International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 9 ǁ September 2022 ǁ PP. 395-403

ECG signal Classification using Machine Learning Techniques

[1]Gayathri Sampath[*],Ayishabanu Syed Ibrahim[1], L. Malathi[1] N. Mohanapriya[1]
Omkar Ashok Banne[2], J. John Rozario Jegaraj[3]

[1]*Department of Computer Science and Engineering, Vivekanandha College of Engineering for Women, Namakkal, Tamilnadu, India.*
[2]*Advanced Manufacturing Technology Development Centre (AMTDC), IIT Madras, Chennai, Tamilnadu, India*
[3]*Defense Research and Development Laboratory (DRDL), Hyderabad, Telangana, India*

Abstract

The Electrocardiogram (ECG) Signal Analysis is an important process for diagnosing the health status of the patient. Analysis of this electrical signal is very challenging due to the involvement of various types of noises, insufficient data gathering, large amounts of data and improper dependencies. So, It is very important to detect heart diseases as early as possible so that the number of deaths can be reduced. This paper aims to analyze and implement an automatic heart disease diagnosis system using MATLAB and Python. The ECG dataset obtained from Bilkent University machine learning repository was used as the main database for training and testing of the machine learning model and the accuracy of each machine learning technique were compared.

Keywords:*Electrocardiogram signal analysis; ECG; Heart diseases; Machine learning; Linear regression; Random Forest classifier; ANFIS.*

Date of Submission: 03-09-2022 Date of acceptance: 17-09-2022

I. INTRODUCTION

Heart diseases are one of the main reasons for human death in many countries. World Health Organization (2022) states that in 2019, 17.9 million people died from heart disease representing 32% of all global deaths, 85% were due to heart attack and stroke. Most cardiovascular diseases can be prevented by finding out some risk factors such as use of harmful alcohol and tobacco. The World Health Organization agreed on a global mechanism of avoiding heart disease by fixing up the target like at least 50% of eligible people should receive drug therapy to reduce the heart related problems by 2025 [1]. It is important to find out the cardiovascular disease as early as possible to reduce the death counts. The growth of heart disease, complications and high-cost treatment caused the medical field to search for an effective solution.

Heart disease refers to a range of conditions that affect the heart. Major Heart diseases include coronary artery disease, arrhythmias, congenital heart defects and heart valve disease.Heart disease symptoms such as high blood pressure, chest pain, shortness of breath, an irregular heartbeat, or heavy heartbeats. Common causes of heart diseases are Diabetes, Drug abuse, Excessive use of alcohol or caffeine, birth Heart defects, High blood pressure, Smoking, Stress [2]. Heart is divided into two portions that are atrium and ventricle. It pumps blood to the lungs through the pulmonary arteries. The lungs give the blood as a supply of oxygen. The wave of electric depolarization spreads from the atrium down through the interventricular septum (IVS) to ventricles. So, the direction of this depolarization is usually from the superior to the inferior aspect of the heart. Recording the heart's electrical activity through Electrocardiogram (ECG) gives a graph line that shows the various changes in the electrical activity of the heart. ECG recording usually takes 5 to 10 minutes. The ECG signal is measured by placing a series of electrodes on the patient's skin. There are 12 sensors often known as electrodes that are fixed to the chest, legs and arms [3]. These electrodes are sticky patches with wires that are connected to the monitor. These electrical signals will make the heartbeat count. A computer records the information and displays it as waves on a monitor or on paper. Hence ECG signal analysis is a major challenge in order to eliminate noise and preprocess the data. Some of the important parameters of ECG Signal are QRS Duration, R wave, PR interval, P wave, QT interval, T wave. The QRS Duration is a combination of three waves Q, R and S waves, it represents the ventricular depolarization. The PR interval is the duration from the start of the P wave to the start of the QRS complex. The Depolarization of the right atrium is responsible for the early part of the P wave, and depolarization of the left atrium is responsible for the middle and terminal portions of the P wave, QT interval is the time from the start of the Q wave to the end of the T wave. It mainly records how often the heart beats (heart rate) and how regularly it beats (heart rhythm).

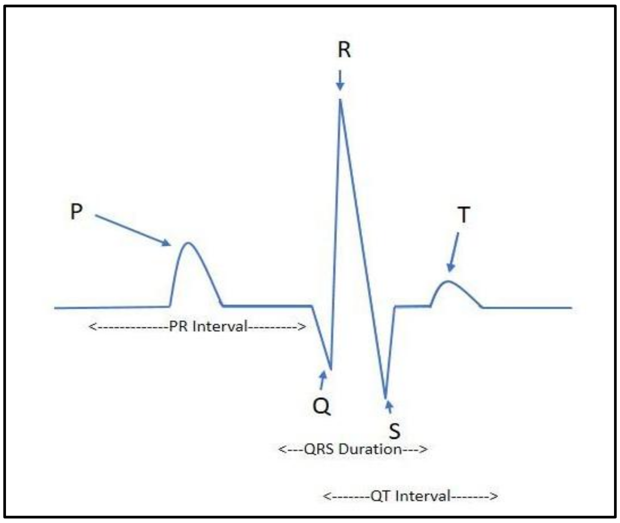

R
P
T
<-------------PR Interval--------->
Q
S
<---QRS Duration--->
<-------QT Interval------->

Figure 1 Typical ECG Signal

ECG Signal Analysis and interpretation is often performed by professional doctors, which extremely depends on their knowledge. Sometimes, even experienced researchers are also not able to get enough information from the ECG signals recordings. The advancement of Machine Learning algorithms and automated diagnostic models plays an important role in the prediction of heart disease. To reduce the time for diagnosis and cost of treatment an approach using machine learning is attempted in this paper. Machine learning model analyzes the ECG signal and classifies it into different types of diseases based on the class number associated with each disease. Normal people refer to 'class 01', ‘class 02 to 15’ refers to different classes of arrhythmia and unclassified diseases refer to class 16 [4]. Machine Learning approaches based on ANFIS, Linear Regression and Random Forest Classifier have been used for the identification and classification of the disease. The effectiveness of these approaches is compared and discussed. In the ANFIS and Linear Regression the heart disease is classified into 16 groups whereas in Random Forest Classifier the heart disease is classified into two groups based on the risk factor.

II. Methodology

ECG signals are often known as time-varying signals which have a less range of amplitude from 10 μV to 5 mV. Their typical value is 1 mV, and their frequencies range from 0.05–100 Hz, mainly concentrated in the 0.05~35 Hz range [5]. There are 6 major input values used in this project such as: Sex, QRS duration, PR interval, Q-T interval, P interval, Heart rate and also some personal information like age, height according to these attributes the interval values change for each person [6]. The workflow of the ECG Signal Analysis is given below.

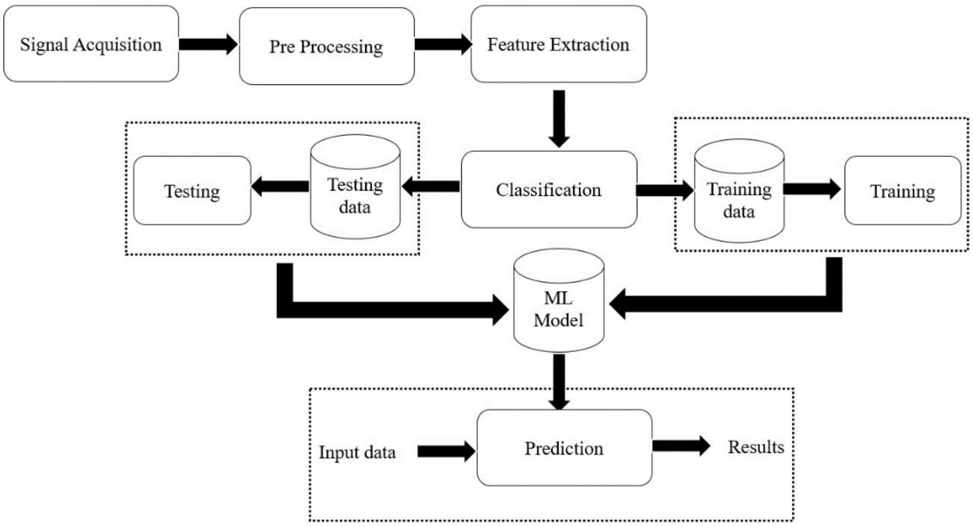


Figure 2 Workflow of ECG Signal Analysis implemented in this project

2.1 Signal Acquisition

An electrocardiograph is a graph which is recorded by electric potential changes that occur between electrodes placed on a patient's body to demonstrate cardiac activity. The electric potential created by contraction and relaxation of the heart wall spreads the currents from the heart throughout the body. This current creates different potentials at various points which can be sensed by electrodes placed on the patient's body. The electrodes which are used in ecg signal recordings are biological transducers made of metals and salts. There are 12 electrodes which are attached to different points on the patient's body. There exists some standard procedure for acquiring ECG signals. The electrocardiogram captures not only weak signals from 0.5 mV to 5.0 mV, but also a DC component of up to ±300 mV and a common mode of up to 1.5 V. The bandwidth of an ECG signal depends on the application and this ranges from 0.5-100 Hz, sometimes reaching up to 1 kHz. Some of the noise includes movement that affects the skin-electrode interface, muscle contractions or electromyographic spikes, respiration, electromagnetic interference (EMI), and noise from other electronic devices that couple into the input [7]. The resources for the dataset can be found at the open ml database. This database contains 279 attributes, 206 of which are linear valued, and the rest are nominal. The aim is to distinguish between the cardiac arrhythmia patients and the normal people and to classify it in one of the 16 groups. Currently there are some computer programs that make such classification [8]. However, there are variations between the cardiologist’s and the computer application’s classification. Taking the cardiologists as a standard this aims to minimize the difference by means of machine learning tools.

2.2 Preprocessing

Data preprocessing helps in enhancing the quality of data to promote the extraction of useful attributes from the dataset. It is the technique for preparing the raw data to make it suitable for building and training machine learning models [9]. In simple words, data preprocessing in machine learning is a data mining technique that transforms raw data into an understandable and readable format. Removal of different kinds of artifacts from the ECG signal is the main objective of the pre-processing [10]. The three important stages of this preprocessing are importing all the crucial libraries, importing the dataset and handling the attributes. The data set consists of 279 attributes and out of these only few are used to train the model and the rest contains the general information of the patient’s details. Among these few data will be definitely impossible and this error will be due to some technical faults and so it is corrected to some random value based on the possibility for example heartrate of a patient cannot be 0 and above 200 so it is randomly corrected to some normal value.

2.3 Feature Extraction

A feature is a particular column of the dataset such as age, height, QRS duration etc. Feature selection is the process of selecting a subset of the original feature set, which efficiently describes the input data among those features [5]. Therefore, the focus of feature selection is to find appropriate algorithms to evaluate the results which decrease the dimension of features to improve the model generalization ability and reduce the overfitting. This selection process is independent of the training process. First, the age column is selected to perform the feature extraction. The count of age feature is calculated using python count function and to make it simple sort index function is used to view the counts in particular order. The zero-age record is six hundred and ten centimeters(610cm) which is not possible so it should change to sixty-one centimeters(61cm). Then, the height column is selected with age equal to 1 where the height is seven hundred and ten centimeters(710cm). It is seventy-one centimeters(71cm).

Table 1 Number of Missing Values for each column

Column Name	No.of. missing values
T	8
P	22
QRST	1
J	375
Heart Rate	1

Now, calculate the missing columns and remove those columns to clean the dataset (V. Rathika rani et al., 2014). Then, replace the one or two missing values with the median of the corresponding feature.

2.4 Classification

Heart disease is classified into sixteen different types so that this machine learning model can make accurate classifications of heart disease (Sandra Smigiel et al., 2021). For each heart disease, the class number is assigned.

Table 2 Classification of Heart Diseases for linear regression and ANFIS

Class No	Disease Name
01	Normal
02	Ischemic changes (coronary artery disease)
03	Old Anterior Myocardial Infarction
04	Old Inferior Myocardial Infarction
05	Sinus tachycardia
06	Sinus bradycardia
07	Ventricular Premature Contraction (PVC)
08	Supraventricular Premature Contraction
09	Left bundle branch block
10	Right bundle branch block
11	1. degree Atrioventricular block
12	2. degree AV block
13	3. degree AV block
14	Left ventricular hypertrophy
15	Atrial Fibrillation or Flutter
16	Others

In the case of Random Forest Classifier, the heart disease is classified into two different types so that this machine learning model can make predictions based on Risk Factor.

Table 3 Binary Classification of Heart diseases for random classifier

Class No	**Risk Factor**
01	Normal
02	Risk

III. SYSTEM DESIGN

3.1 ANFIS Modeling

The automated heart disease diagnosis system is implemented using MATLAB, which is a powerful language for data analysis and visualization. Even though there are many techniques, the reasons for choosing MATLAB as a data mining tool for this paper is portability that the users will have the same range of basic functions at their disposal and the domain specific representations that points out in MATLAB implementation, all data is the form of matrices [11]. The neural networks concepts have got multidimensional data which is hard for the humans to understand, MATLAB makes this easy to deal with by 3D graphs and plot. MATLAB is preferred because of its efficiency in data calculation and visual graphic representation [12]. ANFIS uses various algorithms. Here the hybrid algorithm is used. The least square method and the gradient descent method are combined to solve the problem of backward pass. To tune the parameters the least square method is used. Once these parameters are obtained immediately the backward pass starts. The gradient descent algorithm is used to adjust the antecedent parameters. The output error is used to adapt the antecedent parameter by means of back propagation method. The hybrid method is highly efficient in training of ANFIS [13]. When the raw dataset is uploaded into the training model as the train data and test data, the values are plotted in the plotting area. The fis model is generated with the six major input parameters and the three membership functions. The ANFIS structure can be viewed from the view tab. Then the model is trained using the train data section with the number of epochs during the training the input values undergoes all the five layers and then tested against the test data section. The RMSE value Obtained is 0.4 and the model is obtained with great accuracy compared to other machine learning algorithms.

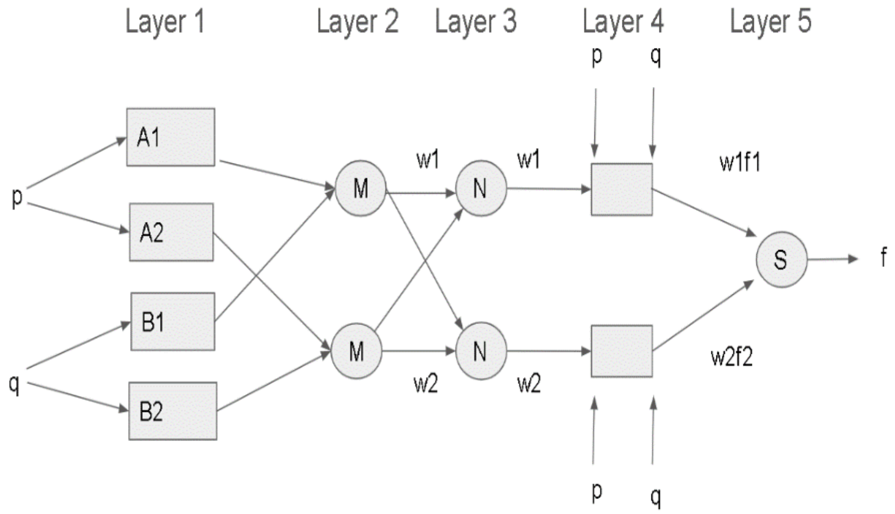


Figure 3 Structure of ANFIS

3.2 Regression Modeling

Linear regression is used to make predictions of a variable based on the value of another variable.The variable needed to predict is called the dependent variable and the variable used to predict the other variable's value is called the independent variable. It is the process of finding a line that best fits the data points available on the plot, so that it can predict output values for inputs. Performance and error rates depend on various factors including how to clean and make the data consistent. There are different ways of improving the performance of the model. However, each one has its own pros and cons, which makes the choice of methods application-dependent. These are used to map numeric inputs to numeric outputs and also fitting a line into the data points. Linear regression is classified into two types as simple linear regression and multiple linear regression.

Let us consider the model of dependent variable y and the independent variable x_i(i=1,2,3……) which will influence the variable y and predict the development of y (Shen Rong and Zhang Bao-wen, 2018).

Simple linear regression model expressed as followed:

$y = a_0 + a_1 x + e$

Where,
y - dependent variable
x - independent variable.
a_0 - constant term
a_1 - regression coefficient
e - random error

To implement the linear regression algorithm for ECG dataset, Python Programming Language is used to make heart disease prediction. Python is used successfully in real-world business applications, large, small and mission critical systems. Python is chosen for this project because it is a well-supported scripting language that extends the core code. Indeed, writing much more code in python than expected, including all in-game screens and the main interface. It was a huge success for the project because writing code in a garbage collection language simply goes faster than C++ coding. Here, the step-by-step process of creating the linear regression model is explained. The required python libraries were imported to the programming. Pandas in python used to do data analysis, data manipulation, numerical tables, time series and data structure operation. NumPy supports matrices, multi-dimensional arrays, and the operation of a large collection of high-level mathematical functions. Matplotlib is a numerical mathematics extension of NumPy. Load the dataset which contains 451 records with the help of drive mounting. By using the values. Count () find the number of records in each type of class. Then find the occurrences of non-normal data. Feature extraction is performed in the age and height column that zero-year-old refers to the child, so its height is changed from 610cm to 61cm. Likewise, one year old also refers to the child, so its height is changed from 780cm to 78cm. Some columns have all values as null values that are zero values that should be dropped. Histogram diagram is viewed for the remaining columns. One or several groups of variables are distributed in the historical format in which each bin is represented as a bar.

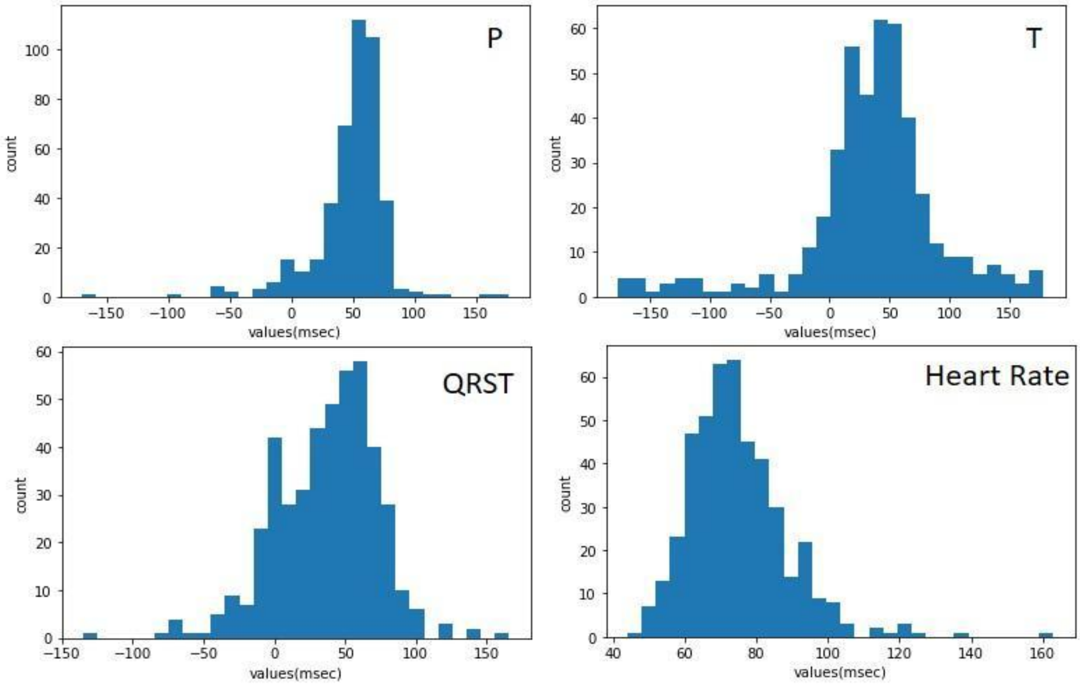


Figure 4 Histogram Diagram of a ECG Signal data - P, T, QRST, Heartrate

Simple Imputer library is imported to replace a value which is not defined (Not a Number - NaN) values with the median of the remaining values present in the column. Once again make sure that there are no missing values. Train_test_split library is imported to split up the whole dataset into training dataset and testing dataset file. To Perform Linear Regression Algorithm, import the Linear Regression library to train the model. Once training is completed start the prediction of the model. Scatter plots are used to observe and represent relationships between variables using dots. The scatter diagram is drawn to view the visual representation of the data. Data frame is created to view the Actual Vs Predicted data and its difference. Finally, accuracy of the linear regression model is calculated as its accuracy is very low compared to other machine learning models for this ECG classification data to predict sixteen types of heart diseases.

3.3 Random Forest Classifier

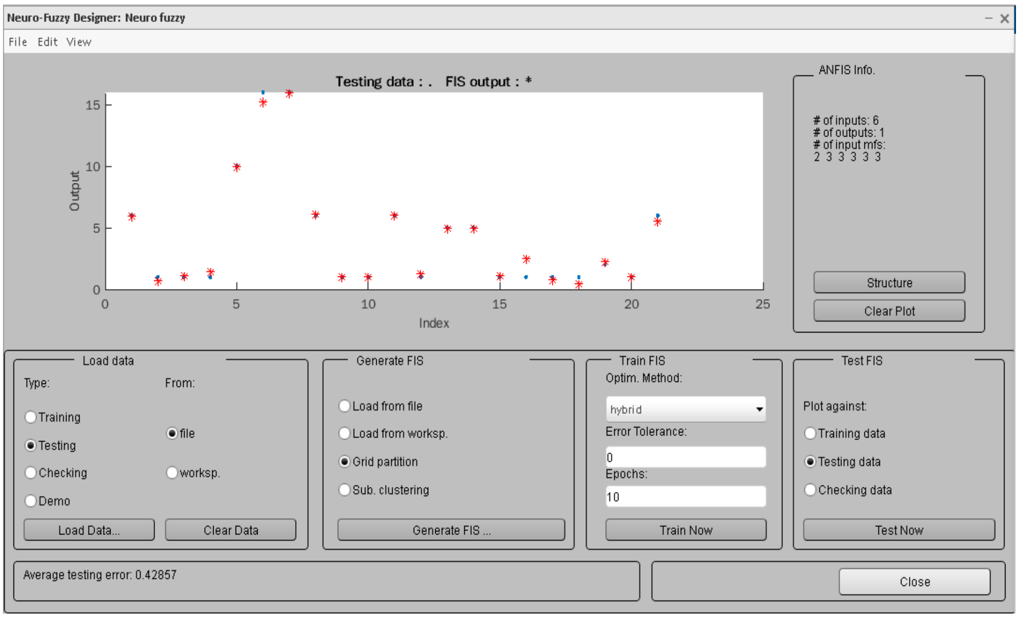
Binary classification is used to categorize data points into one of two buckets: Yes or No,0 or 1, true or false using a random forest classifier. It is a type of supervised learning. Supervised learning is the type of machine learning in which machines are trained using labeled data in which input data is already tagged with proper output. Random forest classifier makes a set of decision trees from randomly selected subsets of the training set that is widely used in Classification [14] and Regression problems [15]. Each tree in the forest is built by random selection of sample data. The Random Forest classifier library should be imported, and the confusion matrix is calculated with accuracy as 78.68%. MinMaxScaler scales all the data features in the range of [0, 1] or [-1, 1]. MinMaxScaler library is imported to normalize the data so that the accuracy is raised to 83.09%. Some Random record in the dataset was given as input to the random classifier model and the corresponding output class is displayed.

Table 4 Random Classifier Output

S. No	age	sex	height	weight	QRS duration	PR interval	Q-T interval	Pinterval	heart rate	Risk Factor
1	56	1	165	64	81	174	401	39	53	Risk
2	55	0	175	94	100	202	380	143	71	Normal
3	30	0	170	73	91	180	355	104	56	Risk
4	48	0	178	80	91	224	331	67	104	Risk
5	40	1	160	52	77	129	377	66	70	Normal

IV. Results and Discussion

This paper implements the method called Adaptive Neuro Fuzzy Inference System which uses fuzzy logic and the neural network techniques along with some other popular machine learning algorithms to analyze the ECG signal and to classify the various arrhythmias. The dataset used is from MIT-BIH arrhythmia database directory for. The dataset is splitted into two: one for training and another for testing. The testing classification accuracy in MATLAB is achieved around 96%. The entire necessary algorithm is implemented on MATLAB for ANFIS and also in Python for other Machine learning techniques. From the comparison chart it is clear that linear regression has low accuracy, Random Classifier has accuracy 83% and ANFIS has high accuracy 96%. Among all the three the ANFIS has achieved the model with greater accuracy.



ANFIS Info.

# of inputs: 6
# of outputs: 1
# of input mfs:
2 3 3 3 3 3

Type:
○ Training
● Testing
○ Checking
○ Demo

From:
● file
○ worksp.

○ Load from file
○ Load from worksp.
● Grid partition
○ Sub. clustering

Optim. Method:
hybrid
Error Tolerance:
0
Epochs:
10

Plot against:
○ Training data
● Testing data
○ Checking data

Average testing error: 0.42857

Figure 5 ANFIS Result obtained in MATLAB

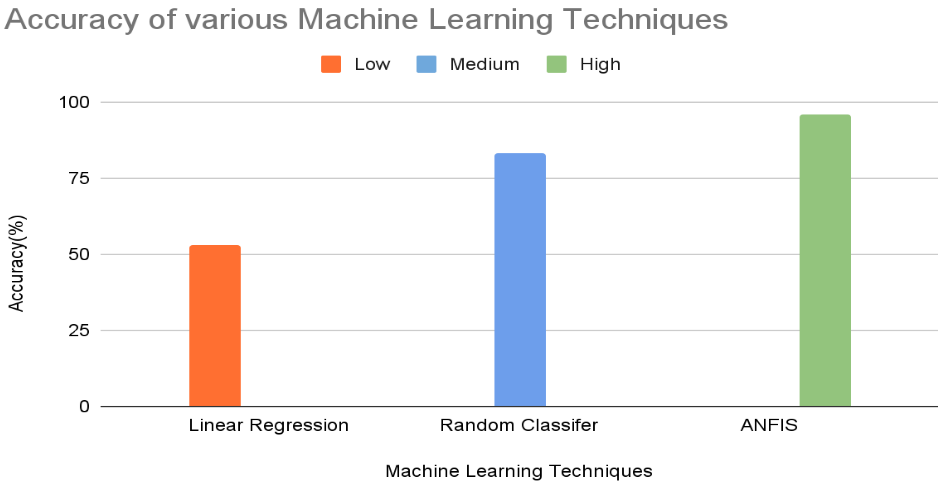


Figure 6 Comparison Chart

V. Conclusions

The application of the ANFIS classifiers is very good in signal analysis problems. Signal processing-based approach is applied for automated diseases classification based on biomedical signal processing in this work. ANFIS is a combination of fuzzy inference system and neural network because of these it gets adaption quality, sensitivity, and smoothness. The ANFIS classifier can efficiently support in recognizing different types of arrhythmias with an average accuracy as high as 96%.

Acknowledgements

We are immensely grateful to the Advanced Manufacturing Technology Development Center (AMTDC) and the sponsor Ministry of Heavy Industries (MHI) for providing us this Internship and we are thankful for your continuous support provided till the completion of our project.

References

[1]. World health organization (2022), ‘cardiovascular diseases (CVDs)’, https://www.who.int/health-topics/Cardiovascular-diseases#tab=tab_1

[2]. Mayoclinic (2022), ‘Overview of Heart diseases’ [online], https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118.

[3]. Muhammad Wasimuddin Khaled Elleithy, Abdelshakour Abuzneid, Miad Faezipour and Omar Abuzaghleh (2021) ‘Multiclass ECG Signal Analysis Using Global Average-Based 2-D Convolutional Neural Network Modeling’, *Journal of Multidisciplinary Digital Publishing Institute,* Electronics, https://doi.org/10.3390/electronics10020170

[4]. Markus Hoglinger (2016), ‘ECG Preprocessing”, Institute of Signal Processing’, *Johannes university linz,* DVR 0093696.

[5]. Liping Xie, Zilong Li, Yihan Zhou, Yiliu He, and Jiaxin Zhu (2020), ‘Computational Diagnostic Techniques for Electrocardiogram Signal Analysis’, *Multidisciplinary Digital Publishing Institute*, Sensors (Basel) v.20(21), https://doi.org/10.3390/s20216318.

[6]. Saira Aziz, Sajid Ahmed and Mohamed-Slim Alouini (2021),‘ECG-based machine-learning algorithms for heartbeat classification’, *Scientific Reports*, https://doi.org/10.1038/s41598-021-97118-5.

[7]. Peiman Shahbeigi-Roodposhti and Sina Shahbazmohamadi (2022), ‘Acquisition and Analysis of an ECG Signal’, Biomedical Engineering Department, *University of Connecticut,* Storrs. Connecticut [online], https://www.jove.com/v/10473/acquisition-and-analysis-of-an-ecg-Electrocardiography-signal.

[8]. Prarthana B. Sakhare and Rajesh Ghongade (2015), ‘An Approach for ECG Beats Classification using Adaptive Neuro Fuzzy Inference System’, *IEEEINDICON* https://doi.org/10.1109/INDICON.2015.7443804, ISSN: 2325-9418.

[9]. S Celin and K. Vasanth (2018) ,’ECG Signal Classification Using Various Machine Learning Techniques’, *Journal of Medical Systems* vol 42, No 241, https://doi.org/10.1007/s10916-018-1083-6.

[10]. Javatpoint (2021), ‘Data Preprocessing in Machine Learning’ [online], https://www.javatpoint.com/data-preprocessing-machine-learning.

[11]. Taiseer Mohammed Siddig and Mohmmed Ahmed Mohmmed (2014), ‘A Study of ECG Signal Classification using Fuzzy Logic Control’, *International Journal of Science and Research (IJSR)* Volume 3 Issue 2,, ISSN (Online): 2319-7064, Paper ID: 02013989.

[12]. Mohammad A. M. Abushariah , Assal A. M. Alqudah, Omar Y. Adwan and Rana M. M. Yousef (2014), ‘Automatic Heart Disease Diagnosis System Based on Artificial Neural Network and Adaptive Neuro-Fuzzy Inference Systems Approaches’, *Journal of Software Engineering and Applications,* Vol.07 No.12, https://doi.org/10.4236/jsea.2014.712093.

[13]. Ali Gharaviri, Mohammad Teshnehlab and H. A. Moghaddam (2008), ‘Ischemia Detection via ECG Using ANFIS’, *Annual International Conference of IEEE Engineering in Medicine and Biology Society, https://doi.org/10.1109/IEMBS.2008.4649368*

[14]. Adam Gacek (2012), ‘An Introduction to ECG Signal Processing and Analysis’, *ECG Signal Processing, Classification and Interpretation: A Comprehensive Framework of Computational Intelligence pp 21–46*, https://doi.org/10.1007/978-0-85729-868-3.

[15]. V. Mahesh, A. Kandaswamy, C. Vimal and B. Sathish (2010), ‘Random Forest Classifier Based ECG Arrhythmia Classification’ ,*International Journal of Healthcare Information Systems and Informatics*, 5(2), 1-10, https://doi.org/10.4018/jhisi.2010040101.