Data Mining Approach to Detect Heartbeat Anomalies using ECG Images.

- Author(s):

Avinash Ramesh avinash.ramesh@sjsu.edu

Nisha M Devadiga nishamohan.devadiga@sjsu.edu

Poojashree NS poojashree.ns@sjsu.edu

- Abstract:

One of the most prominent tools for detecting cardiovascular problems is the electrocardiogram (ECG). The electrocardiogram (ECG or EKG) is a diagnostic tool that is used to routinely assess the electrical and muscular functions of the heart. Even though it is a comparatively simple test to perform, the interpretation of the ECG charts requires considerable amounts of training. Till recently, the majority of ECG records were kept on paper. Thus, manually examining and re-examining the ECG paper records often can be a time-consuming and daunting process.

If we digitize such paper ECG records, we can perform automated diagnosis and analysis. The main goal of this project is to use machine learning to convert ECG paper records into a 1-D signal. This can be achieved by extracting the P, QRS, and T waves that exist in ECG signals to demonstrate the electrical activity of the heart using various techniques. The techniques include splitting the original ECG report into 13 Leads, extracting and converting into the signal, smoothing, converting them to binary images using threshold and scaling. Post-feature-extraction, dimension reduction techniques like Principal Component Analysis are applied to understand the data. Multiple classifiers like k-nearest neighbors (KNN), Logistic Regression, Support Vector Machine (SVM), and Voting Based Ensemble Classifier are implemented, and based on the acceptable criteria on the accuracy, precision, recall, f1-score, and support, the model will be finalized. This final model will aid in the diagnosing of cardiac diseases, to detect whether a patient has/had Myocardial Infarction, Abnormal Heartbeat, or the patient is hale and healthy by inferring the ECG reports.

Introduction

According to the World Health Organization, heart disease is the first leading cause of death in the high and second leading cause of death in low-income countries. It has remained the leading cause of death at the global level for the last 20 years. This paper aims to analyze several data mining techniques implemented in recent years for diagnosing heart disease.

At present, there are plenty of algorithms available that could detect and predict heart anomalies from clinical reports. However, in this project, the focus is more on discovering and extracting patterns from Electrocardiogram (ECG or EKG) image reports. By digitizing ECG records, the need for time-consuming manual intervention for comprehending the report can be eliminated. With digitization, the automation of diagnosis and analysis can be achieved quicker.

Related Work

Many papers related to cardiovascular prediction focused on other features that included diet, age, gender, and many other dimensions, and then predicted for cardiovascular diseases based on these features. Our work is more on predicting diseases by providing the ECG chart to our model.

Data

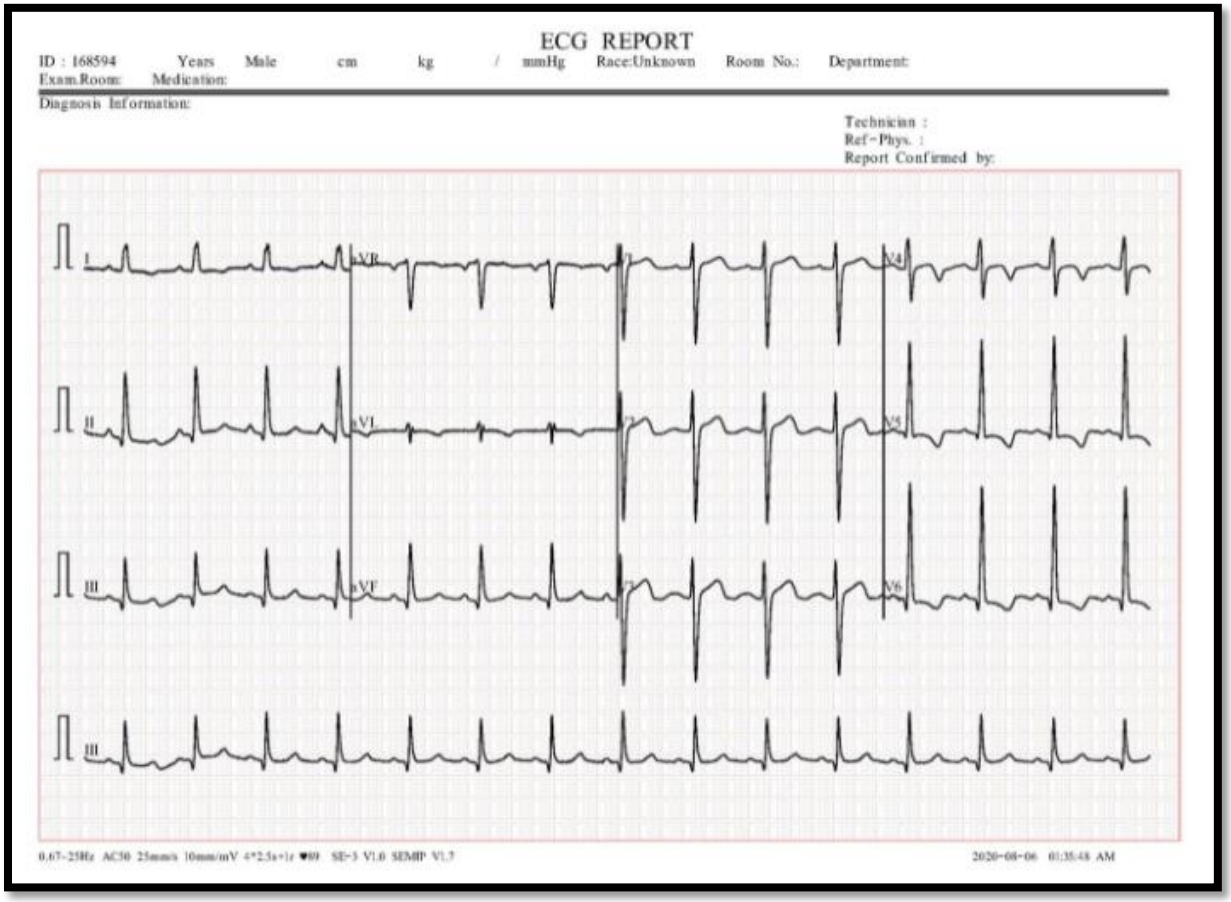
ECG images: https://data.mendeley.com/datasets/gwbz3fsgp8/2

The above dataset contains ECG image signals from both healthy individuals and persons with cardiovascular problems.

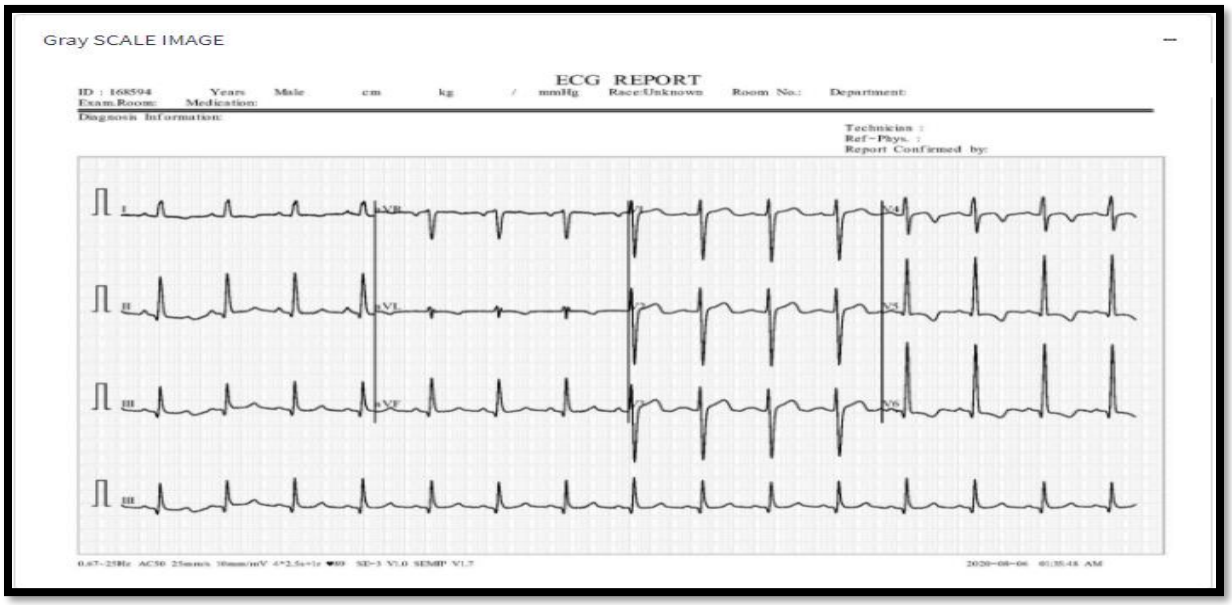
All the ECG images from the above Mendeley dataset are parsed and transformed per our business requirements. This is the most important phrase of our application (data preparation, data cleaning, data engineering and feature extraction).

Each one of the images goes through various processes to extract the data in resultant format.

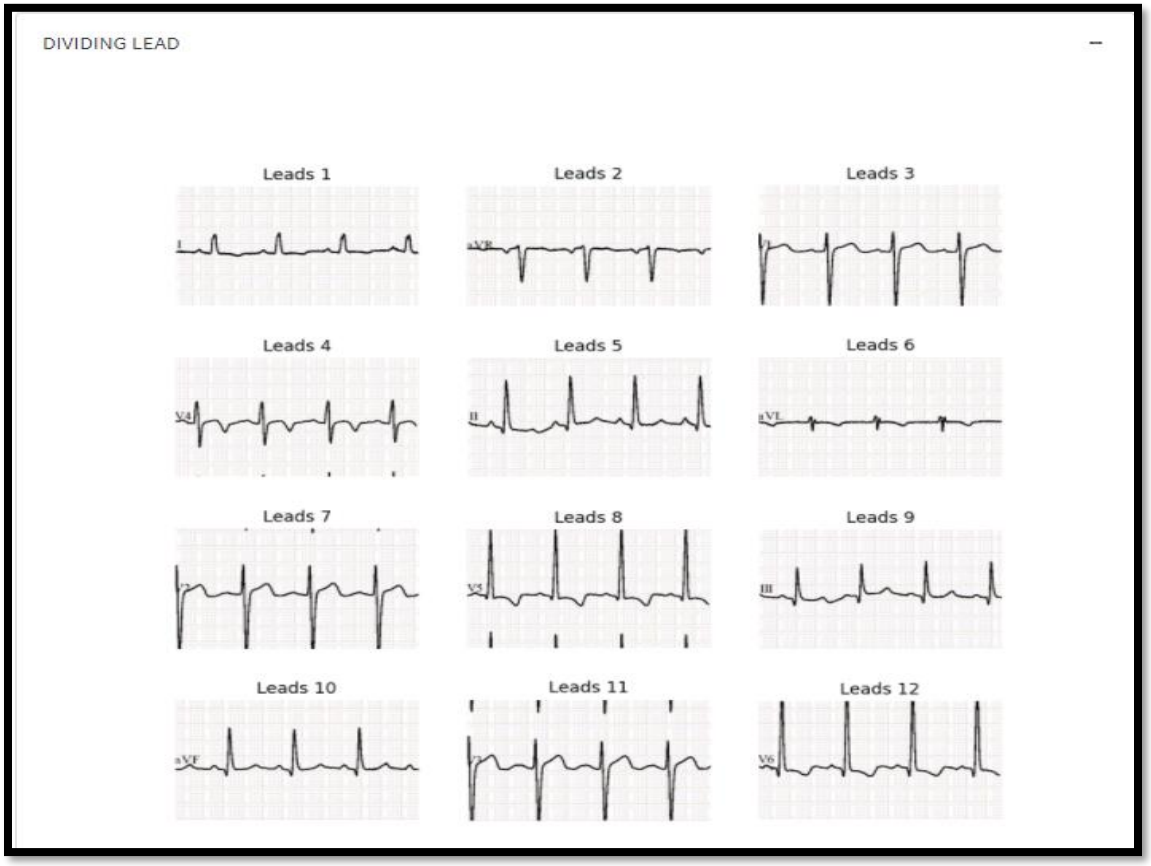
Input Image

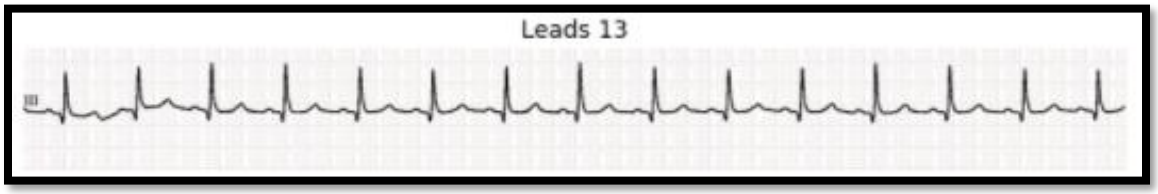


Grey-Scale Image:



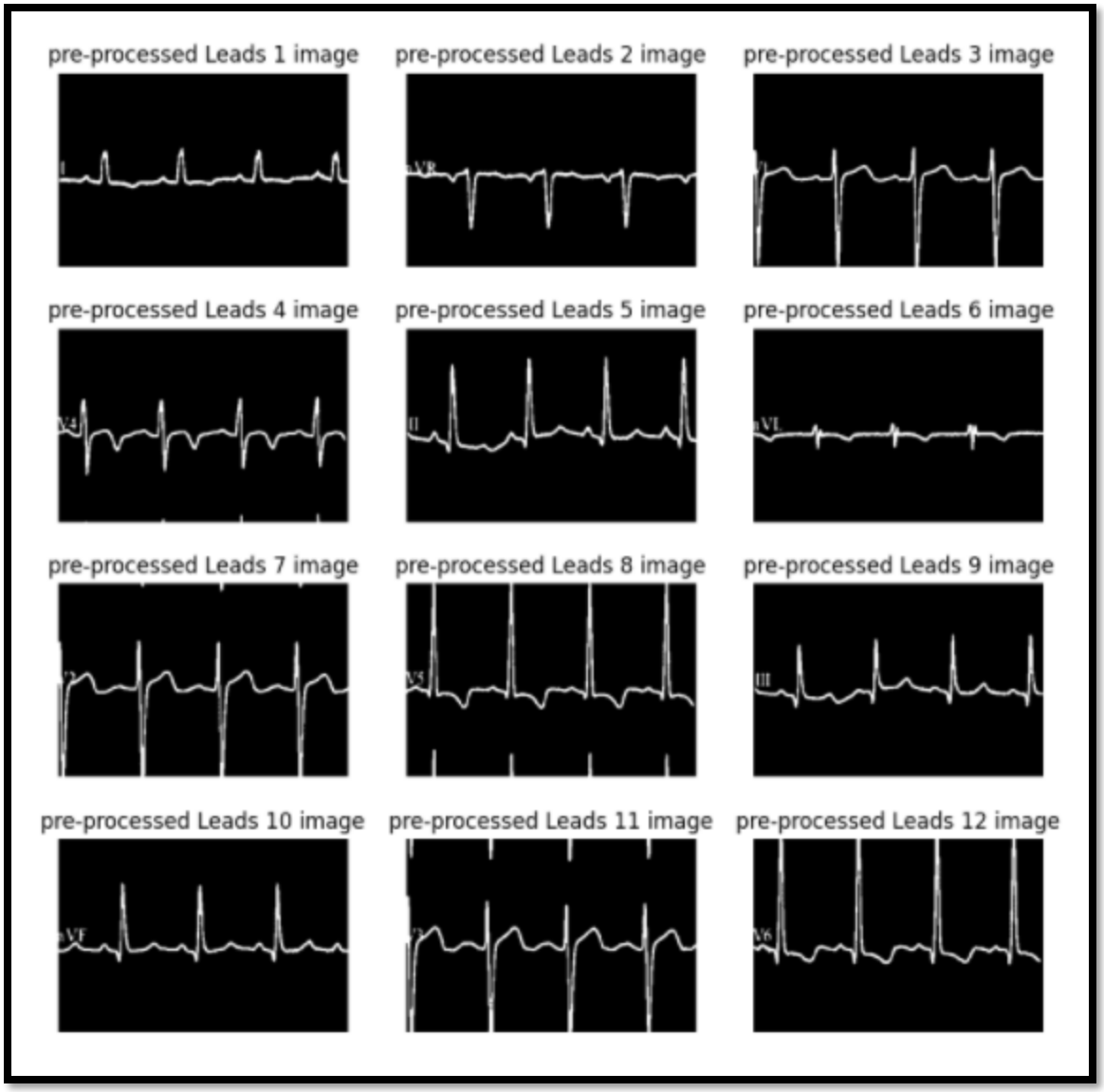
Dividing into different Leads:

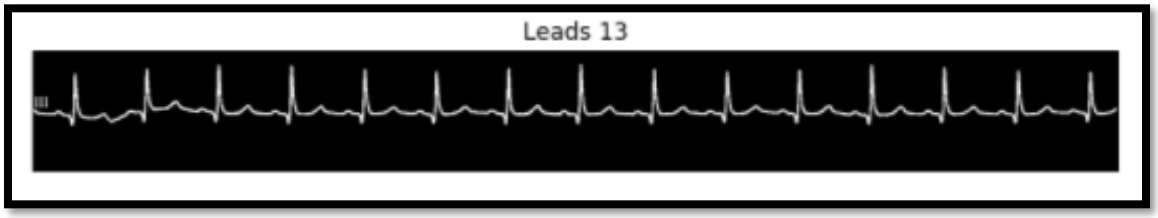




Data Cleaning & Feature Transformation:

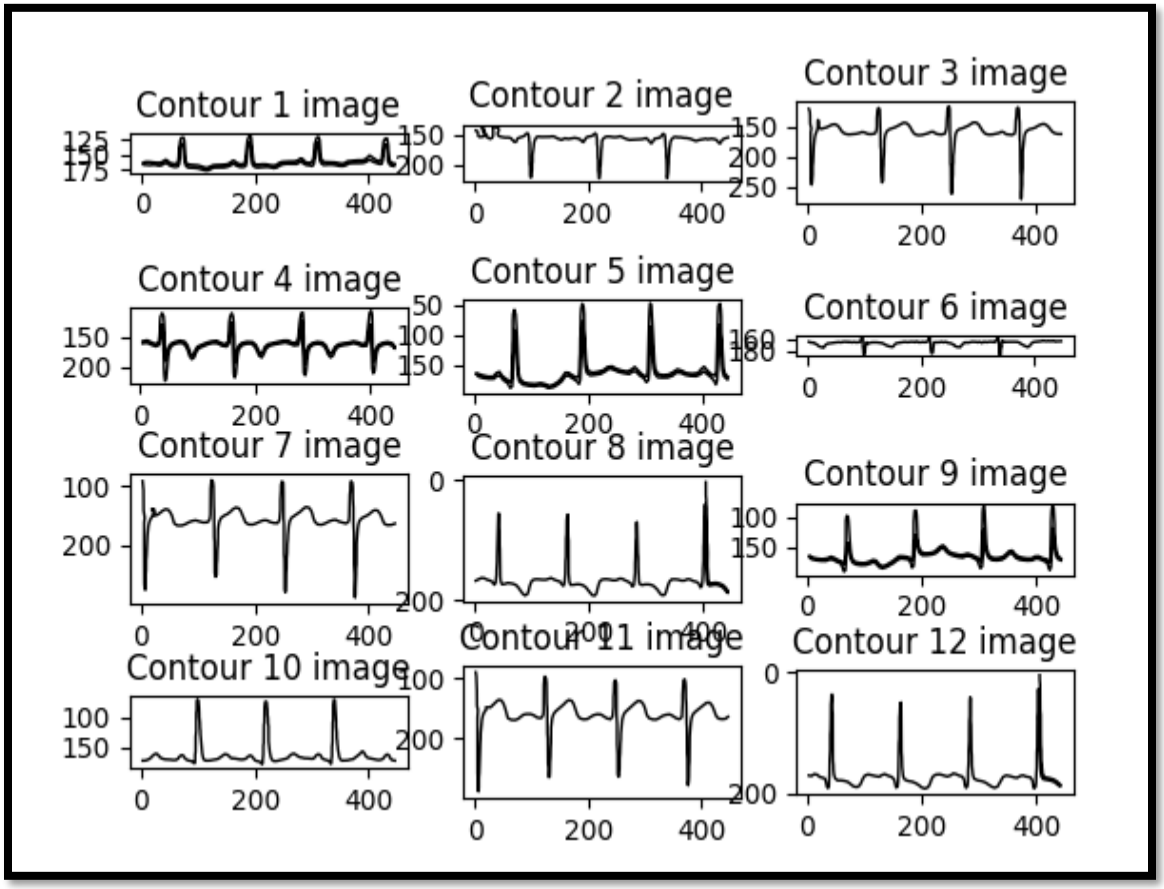
To prepare Leads (1-13) for further processing, each individual lead image is transformed by **removing Gridlines, converting to Grayscale, applying Gaussian filtering, and performing Thresholding** to convert to binary image.





The transformed image is **traced** to extract only the signals from the image using the **contour technique**, and the values are scaled using the MinMax Scalar. The normalized output is saved in CSV format as a 2D signal.

Contour Images:



Normalized 2D Signal

	X	Y
0	0.575342	1.000000
1	0.573973	0.999793
2	0.560274	0.999793
3	0.546575	0.999793
4	0.532877	0.999793
...	...	...
950	0.383562	0.004132
951	0.395890	0.002273
952	0.397260	0.002066
953	0.409589	0.000207
954	0.410959	0.000000

It is noticed from the above observation that the X-axis corresponds to the high and low points and the y axis corresponds to curve/shape. In our analysis, the focus will be more on the low/high points, hence the X-axis will be saved separately as a normalized scaled 1D signal in a CSV file.

Normalized 1D Signal

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
X	0.859944	0.871148	0.87395	0.87395	0.87395	0.87395	0.87395	0.87395	0.885154	0.887955	0.887955	0.887955	0.887955	0.89916	0.901961	0.901961	0.913165	0.915966	0.9159

1 rows × 1731 columns

We transformed all the 1D rows into columns using transpose. With both 1D and 2D CSV files and cropped 1 to 13 lead images, we perform different Supervised classification algorithms: k-nearest neighbors (KNN), Logistic Regression, Support Vector Machine (SVM), and Voting Based Ensemble Classifier on based on CSV DATA.

Combining All Leads

1D Signals

	0	1	2	3	4	5	6	7	
0	0.8781	0.8628	0.7991	0.7105	0.5939	0.4702	0.3533	0.2647	0.297

Once we have extracted all of the image's 12 lead 1D values, we combine them into a single csv for further analysis.

Performance Dimensionality Reduction

Dimensional Reduction

	0	1	2	3	4	5	6	7	
0	-1.5831	4.6774	-0.6037	-3.3532	-1.2244	0.3372	-3.0771	0.6421	

Methods

Many researchers have used different data mining techniques, for identifying and predicting heart anomalies, like Neural networks, Classification based on Clustering, KNN, Decision Trees. In this project, we perform different Supervised classification algorithms: k-nearest neighbors (KNN), Logistic Regression, Support Vector Machine (SVM), and Voting Based Ensemble Classifier based on CSV DATA.

Before performing data modeling, ECG images categorically belonging to four categories of patients i.e., patients with Myocardial Infarction, Abnormal Heartbeat, Myocardial Infarction History, and good Health are combined on the lead level (from 1 to 12) and then convert target column with `array (['No', 'HB', 'MI', 'PM']` into numeric using groups encoder.

Post dimension reduction technique like Principal component Analysis is applied to understand the data and validate the variance explained is under acceptable limit. Here, in this case, `Total Variance Explained: 99.5.`

Post Dimension reduction, following data mining techniques, are applied on 12 leads combined: -

- **k-nearest neighbors (KNN): -**

K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm that can be used for both classifications as well as regression predictive problems. However, it is mainly used for the classification of predictive problems in the industry.

```
Accuracy: 0.793010752688172
              precision    recall  f1-score   support

           0       0.92      0.65      0.76       105
           1       0.95      0.91      0.93        94
           2       0.70      0.86      0.77       112
           3       0.65      0.74      0.69        61

    accuracy                           0.79       372
   macro avg       0.80      0.79      0.79       372
weighted avg       0.81      0.79      0.79       372

Tuned Model Parameters: {'knn__n_neighbors': 1}
```

- **Logistic Regression: -**

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

```
Accuracy: 0.7741935483870968
              precision    recall  f1-score   support

           0       0.83      0.56      0.67       105
           1       0.83      0.91      0.87        94
           2       0.82      0.86      0.84       112
           3       0.59      0.77      0.67        61

    accuracy                           0.77       372
   macro avg       0.77      0.78      0.76       372
weighted avg       0.79      0.77      0.77       372

Tuned Model Parameters: {'lr__C': 0.3593813663804626, 'lr__penalty': 'l2'}
```

- **Support Vector Machine (SVM): -**

Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms that are used both for classification and regression. But generally, they are used in classification problems. SVMs have their unique way of implementation as compared to other machine learning algorithms.

```
Accuracy: 0.9051724137931034
              precision    recall  f1-score   support

           0       0.81      0.92      0.86       119
           1       1.00      1.00      1.00       125
           2       0.91      0.89      0.90       140
           3       0.93      0.78      0.84        80

    accuracy                           0.91       464
   macro avg       0.91      0.89      0.90       464
weighted avg       0.91      0.91      0.91       464

Tuned Model Parameters: {'SVM__C': 10, 'SVM__gamma': 0.01}
```

- **XGBoost: -**

  XGBoost is an implementation of gradient boosted decision trees designed for speed and performance that is dominative competitive machine learning.

```
Accuracy: 0.8836206896551724
              precision    recall  f1-score   support

           0       0.85      0.76      0.81       119
           1       0.98      1.00      0.99       125
           2       0.81      0.93      0.86       140
           3       0.93      0.80      0.86        80

    accuracy                           0.88       464
   macro avg       0.89      0.87      0.88       464
weighted avg       0.89      0.88      0.88       464
```

- **Voting Based Ensemble Classifier with GridSearchCV: -**

  Under Voting-based Ensemble classification, three Machine learning models like k-nearest neighbors (KNN), Support Vector Machine (SVM) and Random Forest Classifier are stacked and voted to pick one model which gives the highest accuracy.

  For tuning the hyperparameters we have used GridSearchCV.

  Based on the voting, the classification report is printed as shown below:

```
Accuracy: 0.9247311827956989
              precision    recall  f1-score   support

           0       1.00      0.94      0.97        80
           1       1.00      1.00      1.00        72
           2       0.84      0.92      0.88        79
           3       0.84      0.79      0.82        48

    accuracy                           0.92       279
   macro avg       0.92      0.91      0.92       279
weighted avg       0.93      0.92      0.93       279

{'SVM__C': 1, 'SVM__gamma': 0.1, 'knn__n_neighbors': 5, 'rf__n_estimators': 300}
```

Once the model is acceptable, we will pickle the model for future use and prediction. In this case, the model is pickled into model_test.pkl.

Experiments and Results

First, the user uploads ECG images to our web app. From there, image conversion techniques such as rgb2gray conversion, denoising, Gaussian Filtering, thresholding and contouring are implemented to extract signals without the grid lines. The signal is then dimensionally reduced, and the necessary waves (P, QRS, T) are extracted using segmentation and fed into our pre-trained model from the analysis. Once the model finishes the analysis, it returns the results back to the user based on the findings.

Conclusion

The empirical results show that we can produce faster and accurate predictions for heart patients by applying the given predictive model to the ECG images of new patients. This study can also be extended to include multiple different heart diseases if the feature extraction from images is done correctly and optimally along with increased accuracy of our model.