iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 Issue: V Month of publication: May 2021

DOI: https://doi.org/10.22214/ijraset.2021.3412

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: 7.429
Volume 9 Issue V May 2021- Available at www.ijraset.com

Prediction of Heart Disease using Machine Learning Algorithms

Aadar Pandita[1], Siddharth Vashisht[2], Aryan Tyagi[3], Prof. Sarita Yadav[4]
[1, 2, 3, 4]*Department of Information Technology, Bharati Vidyapeeth's College of Engineering, New Delhi*

Abstract: Heart diseases have been the primary reason for death all over the world. Majority of the deaths related to cardiovascular problems are caused by heart attacks and strokes. The World Health Organization (WHO) indicates that an approximate 17.9 million people die due to such diseases every year. Therefore, it is essential that we find methods to ensure the minimization of these numbers. In order to minimize the detrimental effects of heart diseases, we must try to predict its presence at earlier stages. Machine Learning algorithms can help us effectively predict such results with a high degree of accuracy which can in turn help doctors and patients detect the onset of such diseases and reduce their impact or prevent them from occurring. Our objective is to create a system that is able to accurately determine the presence of heart disease in a time and cost efficient manner.

Keywords: Cardiovascular Diseases; Support Vector Machine; K- Nearest Neighbour; Naive Bayes; Random Forest; Logistic Regression; Machine Learning; Prediction Model

I. INTRODUCTION

Machine Learning is an extremely useful tool that is based on Artificial Intelligence and its ability to learn from data on its own. This tool can be especially helpful in healthcare systems to develop prediction models so as to help researchers and doctors understand the patterns of various diseases and overcome previous limitations.

In this paper, we are going to focus on heart diseases. Heart diseases are caused due to a number of factors that include hypertension, high blood pressure, old age, high cholesterol, etc. We use the Cleveland heart disease dataset provided by the University of California, Irvine (UCI) repository to create and improve our prediction models.

The dataset consists of 303 instances in total with 76 attributes. Since the data has certain missing values and is also noisy, hence, the data needed to be cleaned and processed in order to be used effectively and produce desirable results.

We used various machine learning algorithms to find the best overall accuracy and thereby predict the outcome for user input data on a web-based application.

II. BACKGROUND

According to the WHO, there are different types of Cardiovascular diseases that affect people. Majority of the related deaths, close to 85%, are caused by heart attack and strokes, while a third of the deaths occur prematurely in people under 70 years of age. Other types of heart diseases include the following :

Types of Heart Diseases

Disease	Description
Coronary heart disease	Disease of the blood vessels supplying the heart muscle
Cerebrovascular disease	Disease of the blood vessels supplying the brain
Peripheral arterial disease	Disease of blood vessels supplying the arms and legs
Rheumatic heart disease	Damage to the heart muscle and heart valves from rheumatic fever, caused by streptococcal bacteria
Congenital heart disease	Malformations of heart structure existing at birth
Deep vein thrombosis and pulmonary embolism	Blood clots in the leg veins, which can dislodge and move to the heart and lungs

III. LITERATURE SURVEY

A comparative study of various algorithms in related work

Year	Author	Purpose	Techniques Used	Accuracy
2019	M. Marimuthu et al. [3]	Analysis of Heart Disease Prediction using Various Machine Learning Techniques	KNN NB Decision Tree SVM	83.50% 80.66% 75.58% 65.56%
2019	Reddy Prasad et al. [4]	Heart Disease Prediction using Logistic Regression Algorithm using Machine Learning	Logistic Regression NB Decision Tree	86.89% 86% 78.69%
2018	Pahulpreet Singh Kohli et al. [5]	Application of Machine Learning in Disease Prediction	Logistic Regression Decision Tree Random Forest SVM AdaBoost	87.1% 70.97% 77.42% 83.87% 83.87%
2019	Erin M. Kunz [6]	Heart Disease Prediction Using Adaptive Network-Based Fuzzy Inference System (ANFIS)	LR KNN SVM NN ANFIS	0.809 0.89 0.84 0.787 0.891

2018	Dinesh K G et al. [7]	Prediction of Cardiovascular Disease Using Machine Learning Algorithms	LR RF NB SVM	0.8651685 0.8089888 0.8426966 0.7977528
2018	Amin Ul Haq et al. [8]	A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms	LR KNN NB Decision Tree Random Forest Artificial Neural Network SVM	84% 76% 83% 74% 83% 74% 75%
2016	Dr. S. Seema et al. [9]	Predict chronic disease by mining the data containing in historical health records	NB Decision Tree SVM	Highest accuracy in case of heart disease 95.556% is achieved by SVM. Highest accuracy in case of diabetes 73.588% is achieved by Naïve Bayes.

2016	Ashok Kumar Dwivedi [10]	Evaluate the performance of different machine learning techniques for prediction of heart disease using tenfold cross-validation	NB Classification Tree KNN Logistic Regression SVM ANN	83% 77% 80% 85% 82% 84%
2017	Syed Muhammad Saqlain Shah et al. [11]	Analysis of Heart Disease Diagnosis based on feature extraction using KFold cross-validation	SVM	91.30% is the highest accuracy obtained
2016	Muhammad Saqlain et al. [12]	Identification of Heart Failure by Using Unstructured Data of Cardiac Patients	LR Neural Network SVM Random Forest Decision Tree Naive Bayes	80% 84.8% 83.8% 86.6% 86.6% 87.7%

IV. METHODOLOGY

A. *Data Pre-Processing*

1) *Cleaning:* The data that is to be processed is not clean and will contain outliers, noise or it may contain missing values. If this data is processed further it will not give best results. Therefore, it is necessary to pre-process the data so as to eliminate unnecessary values from it or to fill missing values according to the remaining values in the dataset.
2) *Transformation:* The processed data is further simplified from one format to another so as to make the model understand it even more by doing scaling of it. The dataset values are brought down to the range of -3 to 3 so as to increase uniformity in the dataset.
3) *Reduction:* Working on complex data is difficult to understand and for the system to understand the data, it is reduced to a format that is easily understandable to the model and gives good results.

B. *Algorithms and Techniques used*

1) Logistic Regression
2) K- Nearest Neighbor
3) Support Vector Machine
4) Naive Bayes
5) Random Forest

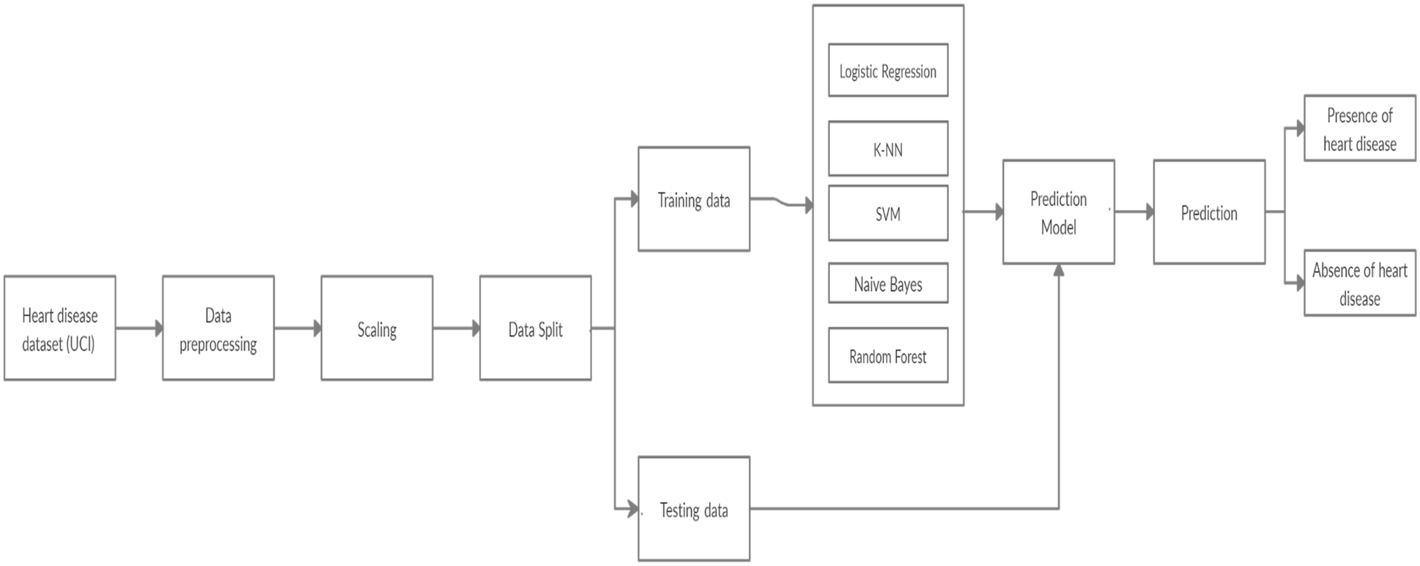
Components of Dataset

Attribute	Representation in Dataset	Attribute Description	Value Type
Age	age	Age in Years	Integer
Sex	sex	Value 1 : Male Value 0 : Female	Integer
Chest Pain Type	cp	Value 1 : Typical angina Value 2 : Atypical angina Value 3 : Non-anginal pain Value 4 : Asymptomatic	Integer
Resting Blood Pressure	trestbps	(in mm Hg on admission to the hospital)	Integer
Serum Cholesterol	chol	Serum cholesterol in mg/dl	Integer
Fasting Blood Sugar	fbs	If Value > 120 mg/dl Value 1 : True Value 0 : False	Integer
Resting Electrocardiographic Result	restecg	Value 0 : normal Value 1 : having ST-T wave abnormality Value 2 : showing probable or definite left ventricular hypertrophy by Estes' criteria	Integer
Maximum Heart Rate	thalach	Maximum heart rate achieved	Integer
Exercise Induced Angina	exang	Value 1 : Yes Value 0 : No	Integer
Old Peak	oldpeak	ST depression induced induced by exercise relative to rest	Real
Slope of peak exercise ST segment	slope	Value 1 : Upsloping Value 2 : Flat Value 3 : Downsloping	Integer
Number of major vessels colored by fluoroscopy	ca	Range : 0-3	Integer
Thalassemia	thal	Value 3 : Normal Value 6 : Fixed Defect Value 7 : Reversible Defect	Integer

V. PROPOSED SYSTEM

This section defines how the problem is solved and what all steps are used in it. Heart Disease dataset from UCI repository is used. Preprocessing is done on the data so as to make it cleaner and obtain best results. The processed data is further split into training and testing data. Furthermore, it is scaled using standard scaler so as to ensure that all the values present in the dataset are within a range i.e. is -3 to 3, 67% values using Standard scaler are in range -1 to 1.

Training Data is passed through various models so that they can predict the outcome of future data entries made by the user by simulating the results they had been instilled with during the training phase.



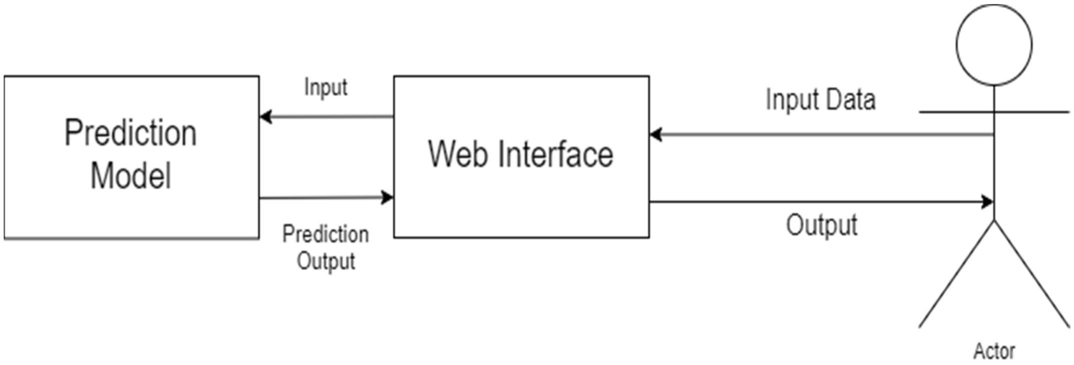
System Architecture

After the model is trained, the testing data is inputted into the model and it will generate some predictions according to the problem statement. These predictions can be compared with the original data so as to calculate the accuracies of our models.

A. Web Application

Since a lot of the predictions made by health practitioners are based on their intuition, the chances of error can be high as they could be based on inaccurate diagnosis.

Our goal is to help patients and doctors in their pursuit for accurate and efficient prediction results backed by data driven models.



Web Application Model

Therefore, a web application is developed using the most accurate model. Flask is used to access tools, libraries and technologies that allows to build a web application. Based on the inputs provided by the user on the web application interface, it displays the prediction on the application itself as the output. It increases user accessibility to make predictions based on the user's own data.

VI. RESULTS

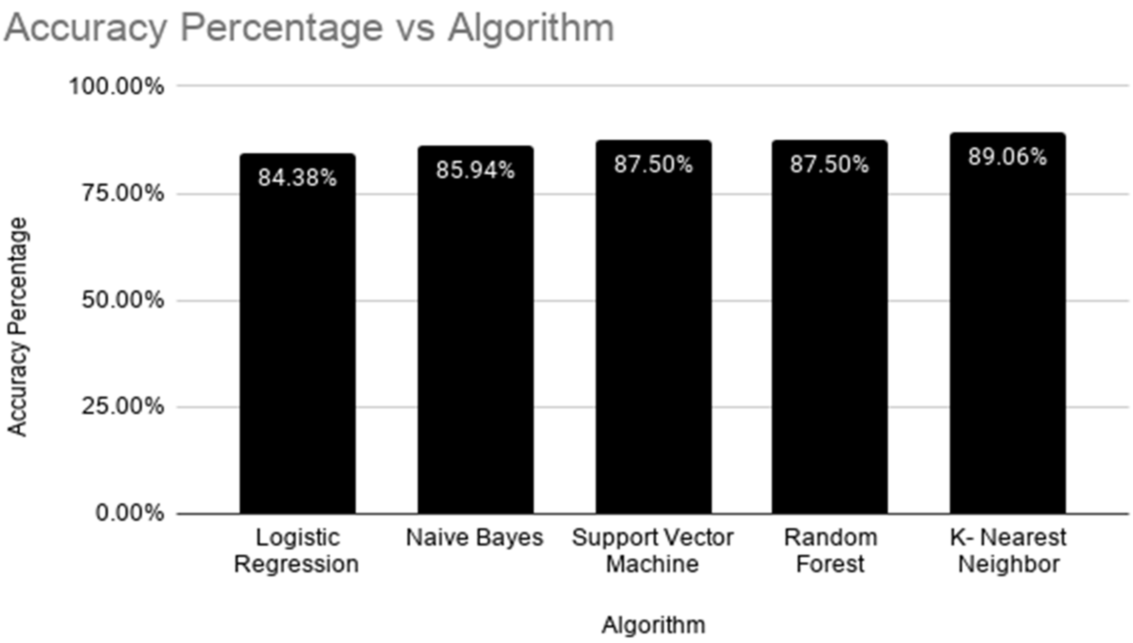
Machine Learning algorithms were successfully applied on the UCI heart disease dataset

.

Algorithm	**Resultant Accuracy**
Logistic Regression	84.38%
K-nearest neighbors	89.06%
Support Vector Machine	87.50%
Naive Bayes	85.94%
Random Forest	87.50%

Comparison Table of Results in terms of Accuracy

Comparison of Classification Results in terms of Accuracy



VII. CONCLUSION

Heart disease is the most common disease that leads to death on our planet according to the World Health Organisation. The technology has been advancing at a rampant rate and it would be unwise to not utilize its full potential in the field of medical sciences where every error leads to a possible loss of life. K-NN gives the best results based on our model with an accuracy of 89.06% while Logistic Regression gives the least accurate prediction at 84.38%.

VIII. FUTURE WORK

The accuracies obtained by various algorithms can be further improved so as to reduce the chances of error in the prediction models. Future researchers can eradicate false positives from the current prediction results. Furthermore, these algorithms can also be used in other types of diseases to predict the pattern of their progression. Web and mobile applications can be developed where users can enter their personal medical details and get a prediction result based on their health conditions.

REFERENCES

[1] WHO Fact Sheet. https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds**)**

[2] UCI Dataset Heart Disease. https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[3] Muthuvel, Marimuthu & Sivaraju, Deivarani & Ramamoorthy, Gayathri (AISGSC 2019) "Analysis of heart disease prediction using various machine learning techniques"

[4] Reddy Prasad,Pidaparthi Anjali, S.Adil, N.Deepa (2019), "Heart disease prediction using logistic regression algorithm using machine learning". International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8, Issue-3S

[5] P. S. Kohli and S. Arora, "Application of machine learning in disease prediction," 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-4

[6] Erin M. Kunz, "Diagnosis of heart disease using adaptive network-based fuzzy inference system (ANFIS)." CS 229 SPRING 2019, Stanford University (unpublished)

[7] Dinesh, Kumar & Arumugaraj, K & Santhosh, Kumar & Mareeswari, V. (2018). "Prediction of cardiovascular disease using machine learning algorithms". 1-7. 10.1109/ICCTCT.2018.8550857

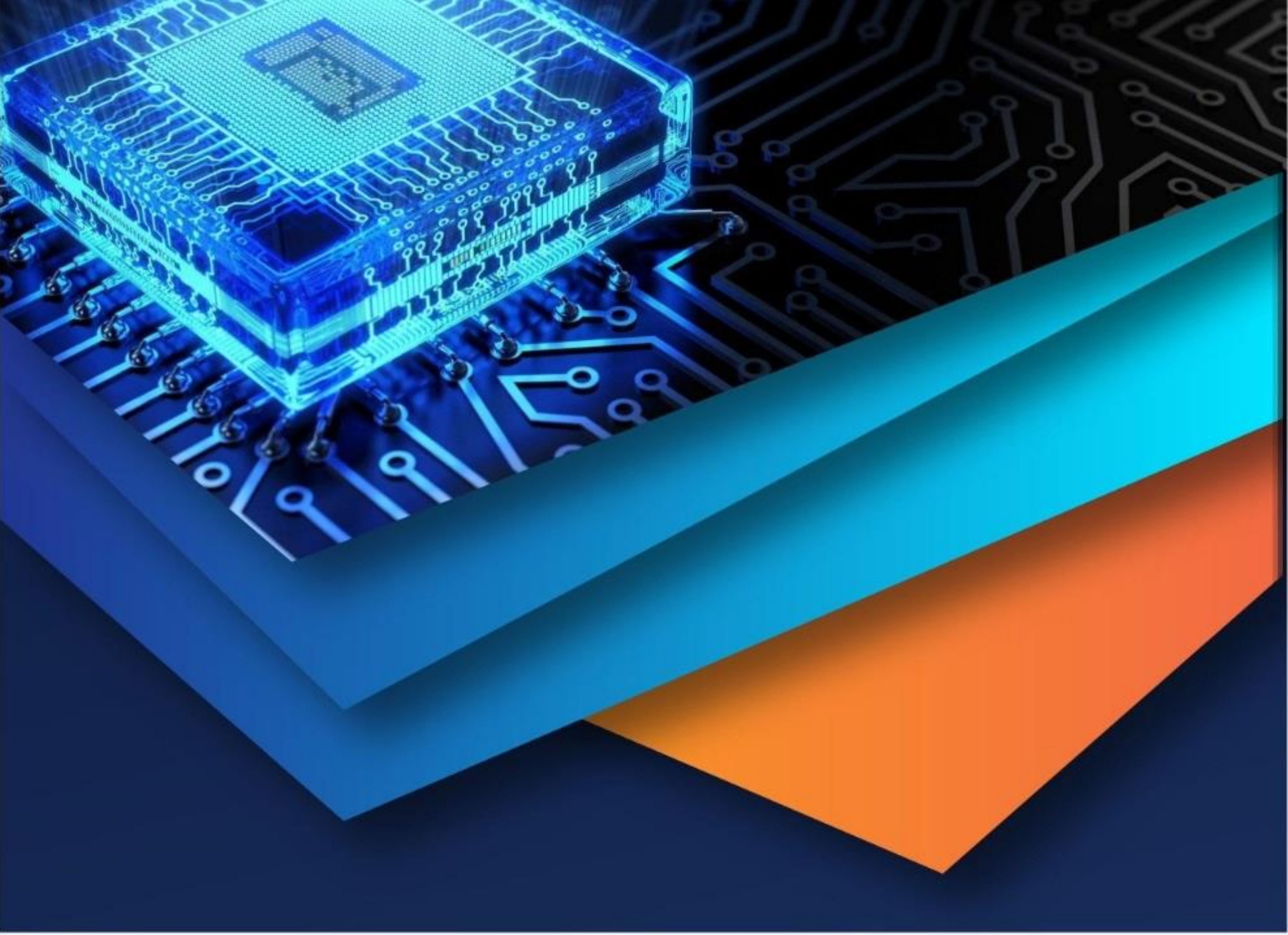
[8] Amin Ul Haq, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, Ruinan Sun, and Iván García-Magariño "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms." Volume 2018, Article ID 3860146

[9] K. Deepika and S. Seema, "Predictive analytics to prevent and control chronic diseases," 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Bangalore, 2016, pp. 381-386

[10] Dwivedi Ashok "Performance evaluation of different machine learning techniques for prediction of heart disease," (2016) Neural Computing and Applications. 29. 10.1007/s0052

[11] M. Saqlain, W. Hussain, N. A. Saqib, and M. A. Khan, "Identification of heart failure by using unstructured data of cardiac patients," 2016 45th Int. Conf. Parallel Process. Work., pp. 426–431, 2016

[12] S. M. S. Shah, S. Batool, I. Khan, M. U. Ashraf, S. H. Abbas, and S. A. Hussain, "Feature extraction through parallel Probabilistic Principal Component Analysis for heart disease diagnosis," Phys. A Stat. Mech. its Appl., vol. 482, pp. 796–807, 2017



10.22214/IJRASET

INDEX COPERNICUS

45.98

IMPACT FACTOR: 7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)