JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A MACHINE LEARNING APPROACH FOR HEALTHCARE IN IDENTIFYING DISEASES

[1]Mr. N.S.R.K. Prasad, [2]Mr.K. Anil Kumar

[1]Assistant Professor, [2]Assistant Professor
[1]Department of Information Technology, [2]Department of Information Technology
[1]Guru Nanak Institutions Technical Campus, Hyderabad, India

ABSTRACT

Now-a-days Healthcare is one of the most important industries in the world offering many advanced Value-Based Healthcare (VBHC) approaches for designing and managing healthcare systems. Machine Learning (ML) approaches in healthcare gives many powerful solutions for efficient diseases forecast, diagnosis, and treatments, improving the overall operations of healthcare. ML uses different strategists to deal with huge amount of healthcare data in order to provide improved healthcare services at lesser costs and enhance patient satisfaction. Heart disease is most severe diseases in the world and many humans are getting affected irrespective of their age. In India and in the world over the last few decades the death rate due to heart related diseases have increased more in number. Diagnosing of heart disease is a major issue observed in many human beings due to their change in life style and the disease need to be identified in advance So, there is a demand of consistent, perfect and realistic process to diagnose heart related diseases in time for appropriate treatment. Machine Learning algorithms and techniques can be applied to different medical datasets to mechanize the analysis of complex health care patient's data. An efficient Machine learning algorithm finds patterns and grounds regarding data. In this paper we presented a machine learning model to identify the patient's health status; monitors health, and suggests necessary steps to be taken in order to prevent heart disease. We considered supervised learning algorithms such as, Gaussian Naive Bayes, Decision Trees (DT), Support Vector Machines (SVM), Linear SVC, and Random Forest (RF) and classified the heart disease and data sets are downloaded from UCI repository site.

Keywords: Gaussian Naive Bayes, Decision Trees (DT), Support Vector Machines (SVM), Linear SVC and Random Forest (RF).

1. INTRODUCTION

In this 21st century of data driven era, health care industry is generating a vast amount of structure, unstructured and semi structured data from patients during their clinical examination, treatment report has become weighty to organize. Due to this awkward nature of patient's data, the decision-making process and treatment is getting delayed. Life-threatening diseases like breast cancer, heart disease, and liver disease are very crucial in medical science. The diagnosing process for these diseases can be done more accurately with the help of machine learning techniques. For effective diagnosing of heart diseases, a computer-based machine learning decision support system will play an important role. Machine learning algorithms are considered into three categories – Supervised, Unsupervised and Reinforcement learning algorithms. Supervised machine learning algorithms uses label data consists of input values and a targeted output value as training data for analysis. Unsupervised machine learning techniques are used to identify hidden patterns from unlabeled data sets. Reinforcement learning approach uses a machine to learn its actions from the feedback received during

the interactions with the external situation. Supervised and unsupervised learning techniques are generally used for data analysis and reinforcement techniques are used for decision making. This paper is mainly focused on various supervised machine learning algorithms so called Support Vector Machine (SVM)[3][4][5], Decision tree[6] Random Forest[7][8][16] and Logistic Regression[9] used in medical diagnosis. we have considered the data sets from UCI for predicting the heart disease from historic medical records. The dataset consists of 583 patients with 10 clinical feature attributes to train different machine learning algorithms. This paper is organized as follows: Section 2 gives the Literature survey to ML approaches to cardiovascular diseases, section 3 describes the work related to Machine Learning algorithms for healthcare, Section 4 describes the proposed work for ML models for identifying heart diseases and diagnosis, section 5 describes about the data sets for diagnosing the heart diseases, section 6 describes the performance measures and section 7 presents results.

2. LITERATURE SURVEY

Machine learning (ML) has been used in many numbers of applications in the healthcare industry. Today, in the 21st century of data driven era ML is used to modernize all the administrative processes in healthcare hospitals. Heart diseases are leading causes to demise among all other diseases, even cancer. Diagnosis process is getting failed due to improper resources and many men & women are facing heart problems and leading to deaths every year.[1]

Hanen Bouali et.al.[3][4][5] contributed a study of different classification techniques on heart diseases usecase. The author applied support vector machine for classification and regression problems for predicting the heart diseases of humans at an early stage.
Seyedamin Pouriyeh, et al [4] in the research paper named a comprehensive investigation and comparison of machine learning techniques in the domain of heart disease have proposed decision trees to solve regression and classification problems.

Shan Xu et al[7] in the research paper named cardiovascular risk prediction method based on CFS subset evaluation and random forest classification framework, have proposed for solving the classification and regression problems for heat diseases by creating multiple Decision Trees on randomly selected data samples.

Manpreet Singh, Levi Monteiro Martins, Patrick Joanis and VijayK. Mago et Al [9]. in the research paper named building a cardiovascular disease predictive model using structural equation model & fuzzy cognitive map they have proposed ML for solving various problems such as disease prediction, cancer detection.

Amandeep Kaur et. Al, compared various algorithms such as artificial neural network, K-nearest neighbor, naïve bayes, Support vector machine on heart disease prediction.
J Thomas, R Theresa Princy., et Al proposed the use of K nearest neighbor algorithm, neural network, naïve bayes and decision tree for heart disease prediction. They used data mining techniques to detect the heart disease risk rate.

Monika Gandhi et. Al, used naïve bayes, decision tree and neural network algorithms and analyzed the medical dataset

Ramandeep Kaur, Er.Prabhsharn Kaur et. Al, [19] have showed that the heart disease data contains unnecessary, duplicate information. Sonam Nikhar et. Al, has built up the paper titled as prediction of heart disease using machine learning algorithms using decision tree classifier and naïve bayes.
Mr. Santhana Krishnan. J and Dr. Geetha. S.et. Al, has written paper that predicts heart disease for male patient using classification techniques.

Utilization of Machine Learning techniques in this regard will be highly beneficial to heart patients. According to the 2016 survey, global burden of disease, in India heart disease killed nearly 1.7 million Indians in 2016. Heart-related diseases increase health-care costs and can reduce an individual's productivity. The World Health Organization (WHO) survey says from 2005 to 2015, India has lost up to $237 billion due to infectious or cardiovascular diseases [2].

2.A. Identifying the cardiovascular diseases:

Coronary artery disease occurs in a situation when the blood supply to the heart muscle is partially or completely blocked.

Coronary artery disease is due to the gradual accumulation of cholesterol and other fatty materials in the wall of a coronary artery. This can be treated as atherosclerosis and can affect many arteries, not just those of the heart.

2.B Symptoms of a heart attack can include:

discomfort, pressure, heaviness, or pain in the chest, arm, or below the breastbone, discomfort radiating to the back, jaw, throat, or arm, fullness, indigestion, or choking feeling (may feel like heartburn) sweating, nausea, vomiting, or dizziness, extreme weakness, anxiety, or shortness of breath, rapid or irregular heartbeats.

Heart Attack Warnings	Heart Diseases
Chest pain	Coronary artery disease (CAD).
Shortness of breath.	Angina pectoris.
Sweating and Fatigue.	Congestive heart failure.
Nausea	Cardiomyopathy.
Indigestion	Congenital heart disease

Table 1: Common Heat Attack warnings and heart diseases

2.C Distribution of CVD deaths due to heart attacks, strokes, and other cardiovascular diseases.

According to global atlas on cardiovascular diseases prevention and control the death chart is showing the impact heart diseases on humans as shown in fig.1

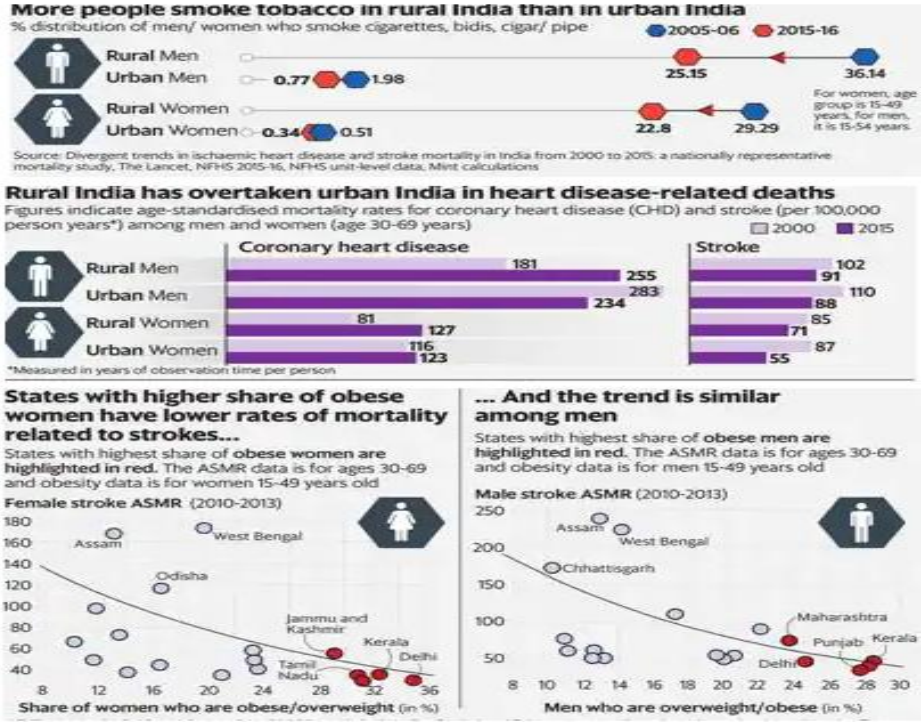

Fig 1. Death Rate in INDIA due to CVD

3. Machine Learning approach for Health Care CVD Diseases.

Machine learning process focuses on developing algorithms based on the machines past experiences. ML Algorithm are used to detect pattern in the input data and builds a model based on input data to make precise predictions for new data.[10]

3.A. Traditional programming approach

Fig 2. Traditional programming approach

3.C.Machine learning ML approaches

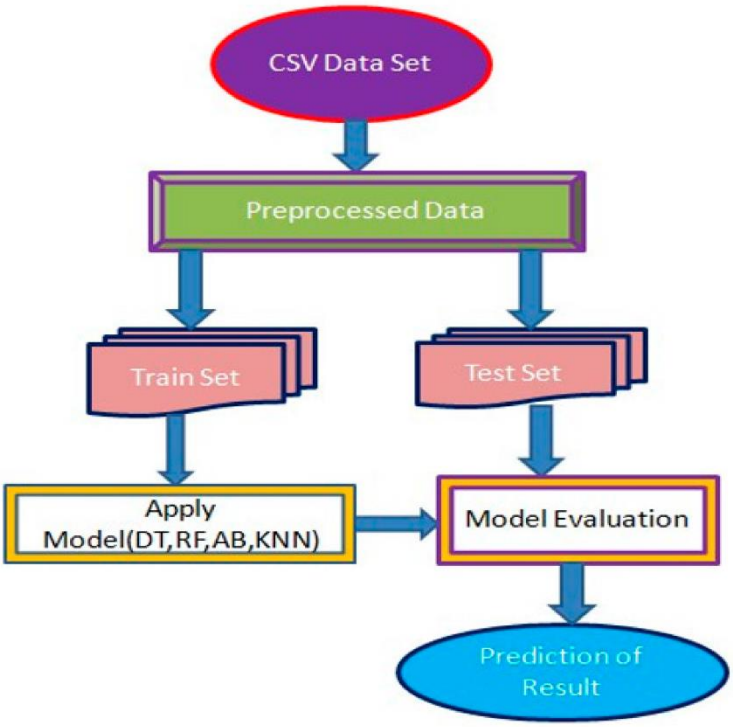

Fig 3. Machine Learning approach

3.D Architecture of ML Model

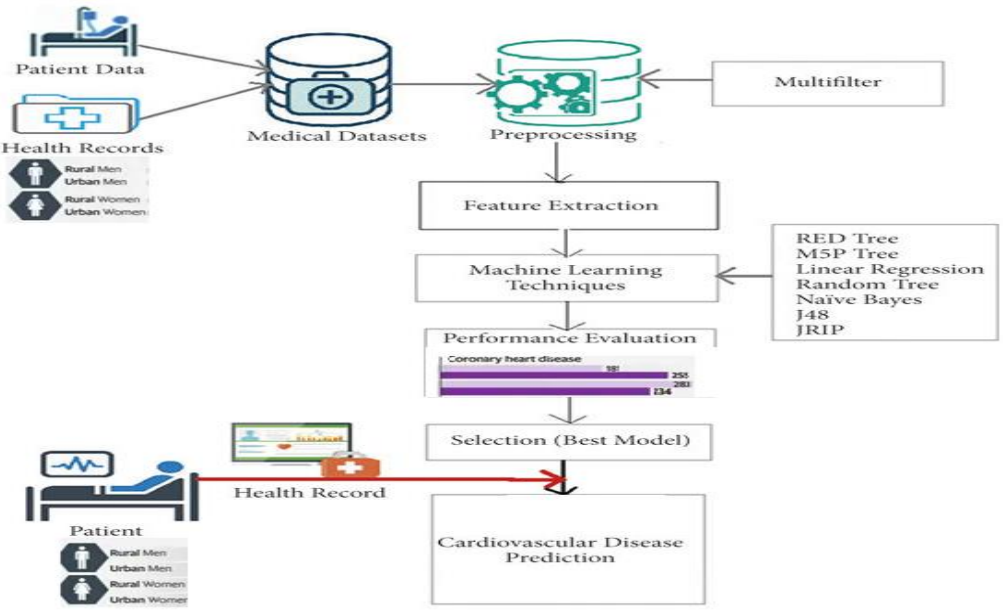

Fig 4. ML Model for CVD Diseases

3.E. Classifications of ML algorithms

Supervised learning algorithms use labeled data. In this paper we have used most popular supervised algorithms for classifying data in health care for diagnosing the heart diseases.[11]

The machine learning techniques can be classified as follows:

1) Supervised
2) Unsupervised [15]

Supervised learning labels for the training data is provided and /or select features to feed the algorithm to learn

Unsupervised learning algorithm is applied on raw data and learns fully automatic.[12]

Algorithm:

Input: heart disease_Input Features

Assign training and testing dataset for heart disease Output: Classification of heart disease as a result of the output

Function: Support_Vector_Machine(Input features F, Label vector V=[1.....n])

Step 1: Decide on the optimal cost and gamma value.

Step 2: Perform while (conditioning)

Step 3: For each set number of input file features, run the training

step. Step 4: Run the classification step for a set number of features in the input file.

Step 5: Come to an end whilst Step

6: Submit the heart disease classification results.

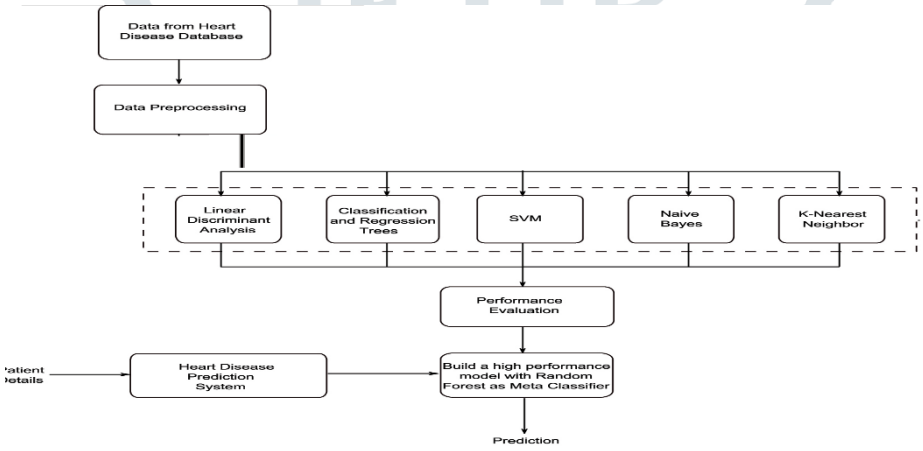

Fig 5. ML Model architecture for identifying CVD heart diseases.

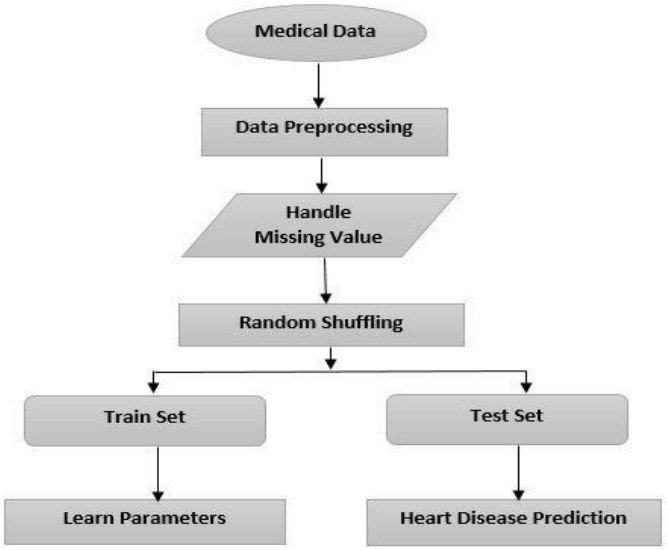

Fig 6: Machine learning classification.

Pre-Processing steps:

1: Import the dataset required: date set is collected from UCI repository Involve gathering of medical information artifacts from several sources like hospitals, discharge slips of patients and from UCI repository

2: Taking consideration of absent data in dataset:

It will remove all the unnecessary data and extract important features from data.
3**:** Encoding categorized data
Splitting Dataset: split the Data set into two parts
1. Training dataset and
2. Testing dataset respectively.
Splitting the dataset enables them to build best machine learning model.
Training data: Model is trained on the dataset of diseases to do the prediction accurately and produce Accuracy.

4. PROPOSED WORK

4.A. Machine Learning Algorithms used for identifying CVD heart diseases

For predicting the heart disease, dataset is downloaded from UCI repository. On applying the ML model on patient's data, ML classifiers predicts the patients disease status and gives a accurate information. The proposed system, uses classification algorithms Support Vector Machine, Gaussian Radial Basis Function (RBF), Decision Tree, Random Forest, logistic Regression algorithms.

4.B Support Vector Machine (SVM)

SVM training procedure develops a strategy that can identify whether an input image belongs to this class or not. To find an accurate decision border, SVM requires a considerable amount of training data, which increases the computational cost. The SVM is a learning algorithm for categorization. It aims to find the optimum separation hyperplane for unobserved sequences with the lowest possible predicted classification error. For linearly non-separable data, the input is transported to a high-dimensional feature set where they can be distinguished by a hyperplane. To achieve this projection onto a high-dimensional feature space, kernels are required.

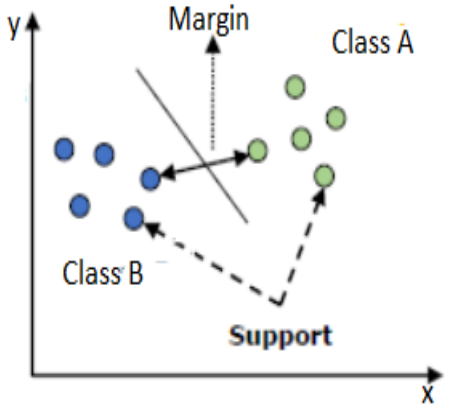

Fig7. Support Vector Machines

Hyper plane –It is a selection plane or space which is isolated between a lot of articles having various classes Support Vector: Data indicating nearer the hyper plane are called support vectors.

4.C Gaussian Radial Basis Function (RBF)

Using this technique to classify our problem we got the accuracy of 68.85% using all the features in Heart disease UCI dataset.

4.D. Linear SVM - For Large data sets Linear SVM is the latest incredibly quick machine learning algorithm to solve classification problems from ultra- large data sets [9]. Using this technique to classify our problem we got the accuracy of 81.97% using all the features in Heart Disease UCI dataset.

4.E. Decision Tree

It is a graph-like flowchart where each internal node indicates a check an element, branch reflects a outcome, and leaf node (terminal node) carries a class name. Decision Tree algorithm measures each and every attribute's entropy first.

The data set is broken with the help of variables to maximum gain of knowledge and minimal entropy.

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

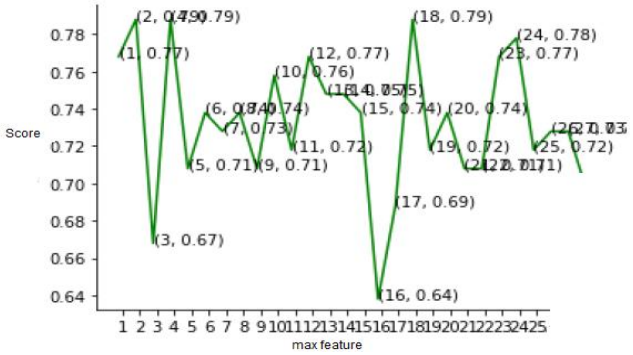

Fig 8. Accuracy Line graph

From the obtained line graph, it has been observed that the maximum accuracy is 79% and is obtained by number of maximum features (2, 4, 18).

Using this technique to classify our problem we got the accuracy of 79% using all the features in Heart Disease UCI dataset.

4.F. Random Forest

It is a ML supervised algorithm used for classification (discrete data) as well as regression (continuous data) techniques. This algorithm/technique builds multiple decision trees & hence mergers them together to get the most correct results along with the stable prediction.[10]

Ensemble Learning aggregates multiple Machine learning models for overall better performance of the system

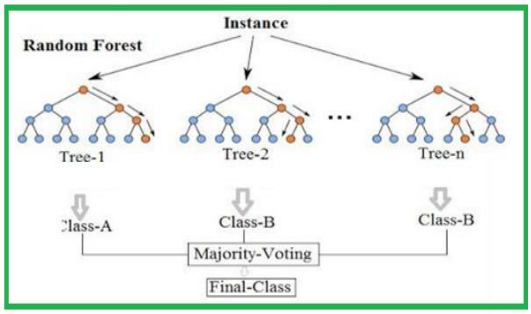

Fig 9. Random Forest algorithm result

Using this technique to classify our problem we got the accuracy of 80.33% using all the features in Heart Disease UCI dataset.

4.G. Logistic Regression: Logistic regression is used and the target variable is categorical. Logistic regression name comes from function used at computational level to compute probabilities;

$$P = \frac{1}{1 + e^{-(a+bX)}}$$

S-shaped curve that can take any number evaluated to a value between 0 and 1.

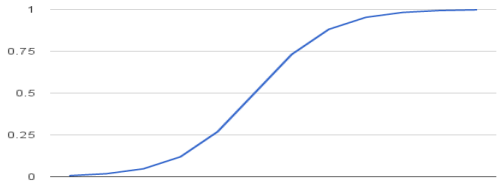

Fig 10. Logistic regression Curve

Using this technique to classify our problem we got the accuracy of 85.25% using all the features in Heart Disease UCI dataset [15].

5. PATIENT DATA SET

The complete of 1090 cases with ten attributes was amassed for the cardiovascular data set from koggle, UCI. The attribute "heart disease" described as the one indicates people having heart disease and Zero Indicates no heart disease.

Table1 suggests the attributes values of heart disease data set. The data set having 539 heart disease no cases and 551 heart disease- yes cases**.**

Table 2: Heart Disease Patient Data Set

Attribute	Remarks
ID	ID num
age	in Days
gender	1-women ,2-Men
height	In Cent Meter
weight	Kilo Grams
systolic blood pressure	Systolic Blood Pressure
Diastolic blood pressure	Diastolic Blood Pressure
Cholesterol	1-Normal ,2-Above Normal, 3-Well Above Normal
glue	1-Normal ,2-Above Normal, 3-Well Above Normal
smoke	Whether Patient Smokes or Not
alco	Binary Feature
active	Binary Feature
cardiovascular	Target Variable

6. PERFORMANCE MEASURES

		Yes	No	
Actual Class	Yes	True Positives (TP)	False Negatives (FN)	P
	No	False Positives (FP)	True Negatives (TN)	N
		P Complement	N Complement	P+N

Table 3: Components of Confusion Matrix Predicted Class

7. RESULTS

Algorithm	Accuracy
Gaussian radial basis function (RBF)	68.85%
Linear SVM	81.97%
Decision Tree	79.00%
Random Forest	80.33%
Logistic Regression	85.25%

Table 4. Comparison table with accuracy values

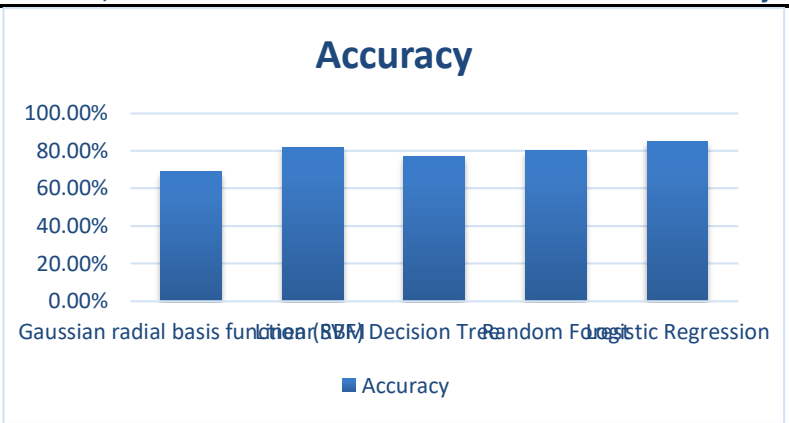

Fig 11. Accuracy of ML algorithms

8. CONCLUSION

Our analysis concludes that predicting cardiovascular diseases offer considerable scope for machine learning algorithms. Gaussian Naïve Bayes and Logistic Regression, performed exceptionally well with 85.25% of accuracy whereas Decision trees performed very poorly with mere accuracy of 79.00%. Random Forest (80.33%) and SVM (81.97%) are performed moderately well, as it overcome the predicament of over fitting by using multiple algorithms (Random Forest multiple decision trees). Naïve Bayes classifier are very efficient in computational terms. Machine learning algorithms and techniques are very accurate in heart-related predictions.

References

[1] Richard C. Becker, MD CDC:, healio.com-cardiology heart disease cancer leading causes of death in 2017
[2] Global Atlas on Cardiovascular Disease Prevention and Control. Geneva, Switzerland: World Health Organization; 2011
[3] Hanen Bouali and Jalel Akaichi et al. "Comparative study of Different classification techniques, heart Diseases use Case.", 2014 13th International Conference on Machine Learning and Applications.
[4] Seyedamin Pouriyeh, Sara Vahid, Giovanna Sannino, Giuseppe De Pietro, Hamid Arabnia, Juan Gutierrez et al. " A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease", 22nd IEEE Symposium on Computers and Communication (ISCC 2017): Workshops - ICTS4eHealth 2017
[5] Houda Mezrigui, Foued Theljani and Kaouther Laabidi et al. "Decision Support System for Medical Diagnosis Using a Kernel-Based Approach", ICCAD'17, Hammamet - Tunisia, January 19-21, 2017.
[6] Simge EKIZ and Pakize Erdogmus et al. "Comparitive Study of heart Disease Classification", 978-1-5386-0440-3/17/$31.00 ©2017 IEEE
[7] Shan Xu ,Tiangang Zhu, Zhen Zang, Daoxian Wang, Junfeng Hu and Xiaohui Duan et al. "Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework", 2017 IEEE 2nd International Conference on Big Data Analysis.
[8] Quazi Abidur Rahman, Larisa G. Tereshchenko, Matthew Kongkatong, Theodore Abraham, M. Roselle Abraham, and Hagit Shatkay et al. "Utilizing ECG-based Heartbeat Classification for Hypertrophic Cardiomyopathy Identification", DOI 10.1109/TNB.2015.2426213, IEEE Transactions on Nano Bioscience TNB-00035-2015.
[9] Manpreet Singh, Levi Monteiro Martins, Patrick Joanis and VijayK. Mago et al. "Building a Cardiovascular Disease Predictive Model using Structural Equation Model & Fuzzy Cognitive Map", 978-1-5090-0626-7/16/$31.00 c 2016 IEEE.
[10] Peter, "Enhancing random forest implementation in WEKA", in: Machine Learning Conference, 2005.
[11] Types of Machine Learning Algorithms, Taiwo Oladipupo Ayodele, University of Portsmouth, United Kingdom.
[12] Dhomse Kanchan B and Mahale Kishor M. et al. "Study of Machine Learning Algorithms for SpecialDisease Prediction using Principal of Component Analysis", 2016 International Conference on GlobalTrends in Signal Processing, Information Computing and Communication.
[13] R.Kavitha and E.Kannan et al. "An Efficient Framework for Heart Disease Classification using Feature Extraction and Feature Selection Technique in Data Mining ", 2016.
[14] Shan Xu ,Tiangang Zhu, Zhen Zang, Daoxian Wang, Junfeng Hu and Xiaohui Duan et al."Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework", 2017 IEEE 2nd International Conference on Big Data Analysis
[15] [15]Hanen Bouali and Jalel Akaichi et al. "Comparative study of Different classification techniques, heart Diseases use Case.", 2014 13th International Conference on Machine Learning and Applications.
[16] Avinash Golande, Pavan Kumar T, (June 2019): Heart Disease Prediction Using Effective Machine Learning Techniques, International Journal of Recent Technology and Engineering (IRTE), ISN: 2277-3878, Volume-8, Issue-1S4.
[17] A. Sahaya Arthy, G.Murugeshwari, (April 2018): A survey on heart disease prediction using data mining techniques.
[18] Amita Malav, Kalyani Kadam, (2018): "A Hybrid Approach for Heart Disease Prediction Using Artificial Neural Network and K – Means", International Journal of Pure and Applied Mathematics. [19] Benjamin EJ et.al, (2018): Heart Disease and Stroke Statistics At-a-Glance.
[19] DhafarHamed, Jwan K.Alwan, Mohamed Ibrahim, Mohammad B.Naeem, (march – 2017): "The Utilization of Machine Learning Approaches for Medical Data Classification" in Annual Conference on New Trends in Information & Communications Technology Applications.
[20] Himanshu Sharma, M A Rizvi, (August 2017): Prediction of Heart Disease Using Machine Learning Algorithms: A Survey.