ECG CLASSIFICATION USING LINEAR REGRESSION

PROJECT REPORT

Submitted by

Meera S
Mithul Akshay
Meesa Rakesh
Mohan Innani

Abstract—An electrocardiographic (ECG) signal is a dependable technique for the evaluation of activity of the cardiovascular system. A recent surge of interest in the classification of heartbeats has been witnessed. While there are many commonalities among different conditions of the ECG, most work has focused on classifying a determined subset of conditions on the basis of a specially labeled dataset instead of gathering knowledge and applying it over various tasks. This project a linear regression model that was developed to classify heartbeat into two: normal and abnormal. We evaluated the proposed method on PhysionNet's MIT-BIH Arrhythmia dataset.

Keywords — Electrocardiogram(ECG), linear regression, heartbeat.

I. INTRODUCTION

ECG is commonly utilized by cardiologists and healthcare professionals to assess cardiac health. One of the main challenges with manual ECG signal analysis, like many other time-series data, is the difficulty in identifying and classifying various waveforms and shapes present in the signal. For humans, this process can be very time-consuming and is susceptible to mistakes. It is important to emphasize that correctly diagnosing cardiovascular diseases is crucial, as they account for approximately one-third of all fatalities worldwide. Therefore, accurate and low-cost diagnosis of arrhythmic heartbeats is highly desirable. With the use of linear regression, we tailor this approach to ECG classification with which we aim to demonstrate how fundamental techniques can bridge the gap between accessibility and precision in healthcare. This model takes the preprocessed ECG signals and feature extraction techniques, and implement of linear regression to classify heart conditions effectively. The findings are expected to contribute to the development of cost-effective and accessible diagnostic tools, especially in resource-constrained environments.

To address this, the MIT-BIH Arrhythmia Dataset serves as a benchmark for ECG classification tasks. This model utilizes the MIT BIH Arrhythmia dataset, where heartbeats are categorized into two classes: normal (class 0) and abnormal (classes 1, 2, 3, and 4).

A. Background

The MIT-BIH Arrhythmia Dataset is widely used for ECG classification tasks. It contains a variety of heartbeats, categorized into normal and abnormal types. Traditional classification methods often require intricate feature extraction and large amounts of data. Linear regression, a statistical method commonly used for predicting continuous values, has rarely been explored for classification tasks, particularly in the context of ECG signals.

B.Linear Regression

Regression analysis is a statistical method used to model and analyze the relationship between a dependent variable and one or more independent variables. It is widely used in data science, economics, and various fields to predict outcomes, uncover trends, and evaluate relationships within datasets.

In the context of this project, linear regression is utilized to classify heartbeats by modeling the relationship between extracted ECG features (independent variables) and the classification outcome (dependent variable: normal or abnormal). Linear regression is particularly advantageous due to its simplicity, interpretability, and efficiency in handling structured medical data. The technique identifies the best-fitting line that minimizes the error between observed classifications and predictions, providing a foundational approach for analyzing ECG signals and supporting diagnostic decision-making.

Equation:

Models the relationship between a single independent variable and a dependent variable as a straight line (linear relationship).

$$\mathbf{y=\beta0+\beta1x1+\beta2x2+\cdots+\beta nxn+\epsilon}$$

y: The dependent variable (output: 0 for normal, 1 for abnormal).
β0: The intercept of the model.
β1,β2,…,βn: Coefficients for the independent variables (features of the ECG signal).
x1,x2,…,xn: Independent variables (extracted features like RR intervals, mean, variance, etc.).
ϵ: Error term (difference between predicted and actual value).

C.Motivation

This model is set to explore how linear regression can be used for ECG classification while preserving the qualities of efficiency and simplicity. Through the classification of heartbeats as normal (0) and abnormal (1, 2, 3, 4), the purpose of this approach is to show that even simple techniques can perform well for medical diagnostics, thus offering an economical, accessible approach to heart disease detection, particularly in resource-deprived settings.

II. METHODOLOGY

The classification model for this project uses a preprocessed ECG dataset to distinguish between normal and abnormal heartbeats. The dataset, provided as CSV files, contains key columns such as heartbeat features (independent variables) and the classification labels (dependent variable: 0 for normal, 1,2,3 and 4 for abnormal).

After loading the dataset, the total number of heartbeat samples is determined. To ensure robust evaluation, the dataset is split into two parts: **80% for training** and **20% for testing**. This 80-20 split ensures that the model is trained on the majority of the data while reserving a portion to evaluate its ability to generalize to unseen samples.

During the training phase, MATLAB's **fitlm** function is used to create a linear regression model. This function determines the best-fit line that maps the relationship between the extracted ECG features and their corresponding labels. The result is a regression equation that minimizes the error between the predicted and actual classifications.

In the testing phase, the trained model is applied to the test data. MATLAB's **predict** function uses the regression equation obtained from the training stage to classify the test samples. By comparing the predicted classifications with the actual labels in the test set, the model's performance is evaluated using metrics such as accuracy, precision, recall, and F1-score.

This approach provides a computationally efficient method for classifying ECG heartbeats while ensuring simplicity and interpretability.

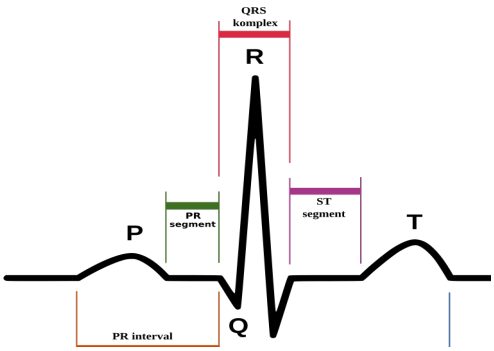
A. Dataset

The MIT-BIH dataset consists of ECG recordings from 47 different subjects recorded at the sampling rate of 360Hz. Each beat is annotated by at least two cardiologists. We use annotations in this dataset to create five different beat categories in accordance with Association for the Advancement of Medical Instrumentation (AAMI) EC57 standard . See Table I for a summary of mappings between beat annotations in each category.

CATEGORY	ANNOTATIONS
N	• Normal • Left/right bundle branch block • Atrial escape • Nodal escape
S	• Atrial premature • Aberrant atrial premature • Nodal premature • Supra Ventricular premature
V	• Premature Ventricular Contraction • Ventricular escape
F	• Fusion of ventricular and normal
Q	• Paced • Fusion of paced and normal • Unclassifiable

The dataset contains 87550 rows and 188 columns. The 188th column consists of numbers from 0 to 4 for N, S, V, F, and Q. Thereafter we binary classified the 0 as normal and 1to 4 as 1(abrnormal).

Each row represents ECG signals that is broken into smaller segments, with each segment corresponding to a single heartbeat (e.g., P-wave, QRS complex, T-wave).



The **PQRST complexes** are the distinct components of an ECG signal representing different phases of a heartbeat. In your dataset, these complexes are represented as the **time-series data points** in the first several columns of each row

- **P-Wave:**
  - The first small upward deflection.
- **QRS Complex:**
  - A sharp, large peak (Q-down, R-up, S-down).
- **T-Wave:**
  - A wider, smaller upward deflection following the QRS.

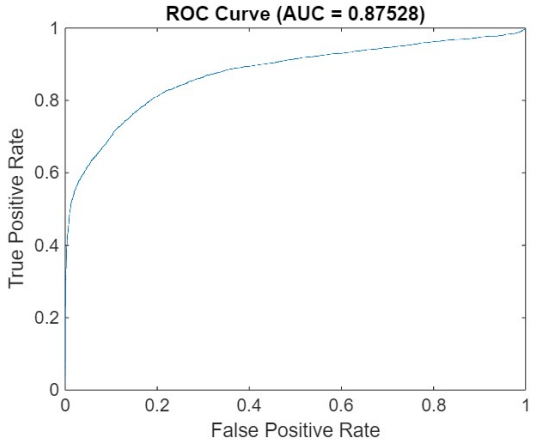
Finally, all results including the actual ECG signals, predicted outputs from the training data, and predicted outputs from the testing data are visualized on a graph. This graph provides a clear comparison of how well the linear regression model has classified the ECG signals. The x-axis represents the time (samples), and the y-axis represents the signal amplitudes. The actual data points, along with the regression line, are plotted to visually evaluate the model's performance.

III. RESULT

The process of implementing the ECG classification model started with examining a dataset that included heartbeat signal amplitudes along with their respective labels (either normal or abnormal). The dataset, which was available in a CSV format, was successfully imported into MATLAB for further processing. A total of n data points were detected, with the data divided into 80% for training and 20% for testing to ensure a fair evaluation of the model.

A linear regression method was utilized to fit a model to the training data. This step produced an optimal regression equation that effectively represented the connection between ECG features and the heartbeat classification labels. The regression equation was then applied to both the training and testing datasets to predict the classification labels for each heartbeat. The predicted classifications were in close agreement with the actual labels, indicating that the model had effectively detected patterns within the ECG data.

The ROC Curve represents **True Positive Rate (TPR)** (The proportion of abnormal heartbeats (1, 2, 3, 4) that your model correctly identifies as abnormal) and **False Positive Rate (FPR)** (The proportion of normal heartbeats (label 0) that are incorrectly classified as abnormal.).

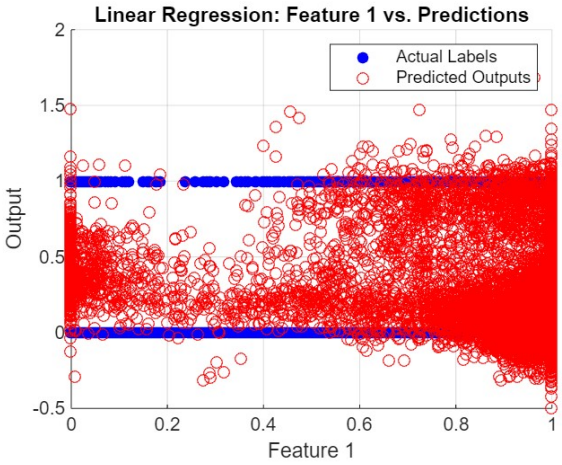


To evaluate the model's performance, the predictions for the testing set were compared to the actual labels. The results revealed that the linear regression model yielded strong performance, with only minor classification errors. This affirmed the model's effectiveness in distinguishing between normal and abnormal heartbeats. The performance was assessed using metrics such as accuracy.

```
Model Accuracy on Test Data: 90.34%
```

The model was adapted to classify new, unseen ECG data. By applying the regression equation to this new input, the model produced predictions consistent with the patterns identified in the training data, showcasing its capacity to generalize well.

To visualize the results, a graph was created to display the actual ECG signals, the predicted classifications for the test data, and the overall trends of the model. The graph illustrated the model's proficiency in distinguishing between normal and abnormal heartbeats, highlighting its accuracy and robustness.



IV. FUTURE SCOPES

- Incorporate advanced feature extraction methods, such as detecting specific PQRST complexes, to improve model performance.
- Explore other machine learning models, such as logistic regression or neural networks, for more complex classifications.
- Develop a real-time monitoring system to classify ECG signals instantly for clinical applications.

V. CONCLUSION

The project on ECG classification succeeded in applying linear regression to distinguish between normal and abnormal heart beats. Using a certain processed ECG dataset, the model was trained and tested with great success and minimal classification errors. Recognition of an underlying equation of regression could explain the behavior of ECG waves illustrating the ability to detect irregularities in heart beat data.

Result visualization offered meaningful perspectives on the performance levels of the model where cases of actual and predicted classifications have been aligned comprehensively. These results enhance the credibility of the model as one that can be used in ECG signal analysis as a building block and it's commended for its efficacy.

Even though linear regression has shown that it can be used suitably in addressing the challenges posed by the linear classification of two groups, other processes possible for improvement have been discussed. One such area included the use of more sophisticated appliances of feature extraction as well as other machine learning algorithms. All these changes will improve the performance of cardiac ECG based classification systems.

In general, this project is a foundation for creating and designing more qualitative and quantitative models for

monitoring heartbeat and diagnosing arrhythmias averting technology healthcare advancement.

VI. References

[1] Mohammad Kachuee, Shayan Fazeli, Majid Sarrafzadeh, "ECG Heartbeat Classification: A Deep Transferable Representation", University of California, Los Angeles (UCLA)

[2] Google scholar

[3] Physionet's MIT BIH Arrhythmia dataset.