iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** I **Month of publication:** January 2024

DOI: https://doi.org/10.22214/ijraset.2024.58216

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 12 Issue I Jan 2024- Available at www.ijraset.com

Cardiovascular Disease Prediction Using Machine Learning Techniques

Arin Singh[1], Anjali Chaudhary[2], Dibyanshi Chaudhary[3], Yatharth Vaish[4], Prof. Sachin Tyagi[5]
Department of Electronics and Communications Engineering, KIET Group of Institutions, Ghaziabad, India

Abstract: ***The study examines the concerning increase in heart disease cases and emphasizes the importance of proactive forecasting. The timely and accurate detection of this growing health issue is essential. The main aim of the study is to determine whether a patient is at the risk of heart disease or not based on the dataset having the medical history of the patient.***
There are a number of linear and non-linear machine learning algorithms including Logistic Regression, K-nearest Neighbor, State Vector Machine and Random Forest Classifier which are used to determine the likelihood of having cardiovascular problems in the patient. The implementation of the algorithms such as Logistic Regression and Random Forest Classifier has shown the great results by providing a higher accuracy in predicting CVDs rather than the traditionally used methods such as Naive Bayes. The proposed method is not only cost effective but also is easier to use. It just requires some basic information of the patient and based on the feeded medical history of the patient , the results are produced. It improves the quality of the medical care as the patient can easily diagnose the possibility of having Cardiac diseases. This is helpful in preventing serious heart problems by providing an alarm to the patient at an early stage.

I. INTRODUCTION

Machine learning is a method by which one can utilize and withdraw anonymous and likely useful information from the provided dataset .With the expansion of its applications day by day, the scope of machine learning is diverging.Machine learning is a mechanism in which various classifiers from ensemble, supervised and unsupervised learning are integrated to predict the accuracy of the given dataset. This gathered information plays a significant role in the prediction of various diseases such as the prediction of heart diseases and lungs related diseases which are growing at an alarming rate nowadays and hence, it is a boon for the entire human race. It has been estimated by the World Health Organization that heart related disorders take a toll of around 17.9 million of the people every year including a larger number of males as compared to the females. It has been observed in recent years that a greater number of adults are getting affected with heart disorders because of unhealthy lifestyle choices. With the help of our project, we will be able to detect the risk of cardiovascular diseases based on the medical history of the patient . The symptoms include increased blood pressure, cholesterol etc. There are four algorithms based on which the data is predicted, these include: (1)Random Forest Classifier, (2)K-nearest neighbors (KNN), (3)Logistic Regression, and (4)Support Vector Machine (SVM) out of which Logistic Regression and the non linear mechanism i.e. Random Forest Classifiers are the ones predicting the most accurate data upto 81.90%. The purpose served by this project is to scrutinize if the person is suffering from any Cardiovascular Disease or not.The dataset is derived from the University of California Irvine machine learning repository. This dataset includes the various attributes along with the medical history of the patient.There are fourteen parameters in total out of which thirteen parameters are provided as inputs which are trained using four machine learning algorithms - Logistic Regression, K-nearest neighbors(KNN), Random Forest Classifier and Support Vector Machine(SVM).The trained parameters include age, sex, Cp(chest pain), Trestbps (Resting Blood Pressure), Chol (Serum Cholesterol), Fbc (Fasting Blood Sugar), Restecg (Resting Electrocardiograph), Thalach(Maximum Heart Rate), Exang(Exercised Induced Angina), Oldpeak(ST Depression when Workout compared to the amount of rest taken), Slope (slope of peak exercise ST segment), Ca(number of coloured vessels by fluoroscopy), Thal(defect type).The fourteenth parameter is ‘target’. It is an output generated based on previous data. The value ‘1’ of ‘target’ is suggestive of cardiovascular disorder whereas the value ‘0’ denotes that the patient is not diagnosed with any heart related medical condition.

II. RELATED WORK

The usage of linear and non-linear machine learning algorithms for detection of various cardiovascular diseases enlights the spark for the effort made behind this project. The study provides a succinct overview of the literature and investigates various algorithms, such as Random Forest Classifier, KNN, SVM, and Logistic Regression, to effectively predict cardiovascular disease. The particular advantages of each algorithm in accomplishing predetermined goals are emphasized in the Results section.

©IJRASET: All Rights are Reserved

The Integrated Heart Disease Prediction System (IHDPS), a recently introduced model, integrates the capacity to define decision limits using both conventional and innovative machine learning and deep learning models. It incorporates important elements including a family history of heart disease. Nevertheless, the IHDPS model's accuracy is not as high as that of other developing models, particularly when it comes to the identification of coronary heart disease through the use of artificial neural networks and other sophisticated machine and deep learning algorithms. In McPherson et al.'s work, risk factors for atherosclerosis of coronary heart disease are identified through the use of a neural network-based implementation algorithm that is built-in. Their method correctly determines if a test patient has the specified illness.

Furthermore, R. Subramanian et al. advance the area by presenting a deep neural network for the diagnosis and prediction of blood pressure and heart disease, among other characteristics. Their model uses 120 hidden layers, which is an important method for guaranteeing accurate findings, particularly when used to Test Datasets. It also integrates pertinent disease-related data. Testing the model with unknown data provided by a physician demonstrates the efficacy of the supervised network in diagnosing cardiac disorders. After being trained on previously acquired data, the model predicts outcomes with precision, proving its dependability and computing accuracy in the diagnosis of cardiac conditions.

III. DATA SOURCE

A carefully selected dataset was put together, with an emphasis on individuals and their extensive medical histories—particularly taking into account their experiences with cardiac issues and other pertinent medical diseases. An array of conditions affecting the heart is referred to as heart diseases. As per the data provided by the World Health Organization, heart disorders are serving as the main cause of death in those belonging to their middle ages. This is the topic of serious concern rising worldwide nowadays. The dataset having the medical history of almost 304 people belonging to different age groups was collected for the thorough study of CVDs.

This dataset included several attributes such as age, blood pressure at rest, blood sugar levels during fasting, and more. The important role of determining if a patient has been diagnosed with heart disease is made easier by the abundance of medical data available. The information, which includes 13 different medical variables for 304 patients, makes it easier to identify people who are at risk of heart disease. It is essential for categorizing patients and separating those who are susceptible from those who are not.

This Heart Disease dataset, which comes from the UCI repository, is useful for identifying trends that may indicate a patient's risk of developing heart disease. Two subsets of the dataset are separated out for testing and training. With 14 columns and 304 rows total—each row denoting a distinct record—the dataset offers a thorough framework for analysis. The characteristics are carefully described in 'Table 1,' providing essential information for additional research and understanding of heart disease prognosis.

Various Attributes used

S.No.	Observations	Interpretation	Data type
1.	Age	Age in years	continuous
2.	Sex	Gender of the patient	Male/female
3.	Cp	Chest Pain	Four types
4.	Trestbps	Resting Blood Pressure	continuous
5.	Chol	Serum Cholesterol	continuous
6.	Fbc	Fasting Blood Sugar	<, or >120 mg/dl
7.	Restecg	Resting Electrocardiograph	Five values
8.	Thalach	Maximum heart rate achieved	continuous
9.	Exang	Exercise Induced Angina	Yes/No
10.	Oldpeak	ST Depression when Workout compared to amount of rest taken	continuous
11.	Slope	Slope of peak exercise ST segment	Up/flat/down
12.	Ca	Number of major vessels coloured by fluoroscopy	0-3
13.	Thal	Defect type	Reversible/fixed/normal
14.	Target	Heart disease	Present(1), not present(0)

Table 1: Attributes for Data Set

IV. METHODOLOGY

A thorough examination of several machine learning methods, namely:- K Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, and Random Forest Classifiers, is presented in this work. In order to accurately diagnose heart problems, practitioners and medical analysts can benefit greatly from the insights provided by these algorithms. To ensure a solid analysis, the study includes a comprehensive evaluation of contemporary journals, published articles, and data on cardiovascular illness.

The research approach utilized in this study functions as a foundation for the suggested model. This methodical procedure consists of a series of actions that convert unprocessed data into observable patterns, improving user understanding. The suggested approach, which is depicted in Figure 1, consists of multiple crucial phases. First, information must be gathered; in the second phase, important values must be extracted. Preprocessing the data, which includes missing value resolution, data cleansing, and algorithm-specific normalization, is covered in the third step.

The model uses classifiers to classify the prepared data once it has undergone data preparation. Notably, the Random Forest Classifier, SVM, KNN, and Logistic Regression are the classifiers used in the suggested model. Using a variety of performance indicators, the accuracy and performance of the suggested model are examined in the last phase of the process.

An Effective Heart Disease Prediction System (EHDPS) is created within this model, utilizing 14 medical characteristics for precise prediction, such as age, sex, blood pressure, cholesterol, chest discomfort, and fasting sugar. The application of various classifiers results in a strong and adaptable predictive model that is intended to improve the precision and effectiveness of heart disease detection.

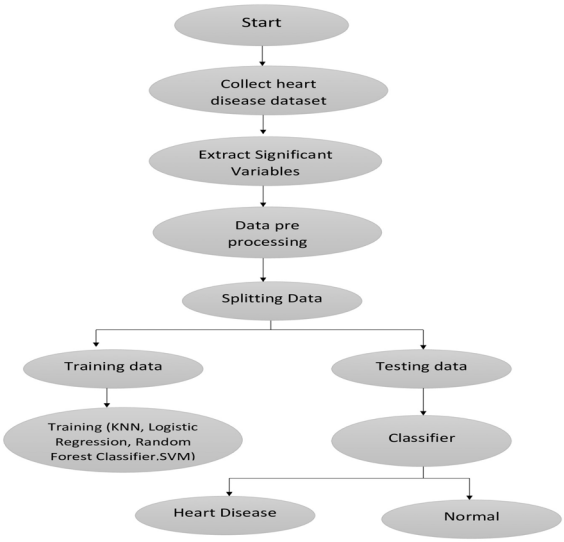


Figure 2: Proposed Model

V. RESULTS AND DISCUSSION

The results show that there has been a notable shift in the algorithms used to identify people with heart disease. Although Support Vector Classifier (SVC) and Decision Trees are frequently used by researchers, our research shows that K Nearest Neighbors (KNN), Random Forest Classifier, and Logistic Regression perform better**.**

With an accuracy of up to 81.90%, Logistic Regression significantly outperforms accuracies found in earlier research. This improvement is a result of our dataset's medical features being used more frequently and similar is the case with non-linear random forest classifier algorithm.From this we can conclude that Logistic Regression and the Random forest classifier are the two best techniques to predict heart diseases accurately.

The prediction of CVDs in the patients from various age groups, sex groups, resting blood pressure levels, and chest pain classes has been represented with the help of figure 2, 3, 4 and 5. These graphical displays provide a concise and informative summary of how well the classifier performed in classifying and predicting patients according to these important criteria. Overall, our work improves the prediction accuracy of heart disease and highlights the superiority of KNN and Logistic Regression over other classifiers.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

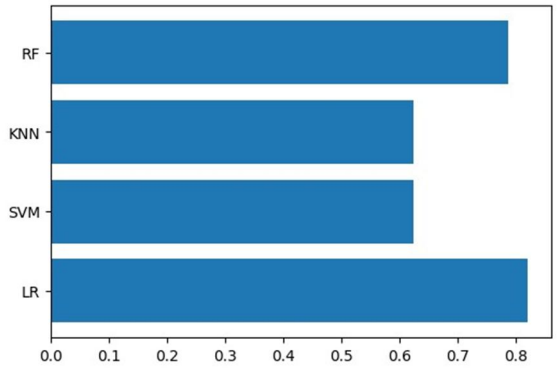


Figure 2: Comparison Table of Used Algorithms

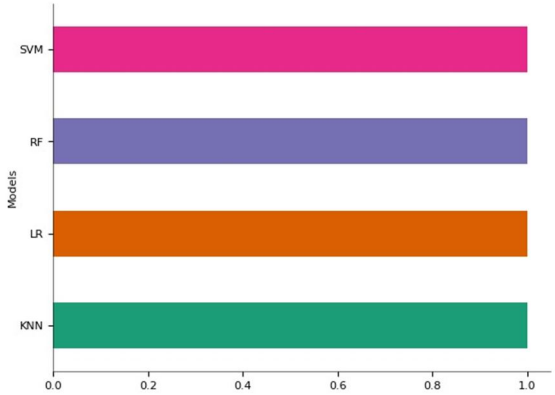


Figure 3: Model Results

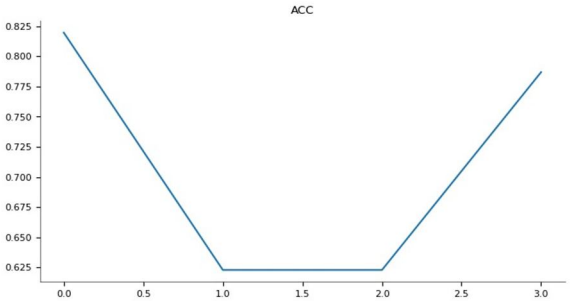


Figure 4: Accuracy Curve

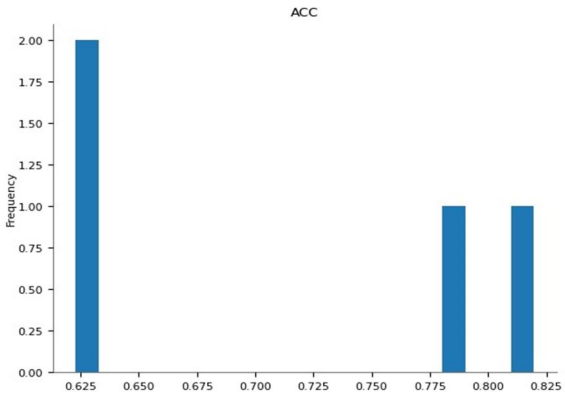


Figure 5: Accuracy

	Models	ACC
0	LR	0.819672
1	SVM	0.622951
2	KNN	0.622951
3	RF	0.786885

Table 2: Accuracy Table

VI. CONCLUSION

In this study, four machine learning classification algorithms are used to introduce a model for the diagnosis of cardiovascular illness. The objective is to identify people who are at risk of cardiovascular disease **by examining** their medical history, which was taken from a dataset that included important details like blood pressure, sugar levels, and chest pain, among other things. Using clinical data from patients, this heart disease detection method determines the probability of a prior heart disease diagnosis.

This model's remarkable 81.96% accuracy is the result of the use of several methods, including K Nearest Neighbors (KNN), Random Forest Classifier, Support Vector Machine (SVM), and by the use of machine learning approaches, which also result in significant cost savings.Logistic Regression. The model's capacity to precisely determine a person's susceptibility to heart disease is improved by the integration of a large training dataset. The prediction process is improved and accelerated

The project demonstrates how machine learning approaches are more effective than human capabilities in terms of prediction accuracy. The availability of numerous medical databases creates opportunities for additional study and the use of these cutting-edge methods in healthcare. Finally, by fine-tuning the dataset and utilizing logistic regression and KNN, our research significantly contributes to the prediction of patients with cardiac illnesses. Our model has an average accuracy of 72.12%, which is an improvement above earlier models that had an accuracy of 71.3%. With an astounding accuracy of 81.90%, Logistic Regression stands out as the most accurate of the four algorithms that were employed. Here, with the help of figure 6, the percentage of people likely to have heart disease and the people who are at a lower risk of CVDs **has** been shown graphically.

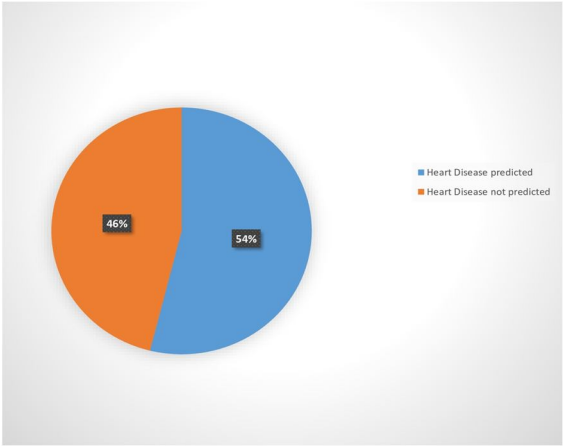


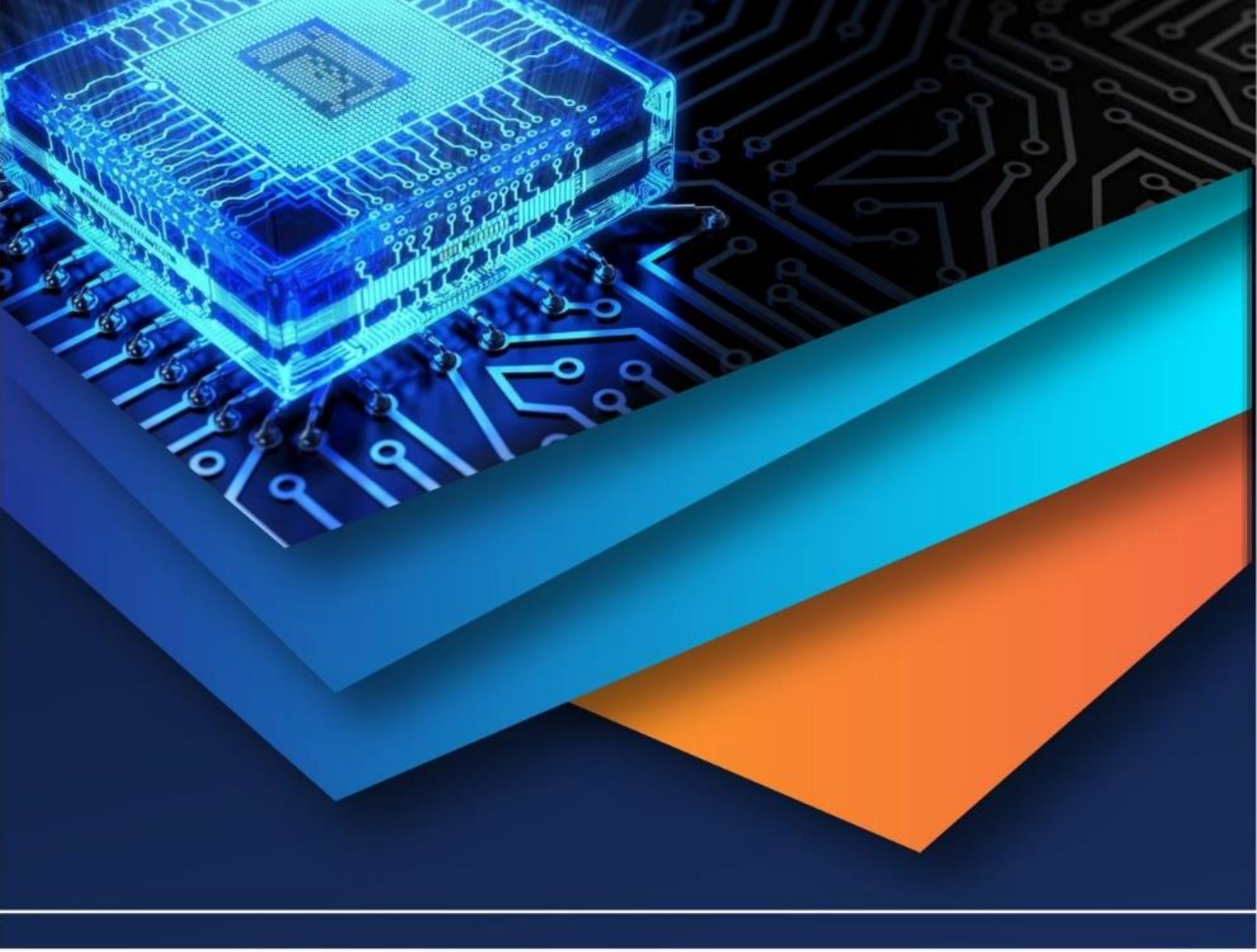
Figure 6: Graph of Patients report with the help of ML Techniques

REFERENCES

[1] Soni, S., & Vyas, O. (2010, July 10). Using Associative Classifiers for Predictive Analysis in Health Care Data Mining. International Journal of Computer Applications, 4(5), 33–37. https://doi.org/10.5120/821-1163
[2] Biomedicine publication information. (2006, January). 10(1), c2–c2. https://doi.org/10.1109/titb.2005.863889
[3] Siddique, S. (2022, July 31). Health Information Exchange using BlockChain and Cardiac Disease Prediction using Naïve Bayes Algorithm. https://doi.org/10.22214/ijraset.2022.45780

[4] Saito, I. (2022, March 1). A Risk Prediction Model of Atherosclerotic Cardiovascular Disease in Japan. Journal of Atherosclerosis and Thrombosis, 29(3), 320–321. https://doi.org/10.5551/jat.ed172
[5] Jabbar, M. A., Deekshatulu, B., & Chandra, P. (2013). Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm. Procedia Technology, 10, 85–94. https://doi.org/10.1016/j.protcy.2013.12.340
[6] Wang, J., Chen, Q., Wang, L., Zhou, S., Cheng, L., Xie, X., Huang, G., Wang, B., & Ma, X. (2011, April). Identifying novel mutations of NKX2-5 congenital heart disease patients of Chinese Minority Groups. International Journal of Cardiology, 148(1), 102–104. https://doi.org/10.1016/j.ijcard.2010.05.041
[7] Parthiban, L., & Subramanian, R. (2009). CANFIS—a computer aided diagnostic tool for cancer detection. Journal of Biomedical Science and Engineering, 02(05), 323–335. https://doi.org/10.4236/jbise.2009.25048
[8] Heart Attack Detection By Heartbeat Sensing using Internet Of Things : IoT. (2018, May 2). International Journal of Modern Trends in Engineering & Research, 5(4), 212–216. https://doi.org/10.21884/ijmter.2018.5124.xutya
[9] Anakal, S., Uppin, C., & Bhadrashetty, A. (2019, January 31). International Journal of Computer Sciences and Engineering, 7(1), 907–910. https://doi.org/10.26438/ijcse/v7i1.907910
[10] Automated search databases at the U.S. patent and trademark office. (1986, January). World Patent Information, 8(4), 309. https://doi.org/10.1016/0172-2190(86)90095-5
[11] TAKCI, H. (2018). Improvement of heart attack prediction by the feature selection methods. TURKISH JOURNAL OF ELECTRICAL ENGINEERING & COMPUTER SCIENCES, 26, 1–10. https://doi.org/10.3906/elk-1611-235
[12] Brown, W. H. (1995, December). Trends in patent renewals at the United States patent and trademark office. World Patent Information, 17(4), 225–234. https://doi.org/10.1016/0172-2190(95)00043-7

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)