ECG based Decision Support System for Clinical Management using Machine Learning Techniques

Sireesha Moturi 1, Dr. Srikanth Vemuru 2, Dr. S. N. Tirumala Rao 3

1Research Scholar, KLEF, Vaddeswaram, India, Assoc. Prof., Narasaraopeta Engineering College, Narasaraopet, India
2Professor, KLEF, vaddeswaram,India
3Professor& HOD, Narasaraopeta Engineering College, Narasaraopet, India
sireeshamoturi@gmail.com
vsrikanth@kluniversity.in
nagitirumalarao@gmail.com

Abstract. Heart disease prediction system using ECG is to predict heart disease using ECG signals. Heart is the next major organ comparing to brain, which has more priority in human body. Heart disease diagnosis is a complex task which requires much experience and knowledge. The huge amount of data generated for prediction of heart disease is too complex and voluminous to be processed by traditional methods. By using traditional methods doctors took lot of time to diagnosis the disease. So, an entropy based feature selection technique is used with classification algorithms in order to reduce the search space. The proposed model was tested on the real time dataset of NRI Hospital medical data. Using this system it is easier to predict the disease. It will also helpful for the doctors to take quick decisions.

1. Introduction
Heart attacks are one of the major reasons behind several deaths happening worldwide [1]. An electrocardiogram is a graphical record of the extent and course of the electrical movement that is created by contraction and relaxation of the ventricular and atria electrocardiogram is used to check the heart rate by placing several electrodes ECG because of their age, weight, high cholesterol level, obesity etc. Any change in the rhythm of the ECG signal is due of heart disease. According to the survey in 2004 1.1 million people died from heart disease, a total of 72 million of these deaths are due to heart disease and 5.7 million were due to stroke A recent survey has found out that about 23.6 million people will die of cardiovascular disease by 2030. Chronic disease requires constant treatment to enhance the quality life of patients. Nowadays, it is estimated that 12% of natural deaths occur accidentally, 88% of which are cardiac. The identification of early heart beat rhythms plays a vital role in preventing heart disease. Now a day’s data is generating more and more [14] so, Content extraction becomes a more challenging task in the today’s world [7]. Information on the web is also of different types which contains structured, semi structured and unstructured kind of data and current websites present a larger wide variety of complexities than traditional ones[13]. Data mining is about describing the past and projecting the future for analysis. It also helps to derive information from large datasets [3] [4]. This involves data planning, evolution, data interpretation, modelling and deployment. To deal this complex data, there is a need of dominant tool like Machine Learning (ML) [2].

2. Literature Survey
The healthcare system produces massive quantities of data every day. Much of it isn't used successfully though. Animesh Hazra et al.[5] suggested that any of the latest work on heart disease prediction using data mining techniques analyses the various combinations of mining algorithms used and conclude which techniques are successful and efficient. A non-stationary signal, the electrocardiogram (ECG) is commonly used to measure the rate and frequency of heartbeats. A comparison of the overall pattern and shape of the ECG waveform helps doctors to identify possible illnesses. Currently, a computer-based diagnosis is conducted using some signal processing to diagnose an ECG-based patient. The feature extraction scheme for subsequent analysis specifies the amplitudes and intervals in the ECG signal or any other features there. Recently various research techniques for analysing the ECG signal have been developed [6].

3. Proposed Model
3.1. Experimental setup: This experiment was conducted on the Intel® Core™ i7 Processors with 64 bit Windows 10Pro machine. Anaconda 5.1.0 jupyter notebook Python distribution is used in this experiment. This dataset is gathered from real-time data. We have collected the real-time information in NRI hospital Vijayawada. We gathered the ECG records of the patients and note down the values in an excel sheet. The dataset consists of 9 attributes such as PR interval [12], QRS duration, QTC interval, QT interval, vent rate, P wave, T wave, QRS wave and problem. Out of these attributes, class variable is problem attribute and remaining attributes are used as the predictor variables. Figure 1 shows the process of the proposed system of Heart Disease prediction using ECG signals.

Figure 2 shows the Electrocardiogram of a healthy heart. Figure 3 shows normal ECG. Figure 4 shows abnormal ECG. Figure 5 shows Borderline ECG. Figure 6 Normal ECG except with rate. Table 1 displays attributes, range and description about the attributes.

| s.no | Attributes  | Standard range | Description                                           |
|------|-------------|----------------|-------------------------------------------------------|
| 1.   | Vent_rate   | 60-100bpm      | It normally refers to the ventricular contraction rate, Nothing but heart beat [13] |
| 2.   | PR_interval | 120z⁻²-200ms   | The PR interval is the time from the beginning of the p wave to the start of the QRS complex. |
3. **QT_interval** 300-440 ms  
   The QT interval is the time from the start of the QRS series, representing ventricular depolarization to the end of the ventricular repolarization of the T wave.

4. **QTC_interval** 400-440 ms  
   It is nothing but corrected QT interval.

5. **P_axes** 110 ms  
   The p wave shows the first positive deflection on the ECG and atrial depolarization.

6. **T_axes** 160 ms  
   T wave reflects ventricular repolarization.

7. **QRS duration** 80-120 ms.  
   The “QRS complex” is a variation of the Q wave, R wave and S. This reflects depolarization to the ventricle.

8. **QRS_axes** <100 ms  
   QRS axes applies to frontal plane ventricular depolarization.

9. **Problem** -  
   Describes whether it is normal ECG or abnormal ECG or borderline ECG or normal except with rate.

| Attribute       | Description                                                                 |
|-----------------|-----------------------------------------------------------------------------|
| 3. QT_interval  | 300-440 ms  
   The QT interval is the time from the start of the QRS series, representing ventricular depolarization to the end of the ventricular repolarization of the T wave. |
| 4. QTC_interval| 400-440 ms  
   It is nothing but corrected QT interval. |
| 5. P_axes       | 110 ms  
   The p wave shows the first positive deflection on the ECG and atrial depolarization. |
| 6. T_axes       | 160 ms  
   T wave reflects ventricular repolarization. |
| 7. QRS duration| 80-120 ms.  
   The “QRS complex” is a variation of the Q wave, R wave and S. This reflects depolarization to the ventricle. |
| 8. QRS_axes     | <100 ms  
   QRS axes applies to frontal plane ventricular depolarization. |
| 9. Problem      | -  
   Describes whether it is normal ECG or abnormal ECG or borderline ECG or normal except with rate. |

Table 1 Attribute Description

---

3.2. **Data Preprocessing**

Data preprocessing is a data mining technique used to convert the raw data into an effective and usable format.

3.2.1 **Handling Missing Data**

Sometime data may contain insignificant and missing values. Data cleaning is required to handle that portion. It includes managing data which are missing, noisy data, etc. It can be done by manually filling in the missing values, by mean attribute, median or by dropping the missing values. But, dropping missing values is not so good technique to opt because we miss the valuable data. From figure 7 we can...
observe that the dataset is having missing values. In Figure 8 we replaced missing values with mean so that we don’t have any missing values.

3.2.2 Correlation
Correlation is a statistical measure that demonstrates how often two or more variables fluctuate together. A positive relationship shows degree to which those factors positively correlated to one other. A negative relationship shows degree to which those factors negatively correlated to one other. From the Figure 9 it is observed that the vent_rate and QT_interval are negatively related to one other. So, the attribute vent_rate got dropped.

3.2.3 Information Gain
This is also one of the pre-processing techniques, which is used to measure the reduction in entropy. It is widely used to build a model like decision tree from a training dataset, to calculate the information benefit of a variable and to choose a variable that maximizes the benefit of information, thereby dividing the dataset into groups for successful classification. From Figure 10 it is observed that PR_interval has the lowest information gain. Information gain can be calculated using the following formula.

\[ I(\theta_1, \theta_2, \ldots, \theta_m) = -\sum_{i=1}^{m} \pi_i \log_2(p_i) \]

4. Implementation
4.1 Splitting the Dataset
Dataset is splitted into two parts such as the train set and test set. It means 80% of data into the train set and the remaining 20% of data into the test set as shown in Figure 11.

**Figure 11. Splitting of dataset**

### 4.2 K-Fold Cross Validation

It is a kind of resampling technique used to evaluate the Machine Learning Model. In every iteration one fold is used for validation or testing and the remaining folds are used for training purpose as shown in Figure 12.

**Figure 12. K-Fold Cross Validation**

**Figure 13. Confusion Matrix**

### 4.3 Algorithms Used

There are many algorithms for classification in Machine Learning. The present paper gives a comparison between the performances of four classifiers: Decision Tree, Random Forest, Support Vector Classifier (SVC), and Gaussian Naïve Bayes. The main aim of researchers is to evaluate the efficiency of these algorithms in terms of accuracy. For Performance Evaluation Confusion matrix is used to evaluate the performance of any classification algorithm as shown in Figure 13.

### 5. Result Analysis

Figure 14 shows the accuracy comparison of three classifiers (Decision Tree, Gaussian Naïve Bayes, SVC). From Figure 14 it is observed that the highest accuracy is 97% for Decision Tree and Gaussian Naïve Bayes. Figure 15 shows the accuracy comparison of three classifiers (Decision Tree, Gaussian Naïve Bayes, SVC) when K-fold cross validation is applied. From fig14 it is observed that the highest accuracy is 98.2% for Decision Tree. It is also observed that all the three classifiers achieves highest performance after applying K-fold cross validation.

**Figure 14. Comparison of Accuracy**

**Figure 15. Comparison of Accuracy**
Figure 16 shows the Comparison accuracy of three classifiers (Decision Tree, Gaussian Naïve Bayes, SVC) before and after applying Information Gain. It is observed that after applying Information Gain the accuracy of decision tree has increased from 97.3 to 98.2.

![Figure 16. Before and after applying Information Gain](image)

6. CONCLUSION
In this paper various pre-processing techniques were applied on the data with three Machine Learning algorithms like Decision Trees, Gaussian Naïve Bayes and SVC in- order to predict presence or absence of heart disease. The accuracy varies for different algorithms. The highest accuracy 98.2% was achieved by Decision tree with Information Gain and K-Fold cross validation methods. By using this system we can reduce medical errors, enhance patient safety and improve patient outcomes. It is easier to predict the disease and it is also helpful for the doctors to make quick decisions.

7. References
[1] M.Sireesha, S.N.TirumalaRao and Srikanth V 2019 optimized Feature Extraction and Hyrid Classification Model for Heart Disease and Breast cancer Prediction International journal of Recent Technology and Engineering vol-7, No 6 pp 1754-1722.
[2] Prakash K.B and Dorai Rangaswamy M.A. 2016 Content extraction studies using neural network and attribute generation. Indian Journal of Science and Technology pp 1-10.
[3] M.Sireesha, Srikanth vemuru and S.N.Tirumalarao 2018 Coalesce based binary table: an enhanced algorithm for mining frequent patterns International journal of Engineering & Technology, vol. 7, no.1, pp 51-55.
[4] M.Sireesha, S.N.Tirumalarao and Srikanth V 2018 Frequent itemset Mining Algorithms: A survey Journal of Theoretical and Applied Information Technology vol 96, No.3 pp 744-755.
[5] Animesh H, Subrata Kumar M, Amit G, Arkomita M and Asmita M 2017 Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review Advances in Computational Sciences and Technology Vol 10, No 7 pp 2137-2159.
[6] Naveen Ku. Dewangan and S. P. Shukla 2015 A Survey on ECG Signal Feature Extraction and Analysis Techniques International Journal Of Innovative Research In Electrical, Electronics, Instrumentation And Control Engineering Vol. 3 Issue 6 pp 12-19.
[7] Prakash K.B and Rangaswamy M.A.D 2016 Content extraction of biological datasets using soft computing techniques Journal of Medical Imaging and Health Informatics pp 932-936
[8] M.Sireesha, S. N. TirumalaRao and Srikanth V 2020 Predictive Analysis of Imbalanced Cardiovascular Disease Using SMOTE International Journal of Advanced Science and Technology Vol 29 No 5 pp 6301 – 6311.
[9] A.Dhanasekar and Dr.R.Mala 2016 Analysis of Association rule for heart disease prediction from large datasets International Journal of Innovative Research in Science, Engineering and Technology Vol 5, Issue 10 pp 18059-18063
[10] M.Sireesha, Srikanth V and S.N.Tirumala Rao 2020 Classification Model for Prediction Of Heart Disease Using Correlation Coefficient Technique International Journal of Advanced Trends in Computer Science and Engineering Vol 9, No 2 pp 2116 – 2123.
[11] A. Calderon, A. Pérez and J. Valente 2019 ECG Feature Extraction and Ventricular Fibrillation (VF) Prediction using Data Mining Techniques IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), Cordoba, Spain, 2019, pp. 14-19
[12] https://www.healio.com/cardiology/learn-the-heart/ecgreview/ecginterpretation-tutorial
[13] Prakash K.B., Dorai R M.A., and Raman A.R. 2010 Proc. Int. Conf. on Trendz in Information
Sciences and Computing pp 28-31.

[14] M. Sireesha, Srikanth Vemuru and S.N. Tirumala Rao 2020 Int. Conf. On Machine Intelligence And Soft Computing.