iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.45507

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

Heart Disease Prediction Using Supervised Algorithms in Machine Learning

Vidyasagar Doddamani[1], Dr. Ravindra S[2]
[1]*PG Student Department of ECE, Dayananda Sagar College of Engineering, Bengaluru, Karnataka, India*
[2]*Asst Professor Department of ECE, Dayananda Sagar College of Engineering, Bengaluru, Karnataka, India*

Abstract: ***There are various curable and incurable diseases which humans encounter in various stages of their life. Now-a-days due to poor nutrition and lifestyle modification, heart diseases are very prevalent. There are so many treatments available in medical field for heart diseases once predicted. However predicting heart disease is a challenging task. Therefore, predicting heart disease early helps people across the world to take the necessary actions before it reaches severe stage. From many years, machine learning approach has been used to deliver effective results in decision making and prediction of heart disease using different datasets available in the medical industry. In this project we use Standard Scaler technique for feature selection. In this project an attempt has been made to predict or detect the presence of heart disease using five most commonly used supervised machine learning algorithms that are, Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Logistic regression (LR) and the K-nearest neighbor (KNN) algorithms. Lastly the performance of these five supervised machine learning algorithms is summarized.***

Keywords: DecisionTree, K-Nearest Neighbor, Logistic Regression, Support Vector Machine, Random Forest.

I. INTRODUCTION

Heart disease is also called as cardiovascular disease, holds a range of cardiovascular conditions. From past few decades heart diseases has been a major cause of death worldwide. Cardiovascular disease is cased due to various underlying health conditions like affected or blocked arteries which lead to heart attack, chest pain, also called as angina or stroke. Other heart conditions, affecting the heart muscle, valve or rhythm, are also a form of heart disease. Symptoms of heart disease are difficult to detect hence we call heart attack as silent killer. Due to poor nutrition, excessive alcohol consumption, poor lifestyle, lack of exercises etc the rate of heart attack is increasing day by day. According to the statistics provided by World Health Organization (WHO), more than 17 million people die every year from heart diseases such as strokes & heart attacks. Heart disease can be classified as two types based on time of occurrence in one's lifetime. If the new born babies gets heart disease it is called congenital Heart Disease and if heart diseases comes at the later ages, then it is called Acquired Heart Disease.

There are number of tests like blood pressure, ECG, auscultation, cholesterol and blood sugar for diagnosis of heart disease. All the tests and procedures mentioned above take very long time for heart disease prediction. This delay in the prediction causes negative impact on the patient. To overcome this negative impact there is a need to find the right, reliable, and logical ways to make an early diagnosis which helps us to achieve quick treatment of the disease. To adopt quickest approach to detect heart disease is need of an hour for us. Hence using machine learning approach is good idea. With machine learning it is much easier to get information on big data which is impossible for humans to analyze.

Arthur Samuel first used the phrase "machine learning" in 1959. Machine learning is a field of study that offers computers the ability to learn without explicit programming. The process of developing algorithms based on prior inputs and experiences is known as machine learning.[12].

II. RELATED WORK

Farzana Tasnim et al. [1] have used KNN, DT, SVM, NN, LR, RF and Gradient Boosting data mining classification techniques to detect the coronary heart disease. The evaluation of this work is done using the dataset from the UCI machine learning repository. The feature selection method used to increase the performance of algorithms. Among the classification algorithms used in this project, Random Forest (RF) algorithm showed the best accuracy of 92.85% for heart disease prediction and classification.

Youness Khourdifi et al. [2] have compared the performances of different algorithms of machine learning. In this paper the supervised algorithms like K-NN, RF and ANN has delivered better results compared to others. The attempt has also been made combine the algorithms and to check the efficiency of merging algorithms. Overall the entire model resulted in better accuracy.

Youness Khourdifi et al. [3] have found the performances and efficiencies of available model using different datasets and algorithms. They have then evaluated the result in which the model showed better performance and effectiveness when K-Nearest Neighbour, Random Forest, Naïve Bayes, Support Vector Machine SVM & Artificial Neural Network algorithms were used.

Komal Saini et al. [4] have reviewed the IoT model for predicting the heart disease. The data of patient's body conditions were collected using wearable sensor device. The data collected is then stored and processed using communication standards, like Bluetooth (BLE) which is established to store large amounts of data generated by healthcare applications and wearable sensors as cloud storage. Lastly, data mining tools and machine learning algorithms were used to analyze data, predict the model and obtain accurate results.

Senthilkumar Mohan et al. [5] Proposed this paper where machine learning techniques were used in which raw data is processed first. This processed data helped them to identify new parameters related with heart disease. They have concluded that it is highly important to extend this study further by focusing the study on the actual dataset, and not just the theoretical approach and simulation. The hybrid HRFLM model which is combination of the capabilities of RF and LM gave good accuracy compared to models run using individual algorithms.

C. Gazeloglu et al [6] have used machine learning algorithms like DT & NB algorithm for detection of heart disease. In first algorithm the True or False decisions were obtained by decision tree algorithm to find out presence of heart diseases. The results from other algorithms like SVM, KNN are obtained on vertical or horizontal split conditions depending on dependent variables. The author has used Cleveland data set to perform the analysis. The dataset is first split into 70% training and 30% testing data. The algorithm NB gave 91% accuracy. It is concluded that the model can handle complicated, nonlinear, dependent data.

III.DATA SOURCE

According to the WHO, the world highest number of deaths in middle age is due to cardiovascular disease. Hence as we discussed above we use rapid method to predict heart disease using machine learning approach. Machine learning approaches need data of patients to learn the test the given data. In this project, the organized personal data sets of patients with history of heart problems were collected from UCI repository.

We have used data set that has data of 5111 different patients of different age groups, gender and different health conditions. This data set consists of 12 important attributes like age, resting blood pressure, and the related medical attributes such as bmi, fasting blood sugar level, etc. This dataset containing 12 medical attributes/parameters of 5111 patients is used to determine if a patient is at risk for Cardio Vascular Diseases i.e, CVD. It also helps us to classify the individuals who are at risk and those who are not at risk.

TABLE I
ATTRIBUTES USED

SL.No	Observations	Description
1	ID	Unique identifier
2	Gender	Male, Female, Other
3	Age	Age of the patient
4	Hypertension	0 - doesn't have hypertension, 1 - has hypertension
5	Heart_disease	0 - doesn't have heart diseases, 1 - has a heart disease
6	Ever_married	No or Yes
7	Work_type	Children, Govt_job, Never_worked, Private or Self-employed
8	Residence_type	Rural or Urban
9	avg_glucose_level	average glucose level in blood
10	bmi	body mass index
11	smoking_status	formerly smoked, never smoked, smokes or Unknown
12	stroke	1 if the patient had a stroke or 0 if not

IV. MACHINE LEARNING APPROACHES

Machine learning combines statistics, artificial intelligence and computer science to analyse and predict data. Machine learning approaches can be broadly classified into 4 categories based on how the data is labelled. The four broad categories are supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. In machine learning approach the supervised, unsupervised and semi-supervised uses a data labelling strategy to solve the problem. The reinforcement learning interacts with the surrounding environment.

A. Supervised Learning

A generic framework for forecasting future events, supervised learning is a machine learning strategy that uses methods that learn from examples that are externally given. The entire learning process is directed by the output variable. An unknown sample is evaluated using a known sample during the learning process. The result is known as classification or regression, and the sample is either an input or an output.

The training dataset is a well-known dataset that is used in this approach. Predictions are now conceivable as a result. The input data and response values make up the training dataset. A model is created during training in an effort to predict the response values of a fresh dataset. Test datasets are frequently used to verify how well machine learning models work. More predictive models are frequently produced by larger training datasets, which can simplify the new dataset. Decision tree (DT), k-nearest neighbours (KNN), support vector machine (SVM), artificial neural networks (ANNs), and random forests are a few examples of supervised learning methods (RF) [9].

The following are the machine learning supervised algorithms:

1) *Logistic Regression:* One of the many supervised learning models, regression seeks to provide a forecast of an output variable based on its input & later recognised variables. The most often utilised algorithms have been found to be stepwise regression, logistic regression, & linear regression. There are also more sophisticated regression algorithms being developed, such as multivariate adaptive regression and ordinary least squares regression.
2) *K- Nearest Neighbor:* In the K-nearest neighbour (KNN) method, the distance between a neighbour and the value of k, which indicates how many neighbours must be checked in order to represent the class of a sample data point, is measured. The two types of nearest neighbour algorithms are structure-based KNN and structure-less KNN. The structure-based approach essentially makes advantage of the data's fundamental structure, which has fewer mechanisms and is also linked to training data samples. The complete set of data is divided into training and sample points in the structure-less approach, and the point with the least distance between them is referred to as the closest neighbour.
3) *Support Vector Machine:* SVMs, or support vector machines, were the first to be developed for application in statistical learning theory. In essence, SVM is a binary classifier that builds a linear separating hyper plane to order the positions of the input points. Clustering, regression, and classification are the main applications of SVMs. SVMs are appealing in a variety of applications because they can handle more challenging issues that occur in high-dimensional spaces when it comes to global optimization. Support vector regression, least squares support vector machine, and successive projection algorithm-support vector machine are three often used SVM algorithms.
4) *Decision Tree:* One of the supervised learning methods used by machine learning algorithms is the decision tree. Both classification and regression use it. Data will be divided in this algorithm based on the parameters. An actual decision tree will include nodes and leaves. We will receive results or decisions at the leaves & data will be separated at the nodes.

There are two types of decision trees:

a) Classification tree - Here, a decision (outcome) variable will be obtained as categorical.
b) A decision (outcome) variable will be obtained as a continuous variable in the regression tree.

5) *Random forest:* One of the supervised machine learning methods, it is also used for regression and classification. However, categorization is where it is most frequently employed. The name alone suggests that it is a forest, and just as a forest is a collection of trees, so too will the random forest algorithm's trees be decision trees. Prediction outcomes will be more precise if we use more decision trees. In order to generate decision trees for each sample, the random forest method first creates random samples from the dataset. From those availabled trees, then choosen tree that would yield the best prediction results.

B. Unsupervised Learning

A form of machine learning technique called unsupervised learning is used to draw conclusions from datasets that contain input data but no labelled outputs. The two main learning goals in this strategy are association and clustering. To discover connections between things in a database, Rakesh Agarwal suggested using related learning. Clustering is used to combine datasets of the same kind and Apriori is the most used technique for association rules. The association rule learning algorithm and k-means clustering are two of the most popular algorithms used in unsupervised learning approaches. [10].

C. Semi-supervised Learning

Semi-supervised learning, which incorporates both supervised & unsupervised learning with some partial labelling of the data, is essentially a machine learning technique. Speech recognition, genetic sequencing, and web page classification are the main applications of this learning. Clustering and classification are the two main learning tasks that fall under semi-supervised learning. [10].

D. Reinforcement Learning

Reinforcement learning is a machine learning method that uses software agents to automatically recognise the whole of an activity in a certain environment and decide how to maximise its performance. The enhancement signal provides the agent with reward feedback to aid in its behavior-learning process. It includes the classification and control learning activities. Applications for reinforcement learning include computer-controlled board games, robotic hands, and self-driving cars. The three most often used algorithms in reinforcement learning are Deep Adversarial Networks, Temporal Difference, and Q-learning.[12]

V. PROPOSED METHODOLOGY

A research method describes the procedure that will be used to conduct the study. It will outline the study's research methods. A comprehensive study that will provide a fresh strategy for integrating machine learning in cardiac disease is necessary to close the research gaps. Exploratory, experimental, and applied research designs will constitute the foundation of the proposed study. It qualifies as exploratory since it involves the discovery of novel concepts and potential insights into the need for more research. By creating the datasets, tests are carried out on the data in order to test the tools. Since the presented work is for heart disease prediction, we may simply apply it in fixing problems with the help of applied research. The proposed methodology includes different steps in a flow chart. Basically we will start with first step as start referred to as importing the data set from the with .csv file which includes 12 parameters/attributes. Next in second stage we going to extract only required parameter or by removing class id and converting the values to 0, 1, 2, 3 by using fit 3rd is the preprocessing stages where we can explore the data. Data pre-process missing values, data cleaning, and normalize depending on the algorithms we use. After preproces of data, we are going to train and test the dataset with 80% & 20% ratio after that we are going to apply on 5 algorithms used in the proposed model are Decision Tree K-Nearest Neighbor, Logistic Regression, Random Forest Classifier, Support Vector Machine. After applying all the algorithms, we are going to predict and see the accuracy and from the accuracy we have found that it has given more as we can refer from the table 2 the prediction accuracy which has the improved accuracy. The suggested model is then put into practise, and its accuracy and performance are assessed. Here, an efficient classifier-based heart disease prediction system has been created. This model makes predictions using 12 medical qualities or factors. In Figure 1, the suggested technique is displayed.

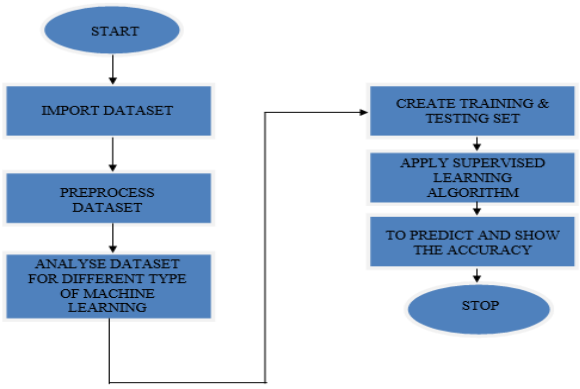


Fig. 1 Proposed Methodology

VI. RESULTS AND DISCUSSIONS

TABLE III
ALGORITHMS PREDICTION ACCURACY

SL. No	Algorithms	Prediction Accuracy
1	Decision Tree	90%
2	KNN	93%
3	Logistic Regression	93%
4	Support Vector Machine	93%
5	Random Forest.	93.63%

From the above table 2 we came to know that the algorithms we used for the prediction and accuracy was assumed as expected was successfully achieved. And we can see the bar graph for the accuracy the different algorithms how much accuracy we have achieved in 0-1 which will be converted to percentage and the blue colour in the bar graph shows the different algorithms with the percentage accuracy gained for the heart disease patients using 5 algorithms as mentioned from the supervised algorithms in machine learning.

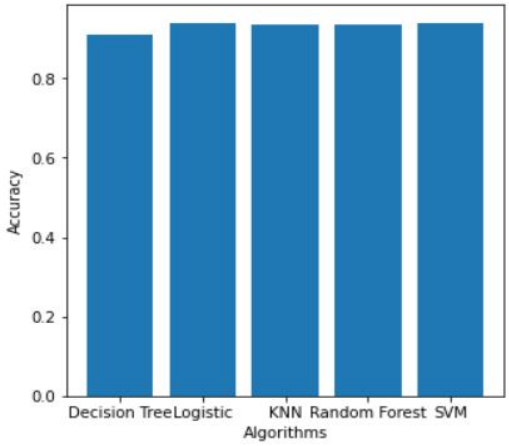


Fig. 2 Bar Graph Chart

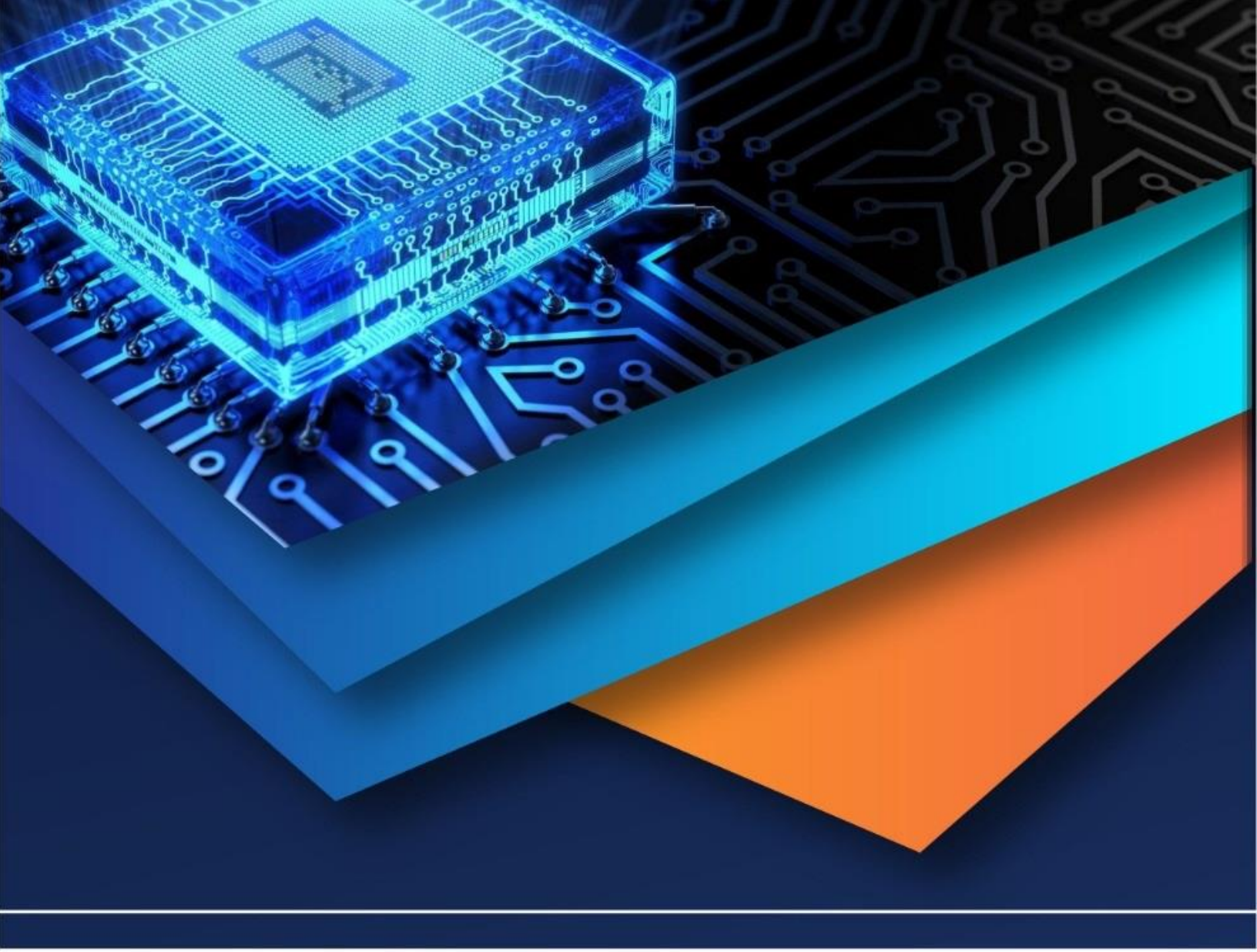
VII. CONCLUSIONS

From above experiment we can come to conclusion that when the data is pre-processed and using the technique with splitting of the data set to train & test and apply on algorithms, and we came to know that Random Forest have highest accuracy followed by SVM, KNN, LR, DT respectively.

REFERENCES

[1] A Comparative Study on Heart Disease Prediction Using Data Mining Techniques and Feature Selection, 2021, IEEE 2nd International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST)
[2] Youness Khourdifi and Mohamed Bahaji (2019). Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization, International Journal of Intelligent Engineering and Systems, Vol.12, No.1, 2019
[3] Youness Khourdifi and Mohamed Bahaji (2019). The Hybrid Machine Learning Model Based on Random Forest Optimized by PSO and ACO for Predicting Heart, ICCWCS 2019, April 24-25, Kenitra, Morocco.
[4] Komal Saini and Sandeep Sharma (2019). Review on the Heart Disease Detection Using IoT Framework, International Journal of Computer Sciences and Engineering Vol.7(3), Mar 2019, E-ISSN: 2347-2693.
[5] Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivasatava (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques, IEEE access volume7,2019.

[6] C. Gazeloglu, “Prediction of heart disease by classifying with feature selection and machine learning methods”, Progress in Nutrition 2020; Vol. 22, N. 2: 660-670, IEEE, DOI: 10.23751/pn.v22i2.9830.
[7] C. B. C. Latha, S. C. Jeeva, “Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques”, Informatics in Medicine Unlocked 16,2019.
[8] Haq, A. U., Li, J., Memon, M. H., Hunain Memon, M., Khan, J, & Marium, S. M (2019). HeartaDisease Prediction System Using Model of Machine Learning and Sequential Backward Selection Algorithm foraFeatures Selection. 2019 IEEE 5th International Conference for Convergenceain Technology (I2CT).
[9] Amin Ul Haq et.al, “Feature Selection Based on L1-Norm Support Vector Machine and Effective Recognition System for Parkinson's Disease Using Voice Recordings” IEEE Access, issue date December 2019, volume: 7, Issue: 1,on page(s)37718-37734, Print ISSN:2169- 3536
[10] Shadman Nashif, Md. Rakib Raihan, Md. Rasedul Islam,and Mohammad Hasan Imam (2018).Heart Disease Detection by Using Machine Learning Algorithms and a Real Time Cardiovascular Health Monitoring System, World Journal of Engineering and Technology, 2018, 6, 854-873.
[11] Kusuma and Divya Udayan (2018). Machine Learning and Deep Learning Methods in Heart Disease (HD) Research, International Journal of Pure and Applied Mathematics Volume 119 No. 18 2018, 1483-1496.
[12] Pooja, Aakansha Sharma and Ankush Sharma (2018). Machine Learning: A Review of Techniques of Machine Learning, Journal of Applied Science and Computations Volume 5, Issue 7, July /2018.ISSN NO: 1076-51



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY