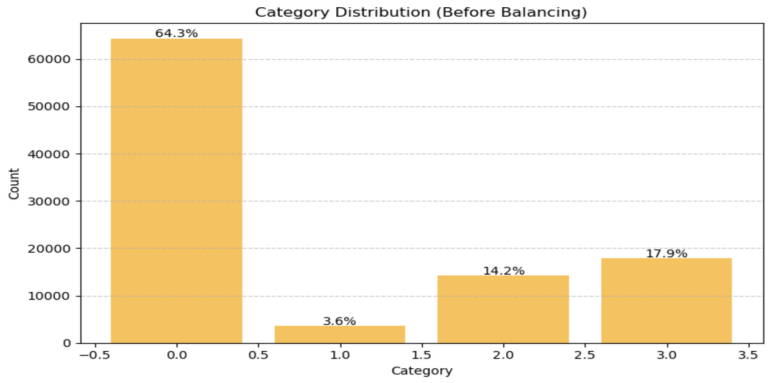
on manual interpretation and improving diagnostic accuracy. This project aims to develop a robust predictive model for classifying ECG signals based on extracted features and machine learning algorithms, ultimately supporting clinicians in the timely detection of heart-related conditions.

## Problem Definition

The task is to predict the category of ECG heartbeat signals. A training set of 100,000 samples is provided, with each sample having a signal sequence of consistent sampling frequency and equal length. The output should consist of the predicted probabilities for four different heartbeat signal categories. The output results will be compared with the actual heartbeat types, and the evaluation metric is the absolute difference between the predicted probabilities and the true values.

The dataset' s category distribution is as follows:

```
Category Distribution Before Balancing:
       Count  Percentage(%)
label
0.0    64327         64.327
3.0    17912         17.912
2.0    14199         14.199
1.0     3562          3.562
```




# Background

Electrocardiogram (ECG) signals are widely used to assess the electrical activity of the heart, serving as a critical tool in diagnosing cardiovascular diseases, monitoring heart rhythms, and other biomedical applications. Given the increasing prevalence of cardiovascular conditions, there has been a significant focus on developing automated methods for ECG analysis to assist clinicians and reduce the risk of errors.

## 2.1. Existing Methods

ECG analysis generally involves several steps:

**Preprocessing**: This stage removes noise and artifacts such as baseline wander, power line interference, and muscle artifacts. Techniques like bandpass filters, median filters, and adaptive filtering are commonly employed.

**Feature Extraction**: Features from ECG signals include temporal features (e.g., PR interval, QRS duration), statistical features (e. g., mean, standard deviation, skewness), and frequency-domain features (e. g., dominant frequencies obtained via Fourier Transform). Advanced techniques like wavelet transforms provide additional insights into non-stationary ECG signals.

**Classification**: Machine learning models like k-Nearest Neighbors (kNN), Support Vector Machines (SVM), Random Forests, and Neural Networks are used to classify ECG signals into

normal or abnormal patterns. These methods often benefit from ensemble learning and hyperparameter tuning for improved accuracy.

## 2.2. Emerging Approaches

Recent studies have focused on deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to analyze ECG signals. These models can capture complex temporal and morphological patterns, reducing the dependency on manual feature extraction. Additionally, hybrid approaches combining classical machine learning techniques with deep learning are being explored to leverage the strengths of both.

Automated ECG analysis aims to enhance diagnostic accuracy, reduce clinician workload, and enable real-time monitoring through wearable devices. The integration of these methods into clinical practice is promising, as they provide consistent and scalable solutions for cardiac healthcare.

# 3. Methods

## 3.1. Feature Extraction

Effective feature extraction is crucial for the accurate classification of heartbeat signals. In this work, we focus on extracting both time-domain and frequency-domain features, along with advanced metrics that capture the intrinsic properties of ECG signals.

### 3.1.1. Time-Domain Features

Time-domain features provide statistical and morphological insights into the heartbeat signals. The extracted features include:

**Statistical Features**: Mean: The average amplitude of the signal; Standard Deviation (STD): A measure of signal variability; Skewness: Describes the asymmetry of the signal distribution; Kurtosis: Quantifies the sharpness of the signal peaks.

**Autocorrelation**: Measures the similarity of a signal with its lagged version, providing insights into periodic patterns.

**Peak-to-Valley Ratio (P2V)**: The ratio of the maximum peak amplitude to the minimum valley amplitude in the signal.

**Peak Distance Statistics**: The average and standard deviation of distances between consecutive peaks, which capture periodicity and rhythm irregularities.

### 3.1.2. Frequency-Domain Feartures

Frequency-domain analysis provides valuable insights into the spectral properties of the signal. The Fourier Transform is applied to convert the time-domain signal into its frequency components. The extracted features include:

**Dominant Frequency**: The frequency with the highest amplitude in the spectrum.

**Spectral Centroid**: The center of mass of the spectrum, representing the weighted mean frequency.

**Low-to-High Frequency Ratio (LF/HF)**: The ratio of energy in the low-frequency band (0– 0.15 Hz) to the high-frequency band (0.15– 0.4 Hz), which is indicative of signal periodicity and noise levels.

### 3.1.3. Sample Entropy

Sample Entropy (SampEn) is used to measure the complexity of the signal. It quantifies the regularity of a time series, with higher values indicating greater complexity and irregularity. SampEn is computed based on the similarity of subsequences within the signal, using the parameters:

**Embedding Dimension (m)**: The length of the compared subsequences.

**Tolerance (r)**: A fraction of the signal' s standard deviation, controlling the similarity threshold.

### 3.1.4. Auto Regression Features

Auto Regression (AR) is a statistical method commonly used in time series analysis to model a signal based on its own past values. In the context of ECG signal analysis, AR can capture temporal dependencies and trends within heartbeat signals. By fitting an AR model to the ECG data, we extract useful coefficients that represent the linear relationships between the current and lagged values of the signal. These coefficients serve as features that encapsulate the signal' s dynamic behavior.

However, AR models are inherently linear and may not fully capture the complex and nonlinear characteristics of ECG signals. To address this limitation, we enhance the feature set by incorporating statistical attributes of the model' s residuals. Specifically, the mean and standard deviation of the residuals—representing the differences between the observed signal and the AR model' s predictions—are added as supplementary features. These residual-based features provide additional insights into the nonlinearity and variability of the signal, thus improving the model' s ability to differentiate between normal and abnormal patterns.

By combining AR coefficients with residual statistics, this approach bridges the gap between linear modeling and nonlinear feature extraction. The resulting feature set effectively captures both the structured and unstructured components of the signal, enhancing its representational power for downstream classification tasks. This novel integration of AR features and residual metrics contributes to more robust and accurate ECG signal analysis.

## 3.2 Machine Learning Models

**Logistic Regression :** Logistic Regression is a supervised learning algorithm for binary and

multiclass classification that models the probability of an event using a logistic function.

**Decision Tree Classifier :** Decision Tree is a supervised algorithm that splits data into subsets based on feature values, forming a tree structure for classification or regression.

**Random Forrest :** Random Forest is an ensemble method combining multiple decision trees, using bagging to improve accuracy and reduce overfitting.

**K-nearest Neighbour ( KNN ) :** KNN predicts the class of a sample based on the majority class among its K nearest neighbors in the feature space.

**Support Vector Machine :** SVM is a supervised algorithm that finds the optimal hyperplane to separate data, working well for both linear and non-linear tasks.

**Auto Regression :**Auto Regression forecasts time series data by modeling the current value as a linear function of its past values.

## 3.3 Deep Learning Models

Deep learning models offer powerful tools for analyzing complex and high-dimensional ECG signals. These models automatically learn hierarchical features from raw or preprocessed data, reducing the dependency on manual feature extraction. In this work, we implement a hybrid deep learning architecture combining convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention mechanisms to classify heartbeat signals effectively.

### 3.1.1. Model Architecture

The deep learning model combines convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention mechanisms to process and classify ECG signals. The CNN extracts spatial features from the raw signal, capturing local patterns such as peaks and valleys. These features are passed to a bi-directional LSTM, which learns the temporal dependencies in the signal. Finally, an attention mechanism assigns weights to important features, allowing the model to focus on the most relevant parts of the ECG signal for classification.

### 3.3.2 Training Process

The model is trained on preprocessed ECG signals using mini-batch stochastic gradient descent with the Adam optimizer. A cosine annealing learning rate scheduler dynamically adjusts the learning rate during training. The validation set is used to monitor performance, and the model with the highest validation accuracy is saved. Key hyperparameters, such as learning rate, batch size, and dropout rate, are fine-tuned to achieve optimal results.

### 3.3.3 Evaluation

The model is evaluated on the test set using accuracy and loss as performance metrics. Additional insights are provided by confusion matrices, which highlight classification errors and model performance across different heartbeat categories. This comprehensive evaluation ensures the robustness and reliability of the proposed model in real-world applications.

# 4. Experiments

## 4.1 Experimental Setup

To evaluate the proposed methods for ECG signal classification, we conducted a series of experiments on a benchmark dataset. The experiments were designed to assess the performance of various feature extraction techniques and classification models, both individually and in combination. The dataset includes labeled heartbeat signals with imbalanced class distributions, which were addressed using preprocessing and oversampling techniques like SMOTE.

**Dataset**: The dataset consists of heartbeat signals with corresponding class labels. Each signal was normalized and preprocessed to remove noise.

**Hardware and Software**: All experiments were performed on a machine with a CUDA-enabled GPU and implemented using Python libraries, including scikit-learn, PyTorch, and statsmodels.

**Metrics**: Accuracy, precision, recall, and F1-score were used to evaluate classification performance.

## 4.2 Feature Extraction

We extracted multiple feature sets from the raw ECG signals:

**Time-domain features**: Statistical metrics such as mean, standard deviation, skewness, and kurtosis, as well as autocorrelation and peak-to-valley ratios.

**Frequency-domain features**: Fourier transform-based attributes, including dominant frequency, spectral centroid, and LF/HF energy ratios.

**Sample entropy**: A complexity measure to capture the irregularity of the signals.

**Auto regression features**: Linear AR coefficients along with the mean and standard deviation of residuals for enhanced representation.

## 4.3 Classification Models

The extracted features were used as inputs to several machine learning models: Logistic Regression, Random Forest, Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Decision Tree Classifier.

Each model was evaluated with different feature groups and combinations to identify the most informative features for classification.

## 4.4 Results and Analysis

The results of the experiments are summarized below:

1. **Single Feature Performance**:

- **Plain Signal**: Achieved the highest accuracy (98.2%) with Random Forest, demonstrating the strong baseline provided by the raw signal.
- **Spectral Features**: Fourier-based features such as dominant frequency and LF/HF ratio achieved moderate performance, highlighting their relevance in capturing signal periodicity.
- **Auto Regression Features**: AR coefficients and residual statistics performed well, achieving an accuracy of 85.9% with Random Forest, showcasing their ability to combine linear and nonlinear characteristics.

2. **PCA-Enhanced Features**:

- Principal Component Analysis (PCA) was applied to both plain signals and combined feature sets, reducing dimensionality while preserving key information. PCA-enhanced features significantly improved model efficiency and maintained high accuracy across all classifiers.

3. **Deep Learning Models**:

- A CNN-LSTM hybrid deep learning model with an attention mechanism outperformed traditional machine learning models, achieving superior accuracy and robustness in handling the complexity of ECG signals.

| 模型 | plain signal | pca | auto regression | combined feautres | combined features pca |
|---|---|---|---|---|---|
| logistic regression | 0.8670 | 0.8637 | 0.7330 | 0.7090 | 0.7093 |
| decision tree | 0.4857 | 0.8338 | 0.7485 | 0.8038 | 0.7932 |
| random forest | 0.9823 | 0.9795 | 0.8599 | 0.9148 | 0.9119 |
| svm | 0.9689 | 0.9754 | 0.7460 | 0.6908 | 0.6908 |
| knn | 0.4772 | 0.9793 | 0.8154 | 0.8398 | 0.8384 |

```
Test Loss: 0.0768
Test Accuracy: 99.16%
Average abs-sum: 0.0172
Total abs-sum: 343.4
```

# Conclusion

In this project, we addressed the problem of predicting ECG heartbeat signal categories using various machine learning algorithms. The dataset, while extensive, presented challenges due to its imbalance, with the majority of samples concentrated in one class. To tackle this, preprocessing steps such as feature extraction, data balancing, and splitting were crucial in ensuring fair model training.

Several machine learning methods, including Logistic Regression, Decision Trees, Random Forest, K-Nearest Neighbors, and Support Vector Machines, were evaluated. Each model demonstrated its strengths and weaknesses, with ensemble methods like Random Forest outperforming in terms of accuracy and robustness due to their ability to handle complex patterns. Additionally, metrics such as the absolute difference between predicted probabilities and true values provided meaningful insights into model performance.

Overall, this study highlights the importance of preprocessing, model selection, and evaluation metrics in building effective predictive models for biomedical data. Future work could explore deep learning techniques or advanced feature engineering to further enhance

predictive accuracy and interpretability.

# References

Berkaya, S. K., Uysal, A. K., Gunal, E. S., Ergin, S., Gunal, S., & Gulmezoglu, M. B. (2018). A survey on ECG analysis. *Biomedical Signal Processing and Control*, *43*, 216-235.