iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IX **Month of publication:** September 2024

DOI: https://doi.org/10.22214/ijraset.2024.64051

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 IX Sep 2024- Available at www.ijraset.com

Heart Disease Prediction Using Machine Learning

Gupta Raj[1], Kumar Shashi[2], Anand Animesh[3], Sneha HR[4], Sushma V[5]
Department of ISE, NMIT, Bangalore, India

Abstract: ***The number of heart disease cases is rising quickly every day, making it crucial and worrisome to anticipate any such illnesses in advance. This diagnosis is a challenging task that requires accuracy and efficiency. The primary goal of the research work is to determine which patient, based on different medical parameters, is more likely to have a cardiac condition. Using the patient's medical history, we developed a heart disease prediction algorithm to determine the likelihood of a heart disease diagnosis or not. To forecast and categorize the patient with heart disease, we employed a variety of machine learning algorithms, including logistic regression and kernel neighborhood network. A very helpful method was employed to control the model's ability to increase any person's heart attack prediction accuracy. In comparison to the previously utilized classifiers, such as naive bayes, etc., the proposed model's strength was quite satisfying. It could predict signs of heart disease in a specific individual by utilizing KNN and Logistic Regression, which demonstrated a good accuracy. Thus, by utilizing the provided model to determine the likelihood that the classifier can correctly and precisely diagnose cardiac illness, a sizable amount of pressure has been released. The Given heart disease prediction system lowers costs and improves medical care. This research provides us with important information that can be used to forecast the patients who will have heart disease. It is used with the.pynb file format.***

Index Terms: supervised; unsupervised; reinforced; linear regression; decision tree; python programming; jupytor Notebook; confusion matrix;

I. INTRODUCTION

Heart is one of the most extensive and vital organ of human body so the care of heart is essential. Most of diseases are related to heart so the prediction about heart diseases is necessary and for this purpose comparative study needed in this field, today most of patient are died because their diseases are recognized at last stage due to lack of accuracy of instrument so there is need to know about the more efficient algorithms for diseases prediction. Machine Learning is one of the efficient technology for the testing, which is based on training and testing. It is the branch of Artificial Intelligence(AI) which is one of broad area of learning where machines emulating human abilities, machine learning is a specific branch of AI. On the other hand machines learning systems are trained to learn how to process and make use of data hence the combination of both technology is also called as Machine Intelligence. As the definition of machine learning, it learns from the natural phenomenon, natural things so in this project we uses the biological parameter as testing data such as cholesterol, Blood pressure, sex, age, etc. and on the basis of these, comparison is done in the terms of accuracy of algorithms such as in this project we have used four algorithms which are decision tree, linear regression, k-neighbour, SVM. In this paper, we calculate the accuracy of four different machine learning approaches and on the basis of calculation we conclude that which one is best among them. Section 1 of this paper consist the introduction about the machine learning and heart diseases. Section II described, the machine learning classification. Section III illustrated the related work of researchers. Section IV is about the methodology used for this prediction system. Section V is about the algorithms used in this project. Section VI briefly describes the dataset and their analysis with the result of this project. And the last Section VII concludes the summary of this paper with slight view about future scope of this paper. The project's objective is to determine whether a patient is likely to be diagnosed with any cardiovascular disease based on attributes such as gender, age, chest pain, and fasting sugar level. The dataset, selected from the UCI repository, contains medical histories and attributes of patients. Using this dataset, we predict the likelihood of heart disease. The 14 medical attributes are trained using three algorithms: logistic regression, KNN, and Random Forest Classifier. KNN is the most efficient, providing an accuracy of 87.52% Ultimately, this system classifies patients at risk of heart disease in a cost-efficient manner.

II. MACHINE LEARNING

Machine Learning is one of efficient technology which is based on two terms namely testing and training i.e. system take training directly from data and experience and based on this training test should be applied on different type of need as per the algorithm required. There are three type of machine learning algorithms:

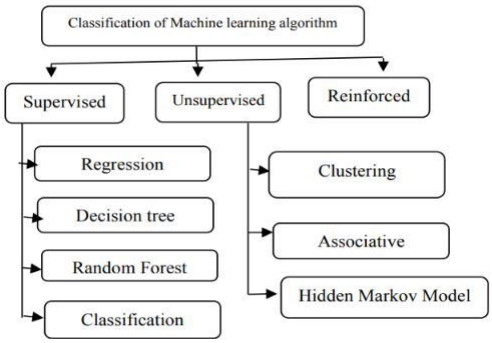


Fig. 1. Classification of machine learningF

III. RELATED WORK

Heart is one of the core organ of human body, it play crucial role on blood pumping in human body which is as essential as the oxygen for human body so there is always need of protection of it, this is one of the big reasons for the researchers to work on this. So there are number of researchers working on it .There is always need of analysis of heart related things either diagnosis or prediction or you can say that protection of heart disease .There are various fields like artificial intelligence, machine learning, data mining that contributed on this work. Performance of any algorithms depends on variance and biasness of dataset[4]. As per research on the machine learning for prediction of heart diseases himanshu et al.[4] naive bayes perform well with low variance and high biasness as compare to high variance and low biasness which is knn. With low biasness and high variance knn suffers from the problem of over fitting this is the reason why performance of knn get decreased. There are various advantage of using low variance and high biasness because as the dataset small it take less time for training as well as testing od algorithm but there also some disadvantages of using small size of dataset. When the dataset size get increasing the asymptotic errors are get introduced and low biasness, low variance based algorithms play well in this type of cases. Decision tree is one of the nonparametric machine learning algorithm but as we know it suffers from the problem over fitting but it cloud be solve by some over fitting removable techniques [14]. Support vector machine is algebraic and statics background algorithm, it construct a linear separable n-dimensional hyper plan for the classification of datasets.

The nature of heart is complex, there is need of carefully handling of it otherwise it cause death of the person. The severity of heart diseases is classified based on various meth-ods like knn, decision tree, generic algorithm and na¨ıve bayes

[3]. Mohan et al.[3] define how you can combine two different approaches to make a single approach called hybrid approach which have the accuracy 88.4% which is more than of all other.

Some of the researchers have worked on data mining for the prediction of heart diseases. Kaur et al.[6] have worked on this and define how the interesting pattern and knowledge are derived from the large dataset. They perform accuracy comparison on various machine learning and data mining approaches for finding which one is best among then and get the result on the favor of svm.

Kumar et al.[5] have worked on various machine learning and data mining algorithms and analysis of these algorithms are trained by UCI machine learning dataset which have 303 samples with 14 input feature and found svm is best among them, here other different algorithms are naivy bayes, knn and decision tree.

Gavhane et al.[1] have worked on the multi layer perceptron model for the prediction of heart diseases in human being and the accuracy of the algorithm using CAD technology. If the number of person using the prediction system for their diseases prediction then the awareness about the diseases is also going to increases and it make reduction in the death rate of heart patient.

Some researchers have work on one or two algorithm for predication diseases. Krishnan et al.[2] proved that decision tree is more accurate as compare to the na¨ıve bayes classifi-cation algorithm in their project.

Machine learning algorithms are used for various type of diseases predication and many of the researchers have work on this like Kohali et al.[7] work on heart diseases prediction us-ing logistic regression, diabetes prediction using support vector machine, breast cancer prediction using Adaboot classifier and concluded that the logistic regression give the accuracy of 87.1%, support vector machine give the accuracy of 85.71%, Adaboot classifier give the accuracy up to 98.57% which good for predication point of view[11]. A survey paper on heart diseases predication have proven that the old machine learning algorithms does not perform good accuracy for the predication while hybridization perform good and give better accuracy for the predication[8].

IV. DATA RESOURCE

Used in this research, the clinical heart disease data were from 303 patients at the Cleveland Clinic Foun-dation (CCF) located in Cleveland, Ohio in the United States. The dataset was obtained from the Heart Disease Database made available in the UCI Machine Learning Repository [15]. Each of the 303 clinical instances contained 75 attributes and a target attribute. The target attribute represented an integer valued from 0 to 4, signifying absence [0] or presence [1, 2, 3] of heart disease in patients. For this research, binary values of 0 and 1 were reassigned to the target attributes for the absence or presence of heart disease in patients, respectively. The dataset included 91 female patients (30.03%) and 212 male patients (69.97%), and their ages ranged from 29 to 77 years with the average being 54 years old of the 303 clinical instances from the Cleveland Clinic Dataset, 282 clinical cases were utilized and the remainder were excluded from the research due to missing data values. Of the 282 total clinical instances, 125 of the cases (44.33%) had heart disease while 157 were cases (55.67%) that were absent of heart disease. Each clinical instance was described with 76 raw attributes. However, only 29 of the raw attributes were utilized in the development of the deep neural network models due to missing values among the other raw attributes. Details regarding the 29 raw attributes are listed in Table 1. In the development of the deep neural network model, the entire data set of 282 total clinical instances was randomly separated into a training data set of 135 clinical instances (47.87%) and testing data set of 147 clinical instances (52.13%).

V. METHODOLOGY OF SYSTEM

Processing of system start with the data collection for this we uses the UCI repository dataset which is well verified by number of researchers and authority of the UCI [15].

A. Data Collection

The Cleveland heart dataset from the UCI machine learning repository has been used for the experiments. The dataset consists of 14 attributes and 303 instances. There are 8 categorical attributes and 6 numeric attributes. The description of the dataset is shown in the table 1. Patients from age 29 to 79 have been selected in this dataset. Male patients are denoted by a gender value 1 and female patients are denoted by gender value 0. Four types of chest pain can be considered as indicative of heart disease. Type 1 angina is caused by reduced blood flow to the heart muscles because of narrowed coronary arteries. Type 1 Angina is a chest pain that occurs during mental or emotional stress. Nonangina chest pain may be caused due to various reasons and may not often be due to actual heart disease. The fourth type, Asymptomatic, may not be a symptom of heart disease. The next attribute trestbps is the reading of the resting blood pressure. Chol is the cholesterol level. Fbs is the fasting blood sugar level; the value is assigned as 1 if the fasting blood sugar is below 120mg/dl and 0 if it is above. Restecg is the resting electrocardiographic result, thalach is the maximum heart rate, exang is the exercise induced angina which is recorded as 1 if there is pain and 0 if there is no pain, oldpeak is the ST depression induced by exercise, slope is the slope of the peak exercise ST segment, ca is the number of major vessels colored by fluoroscopy, thal is the duration of the exercise test in minutes, and num is the class attribute [10]. The class attribute has a value of 0 for normal and 1 for patients diagnosed with heart disease.

TABLE.1 Attributes of the Dataset

S. No.	Attribute	Description	Type
1	Age	Patient's age (29 to 77)	Numaric
2	Sex	Gender of patient(male-0 female-1)	Nominal
3	Cp	Chest pain type	Nominal
4	Trestbps	Resting blood pressure(in mm Hg on admission to hospital ,values from 94 to 200)	Numerical
5	Chol	Serum cholesterol in mg/dl, values from 126 to 564)	Numerical
6	Fbs	Fasting blood sugar>120 mg/dl, true-1 false-0)	Nominal
7	Resting	Resting electrocardiographics result (0 to 1)	Nominal
8	Thali	Maximum heart rate achieved(71 to 202)	Numerical
9	Exang	Exercise included agina(1-yes 0-no)	Nominal
10	Oldpeak	ST depression introduced by exercise relative to rest (0 to .2)	Numerical
11	Slope	The slop of the peak exercise ST segment (0 to 1)	Nominal
12	Ca	Number of major vessels (0-3)	Numerical
13	Thal	3-normal	Nominal
14	Targets	1 or 0	Nominal

Fig. 2. TABLE.1 Attributes of the Dataset

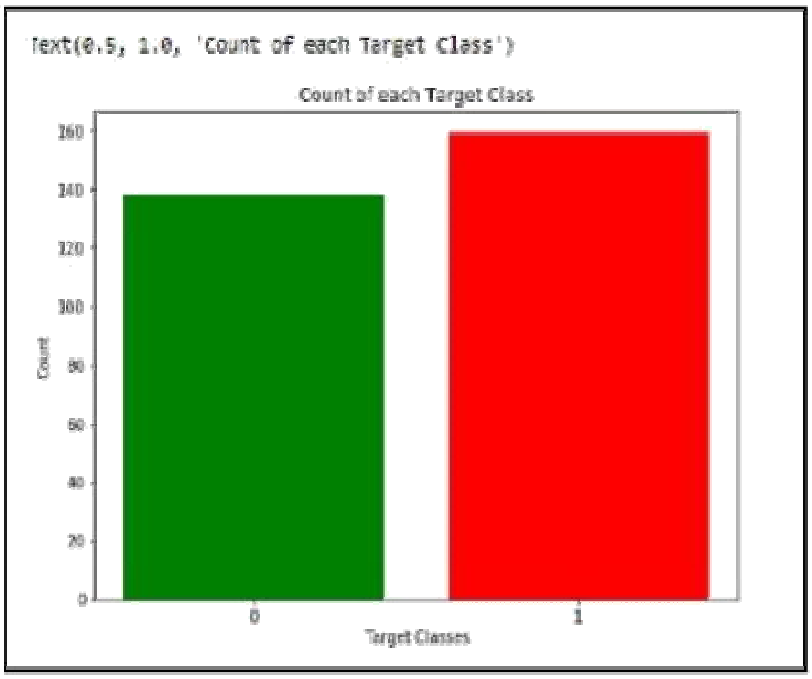


Fig. 3. Target class view

B. Attribute Selection

Attribute of dataset are property of dataset which are used for system and for heart many attributes are like heart bit rate of person, gender of the person, age of the person and many more shown in TABLE.1 for predication system.

C. Preprocessing of Data

Preprocessing needed for achieving prestigious result from the machine learning algorithms. For example Random forest algorithm does not support null values dataset and for this we have to manage null values from original raw data. For our project we have to convert some categorized value by dummy value means in the form of “0”and “1” by using following code:

D. Data Balancing

Data balancing is essential for accurate result because by data balancing graph we can see that both the target classes are equal. Fig.3 represents the target classes where “0” represents with heart diseases patient and “1” represents no heart diseases patients.

E. Histogram of Attributes

Histogram of attributes shows the range of dataset attributes and code which is used to create it. dataset.hist()

VI. MACHINE LEARNING ALGORITHMS

A. Decision Tree

On the other hand decision tree is the graphical repre-sentation of the data and it is also the kind of supervised machine learning algorithms.

B. Support Vector Machine (SVM)

For problems involving regression and classification, super-vised learning techniques like the Support Vector Machine (SVM) algorithm are employed. It operates by locating the hyperplane in the feature space that optimally divides the data points of various classes. Given its excellent accuracy and resilience, even with tiny datasets, support vector machines (SVM) are especially helpful for medical diagnosis.

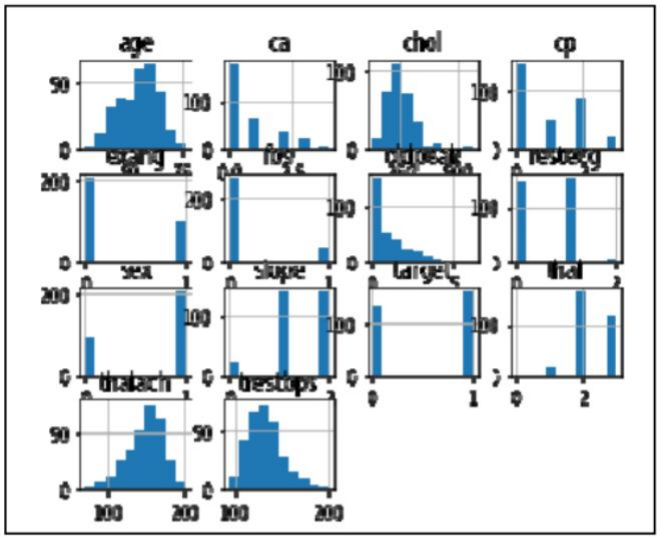


Fig. 4. Histogram of attributes

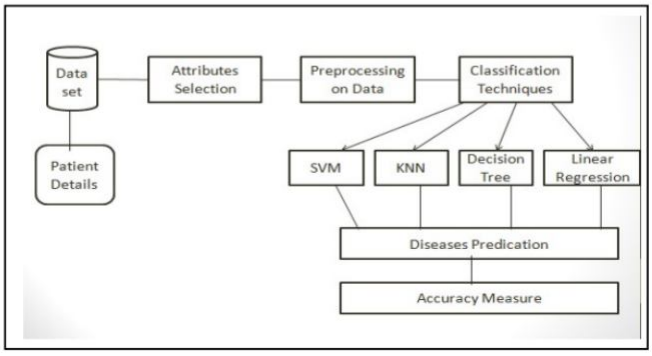


Fig. 5. Architecture of Prediction System

C. Logistic Regression for Classification

One statistical technique for binary classification problems is logistic regression. By utilizing a logistic function to estimate probabilities, it models the relationship between a dependent variable and one or more independent variables.

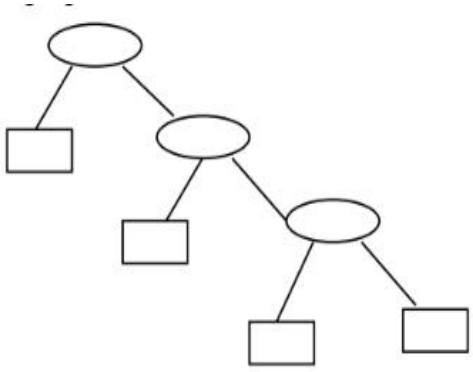


Fig. 6. Working of Decision Tree

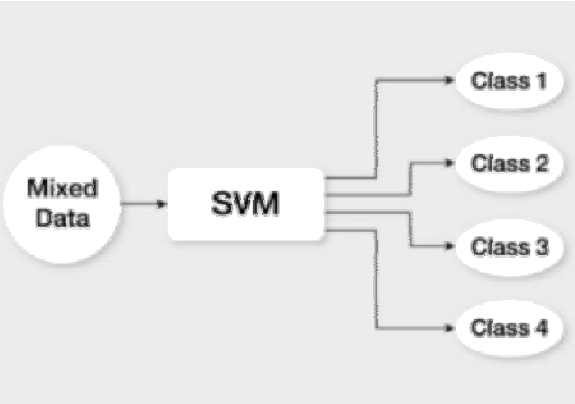


Fig. 7. Working of Support Vector Machine

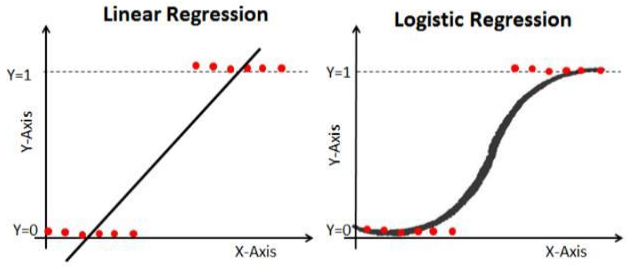


Fig. 8. Working of Logistic Regression

D. K-Nearest Neighbors (KNN) Algorithm

A supervised learning technique called K-Nearest Neigh-bors (KNN) is applied to regression and classification prob-lems. It functions on the tenet that comparable data points typically fall into the same class or have comparable numerical values. As an instance-based learning technique that is non-parametric and does not assume anything about the distribution of the underlying data, KNN saves training data instances for prediction as opposed to building a model [12].

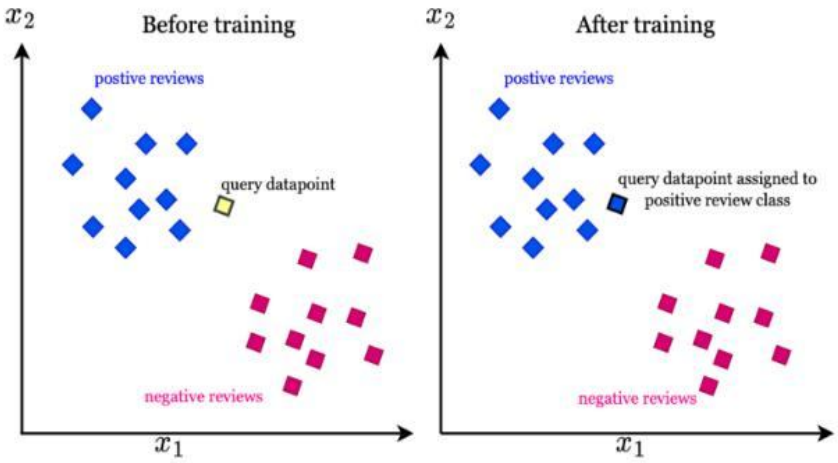


Fig. 9. Working Of KNN

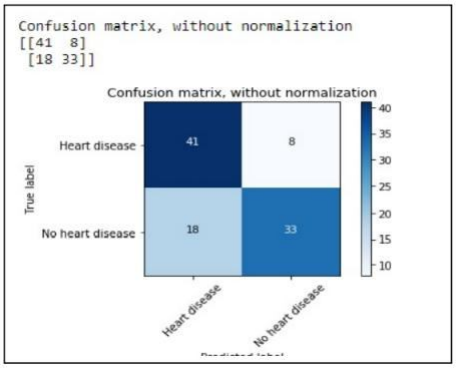


Fig. 10. Confusion matrix for Decision tree

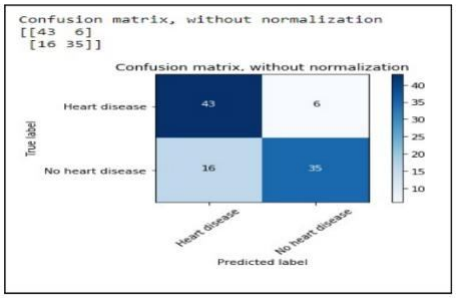


Fig. 11. Confusion Matrix for linear regression

E. Accuracy Calculation

Accuracy of the algorithms are depends on four values namely true positive(TP), false positive(FP), true negative(TN) and false negative(FN).

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

Where:

- TP = Number of persons with heart diseases
- TN = Number of persons with no heart diseases
- FP = Number of persons with no heart diseases but incorrectly classified as having heart diseases
- FN = Number of persons with heart diseases but incor-rectly classified as not having heart diseases

F. Result

After performing the machine learning approach for testing and training we find that accuracy of the knn is much efficient as compare to other algorithms. Accuracy should be calculated with the support of confusion matrix of each algorithms as shown in Fig.6 and Fig.7 here number of count of TP, TN, FP, FN are given and using the equation (2) of accuracy, value has been calculated and it is conclude that knn is best among them with 87% accuracy and the comparison is shown in TABLE.2

Algorithm	Accuracy
Support Vector machine	83%
Decision tree	79%
Linear regression	78%
k-nearest neighbor	87%

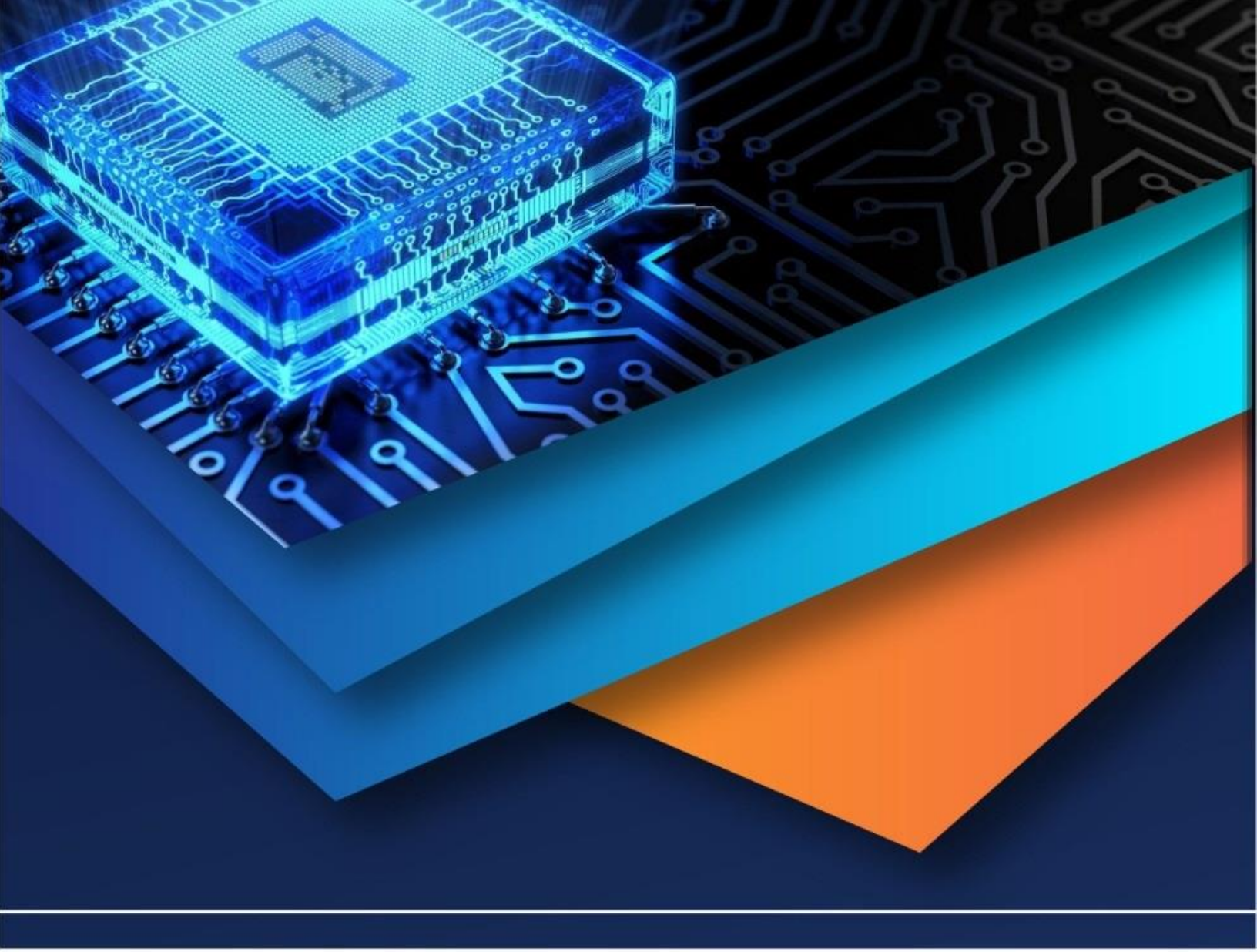
Fig. 12. TABLE.2 Accuracy comparison

VII. CONCLUSION AND FUTURE WORK

Heart is one of the essential and vital organ of human body and prediction about heart diseases is also important concern for the human beings so that the accuracy for algorithm is one of parameter for analysis of performance of algorithms [13]. Accuracy of the algorithms in machine learning depends upon the dataset that used for training and testing purpose. When we perform the analysis of algorithms on the basis of dataset whose attributes are shown in TABLE.1 and on the basis of confusion matrix, we find KNN is best one. For the Future Scope more machine learning approach will be used for best analysis of the heart diseases and for earlier prediction of diseases so that the rate of the death cases can be minimized by the awareness about the diseases.

REFERENCES

[1] Effective heart disease prediction using hybrid machine learning tech-niques, Mohan, Senthilkumar and Thirumalai, Chandrasegar and Srivas-tava, Gautam
[2] Heart disease prediction using machine learning techniques Sharma, Vijeta and Yadav, Shrinkhala and Gupta, Manjari
[3] Heart disease prediction using machine learning techniques: a quantita-tive review ,Riyaz, Lubna and Butt, Muheet Ahmed and Zaman, Majid and Ayob, Omeera
[4] Heart Diseases Prediction using Machine Learning,Yadav, Anup Lal and Soni, Kamal and Khare, Shanu
[5] M. Nikhil Kumar, K. V. S. Koushik, K. Deepak, "Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools" International Journal of Scientific Research in Computer Science, Engineering and Information Technology ,IJSRCSEIT 2019.
[6] Amandeep Kaur and Jyoti Arora,"Heart Diseases Prediction using Data Mining Techniques: A survey" International Journal of Advanced Research in Computer Science , IJARCS 2015-2019.
[7] Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison Ridwan Taiwo1*, Idris Temitope ,Ali, Md Mamun and Paul, Bikash Kumar and Ahmed, Kawsar and Bui, Francis M and Quinn, Julian MW and Moni, Mohammad Ali
[8] Classification models for heart disease prediction using feature selection and PCA rate-Escamila, Anna Karen and El Hassani, Amir Hajjam and
[9] A hybrid framework for heart disease prediction using machine learning algorithms ,Chandrika, L and Madhavi, Karanam
[10] Prediction of coronary heart disease using machine learning: an ex-perimental analysis Gonsalves, Amanda H and Thabtah, Fadi and Mohammad, Rami Mustafa A and Singh, Gurpreet
[11] A hybridized model for the prediction of heart disease using ML algorithms Naidu, T Penchala and Gopal, K Amar and Ahmed, Sk Rameez and Revathi, R and Ahammad, Sk Hasane and Rajesh, V and Inthiyaz, Syed and Saikumar, K, booktitle=2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)
[12] Heart disease prediction using machine learning and data mining, Srivastava, Keshav and Choubey, Dilip Kumar
[13] Heart disease prediction using hybrid machine learning model ,Kavitha, M and Gnaneswar, G and Dinesh, R and Sai, Y Rohith and Suraj, R Sai
[14] Implementation of a heart disease risk prediction model using machine learning, Karthick, K and Aruna, SK and Samikannu, Ravi and Kup-pusamy, Ramya and Teekaraman, Yuvaraja and Thelkar, Amruth Ramesh and others
[15] https://www.sciencedirect.com/science/article/pii/S0926580523004478



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY