iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** XII **Month of publication:** December 2020

DOI: https://doi.org/10.22214/ijraset.2020.32591

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 8 Issue XII Dec 2020- Available at www.ijraset.com

A Comparative Study of Machine Learning Algorithms for the Prediction of Heart Disease

Hidayatullah Arghandabi[1], Parvaneh Shams[2]
[1, 2]*Computer Engineering Department, Istanbul Aydin University*

Abstract: ***The heart is a significant organ of the human body. These days heart disease has become an important issue in humans, nearly one person dies per minute. To deal with this problem we need a strong prediction system for this condition. Machine learning is used to solve many data science problems. The traditional use of machine learning is to predict an output from input data. Here machine learning is applied to medical records to predict the diseases of each patient. The ML algorithms learn patterns from the given input data. And using them with real-life data to forecast the disease's existence. Some machine learning algorithms predict with great accuracy while others predict with insufficient accuracy. Here we predict heart disease patients in the early stage by interpreting the medical records accurately and compare the algorithm accuracy.***
In this research paper, we checked and compared various machine learning algorithms, such as decision tree, logistic regression, k-nearest neighbors, support vector machine, gradient boosting. We Used the UCI heart dataset and for the implementation Jupyter notebook and Python programming language. This research also aims to forecast the possibility of the patient's heart disease at an early stage, selecting the algorithm for this problem with the highest precision. The detection of the disease at an early stage can help to provide the patient with the necessary treatment and care.
Keywords: Machine learning, Heart disease prediction, Comparing algorithms, Machine learning algorithms, Testing algorithms

I. INTRODUCTION

The heart is an essential component of the body. It pumps blood and carries oxygen, nutrients and other resources to different areas of the body [1]. Therefore, this muscle system plays a crucial role in the body. Heart problems also affect the normal function of other components in the body. It is one of the key causes for the deaths of humans. They are the main responsible for one-third of all human deaths in the world [2]. Thus, it is very significant to diagnose the illness at an early stage. Illness diagnosis and providing effective treatment to the patient plays an important role in healthcare. Inadequate diagnosis causes harmful outcomes that are not acceptable [1]. The continuous development of technologies helps the researchers to develop new methodologies for prediction. Healthcare has collected a large amount of data and, which needs to process using certain mining techniques. Data mining is extracting information from a large amount of data. It is also an essential step in finding knowledge from databases [6]. Data-mining techniques can be implemented through machine learning algorithms.

There are various machine algorithms used for the prediction of the heart disease. Some algorithms give higher accuracy and some poor accuracy. Therefore, a comparative study about ML algorithms, becomes crucial to predict with high accuracy in the early stage. it is a crucial step to diagnose the illness at an early stage so that effective treatment is given to the patients.

Several Machine learning algorithms are used to detect the occurrence of the heart disease in patient at an early stage. K-nearest neighbors, logistic regression, support vector machine, decision tree and gradient boosting are the recent literature used ML algorithms for this problem. These algorithms take several inputs from the medical data. Effective datasets help the algorithms to predict the output with high accuracy. We use the UCI heart disease dataset. Up to this day machine learning developers have been using it, too. First, we train the ML algorithms to predict patients with heart diseases. After training the models, we tested the prediction accuracy of algorithms. According to their accuracy score, we selected the most efficient machine learning algorithm for this problem.

In this paper, we are mainly focused on discussing and comparing the machine learning algorithms models' performance that are used for prediction heart problem. We compare and implement 5 machine learning algorithms. The training and test environment are developed in Jupyter notebook in Python programming language.

This paper consists of 5 sections. Section 1 consists of introduction to the heart disease and machine learning. Section 2 describes the literature review, section 3 describes the experiments setup used in this research, section 4 describes the machine learning algorithms and the result of training and testing, section 5 shows results and evaluation, and section 6 describe conclusion and future scope.

II. LITERATURE REVIEW

Heart disease has been the largest cause of human deaths in developed and non-developed countries [3]. The main reason for these deaths is the risks that are not recognized or recognized at a later stage of the disease. The presence of risk and importance of this organ makes it clear that it needs care and protection. Machine learning algorithms can be useful to overcome this problem. Machine learning gives effective results in various real-life problems. Healthcare is an important application area because it has large data stores, which are challenging to handle manually [4]. Many researchers use machine learning algorithms for heart disease prediction problem. Chu-Hsing Lin et al. [7] have compared Convolutional Neural Network and regular Neural Network to predict heart disease, using the Cleveland data set. They have found that the CNN model was more stable with 80% accuracy in heart disease prediction. By tuning the neurons and number of layers they reach 93% accuracy. The result shows that shows CNN model performance is better than the NN model. Some researchers have used machine learning recent literature algorithms. B. Keerthi Samhitha et al. [8] have explained that ML algorithms rely on current data. They have checked the performance of various ML algorithms used for heart disease patient detection. In result, they find higher accuracy in K-Nearest Neighbor's model.

Elzhan Zeinulla et al. [9] have worked and developed a framework for the incomplete medical data, to diagnose heart disease patients. They have explained that medical data are dirty and incomplete. They utilize the Synthetic minority over sampling technique. They have trained the ML recent literature algorithms and since they applied data processing and engineering to eliminate the incomplete entries, as result, they get 93.4% accuracy in Fuzzy Random Forest model.

Riddi Kasabe et al. [10] have evaluated different classification techniques to diagnose heart disease in patients. Before the experiment implementation the data set was pre-processed to prepare for the ML algorithms, as the result they find higher accuracy in Random Forest model. Montu Saw et al. [11] have explained the improvement of logistic regression model to predict the heart disease using the healthcare data for training and testing. They have also described the importance of heart patients caring the in early stages. The performance of model was check by the Confusion matrix. As result they get 80% accuracy in logistic regression model. Noor Basha et al. [12] used various ML algorithms to predict and analyse the heart disease presence in patients in early stage of illness. They compared the ML algorithms in the recent literature, as result, they get 85% accuracy with KNN, it was selected as the efficient model among others. Rahul Katarya et al. [13] describes that some supervised ML algorithms can be used for prediction of heart problem. They have proposed a used of recent algorithms in the literature, to create an automated system for the patient of heart attacks. As result, they have examined that features should be chosen carefully in order to get a better result from the ML models and summarized the use of the Hybrid Grid Search algorithm for such a case.

Laxin Miao et al. [14] created a risk analysis mode for heart disease prediction using the support vector machine, and k-nearest neighbors ML algorithms. They have balanced the data before the training process. As result they found a high accuracy of 96.5% in the support vector machine model in a 1000 times training iteration. Halima El Hamdaoui et al. [15] developed a system for the healthcare to make a better decision. They used the recent machine learning literature algorithms. As result they found 84.28% accuracy in Naïve Bayes. Tulay Karayilan et al. [16] explains the importance of heart disease diagnosis at early time. They have proposed a model trained by an artificial neural network (ANN) and they have found 95% accuracy in prediction.

A survey paper on the diagnosis of heart attacks shows that traditional ML algorithms don't give good accuracy while hybridization gives good results [17].

III.METHODOLOGY

A. Dataset

The dataset of UCI for heart disease is used in this research. It is an important dataset and widely used by machine learning researchers. We use this dataset to train and test the ML algorithms. As the result, we aim to find the presence of heart disease in patients. We have used 73% of dataset for training and 37% for testing. The dataset consists of 14 attributes, that are described in the table 1.

B. Architecture Diagram of Prediction System

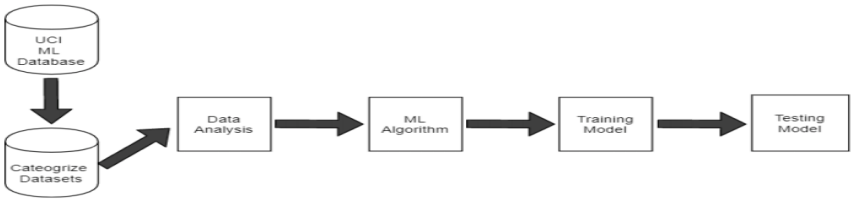


Fig. 1.Architecture diagram

TABLE I
Attributes of Dataset

S. No.	*Attribute*	*Description*	*Values*
1	Age	Age of patient's in years	20 - 77
2	Sex	Gender of patient (1: Male, 0: Female)	0, 1
3	Cp	Chest pain type	1, 2, 3, 4
4	Trestbps	Resting blood pressure in mm Hg	94 - 200
5	Chol	Serum cholesterol in mg/dl	126 - 264
6	Fbs	Fasting blood in mg/dl sugar>120 then 1(true): else 0 (false)	0, 1
7	Resting	Resting electrocardiographic Result	0, 1, 2
8	Thalach	Maximum Heart Rate Achieved	71 - 202
9	Exang	Exercise Included Angina (1:Yes, 0:No)	0, 1
10	OldPeak	ST Depression Introduced by Exercise Relative to Rest	1 - 3
11	Slope	Slope of the Peak Exercise ST Segment (1:up-sloping, 2:flat, 3: down-sloping)	1, 2, 3
12	Ca	Number of Major Vessels	0 - 3
13	Thal	3- Normal, 6-Fixed Defect, 7-Reversible Defect	3, 6, 7
14	Target	Presence heart disease (0: No, 1, 2, 3 4:Yes)	0, 1, 2, 3, 4

C. Data Analysis

We import the data to the Jupyter notebook and visualize the correlation between attributes to understand the importance of each feature using the heat map. We observed that the main column target is positively correlated with the age and negatively correlated with ca, old peak as shown in figure 3.

From the heat map, we observed that age is an important feature concerning the target. It has a positive correlation with the target.

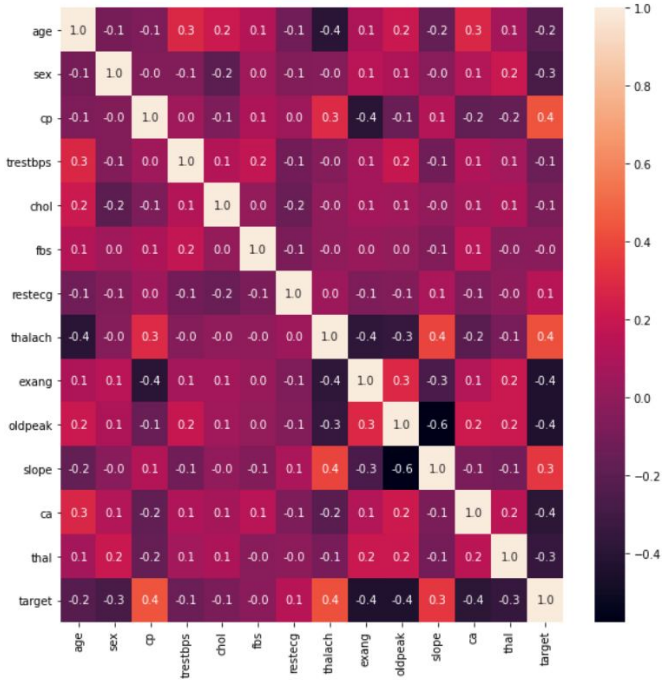


Fig. 2. Correlation of Attributes

We check the target value in the dataset, and we found that 165 patients have heart diseases and 138 patients have no heart disease as shown in figure 4. For a better analysis, we divided age into three categories such as young age (29-40), middle-age (40-55), elder age (55-...). According to this category, we have 16 young ages, 128 middle ages, and 159 elder ages patients.

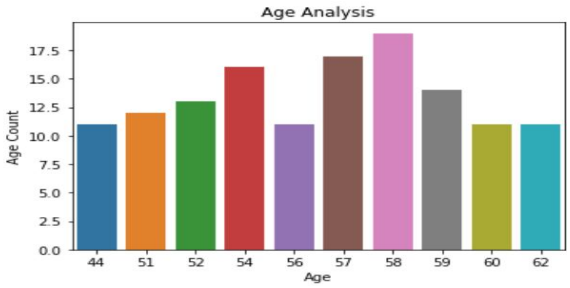


Fig. 3. Age Analysis of Patients

According to figure 3, we observe, most patients with presence of heart disease belong to age 58 and 57. Therefore, patients relating to age of higher than 50 are suffering from heart disease. Later, we observed the representation of sex and age for every target class. We found that 31.68% of patients are female and, 68.32% male. As result, we predicted that heart disease is more in male patients.

IV. MACHINE LEARNING ALGORITHMS

A. *K-Nearest Neighbors*

It is a supervised classification ML algorithm. It predicts the output using the similarity of the data provided in the training [5]. It compares input data with the features of existing data. This algorithm compares unclassified data with classified data by calculating the Euclidean, Manhattan, or Minkowski distance between the feature points. As in the example below:

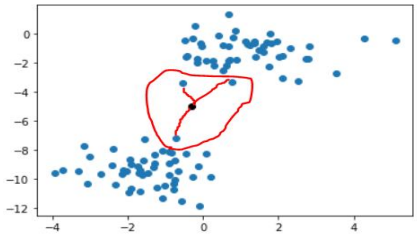


Fig. 4. KNN and k=3

In the above picture, we have a black dot and k =3 we should calculate the distance of the 3 nearest points. If the calculated distance is close to data at the top, then this point belongs to data at the top, or else it belongs to data at the bottom. KNN is also called a lazy learner, therefore we have the chance to change the parameter's weight. We may assume some parameters more important than others. We choose the best k point which will also help in the optimization of this problem.

B. *Logistic Regression*

It is an ML algorithm, similar to linear regression. It creates an equation with one or more input variables (x), which predicts an outcome for the binary variable (y). This equation also shows the relationship between the input variable and the output binary variable. Output variable is 0 or 1 for two categories [18]. For example, the prediction of an email in category output of whether it is spam (1) or not (0).

Logistic regression does the estimation for an event to happen based on the given data. It works with binary data (1, 0), therefore if an incident occurs 1 if not 0.

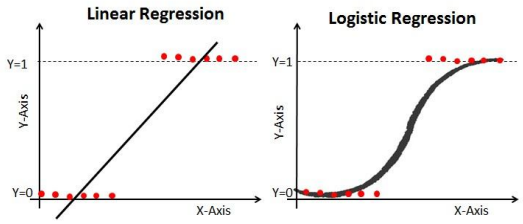


Fig. 5. Linear, Logistic Regression (Sigmoid function Applied)

C. Support Vector Machine

It is a supervised ML algorithm. It is mostly used for the classification problems. It works on the principle of hyperplane; it aims to classify the given data by building a hyperplane [9]. It implements a hyperplane in an n-dimensional space to distinguish various classes. In result an optimal hyperplane is implemented, in an iterative way in order to reduce the errors. The aim in here is to create a hyperplane with maximum margins which separates the dataset efficiently.

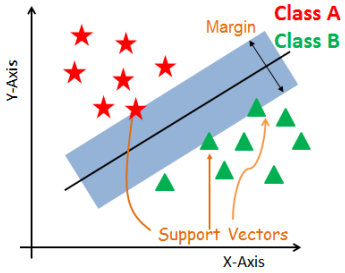


Fig. 6. Support Vector Machine

D. Decision Tree

It a supervise ML algorithm. It is a tree created from three nodes such as chance, decision and the end node [21]: Chance node display the possible output of a certain node. Decision node display the output-based decision by that node. The final result of a path is the end node.

The decision tree begins from the root node and it is divided into several nodes. The root separation happens based on the possibilities. Each node collects information about data characteristics. And every link shows a decision taken on the node [21].

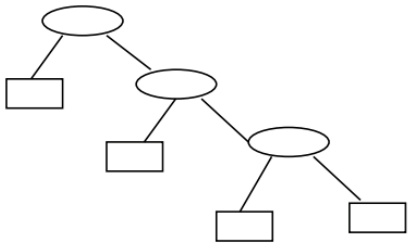


Fig. 7. Decision Tree

Data's entropy is used to construct the tree and drawn the nodes. It is calculated by the following equation (1), Pij is probability of each node in the tree.

$$Entropy = -\sum Pij \ \ logPij \quad (1)$$

The node with the maximum entropy calculated, is chased the root node. This process is repeated in order to construct the tree completely. It is important to note that decision tree has problem with overfitting [5].

E. Gradient Boosting

It is a powerful machine learning algorithm. The idea of boosting comes from the modification of a weak learner to become better. This algorithm has shown efficient performance in various applications [22]. Gradient boosting is learning by optimizing the loss function, it used the gradient descent method. It uses two types of base estimators, the average type model and decision tree with full depth. It is utilized for classification and regression.

V. EVALUATION AND RESULT

A. Evaluation

We compare the performance of algorithm using the confusion matrix. It displays accurate and inaccurate predicted outcome by the algorithms. Accuracy of an algorithm depends on 4 values explained bellow:

1) *TP:* The number of heart attack patients (correctly predicted)
2) *TN:* The number of patients with no heart attack (correctly predicted)
3) *FN:* The number of patients with no heart attack (incorrectly predicted)
4) *FP:* The number of patients with heart attack (incorrectly predicted)

$$Accuracy = (FN+TN) / (TP+FP+TN+FN) \quad (2)$$

$$Accurately\ Predicted\ Patients = TP+TN \quad (3)$$

$$Inaccurately\ Predicted\ Patients = FP + FN \quad (4)$$

The sum of TP and TN indicates the number of patients found accurately and the sum of FP and FN is the number of patients found inaccurately by the algorithm, shown in equation (3) and (4).

B. Result

In result of train and testing on the machine learning algorithms, we found high accuracy for knn 85.7%. For this problem this algorithm is efficient others. We calculated the accuracy using the confusion matrix for each algorithm. Accuracy is explained in the equation (2). The aim of this work was to compare the machine learning algorithms accuracy and find the patients in early stage using the efficient algorithm. The accuracy scores are shown in the figure 8.

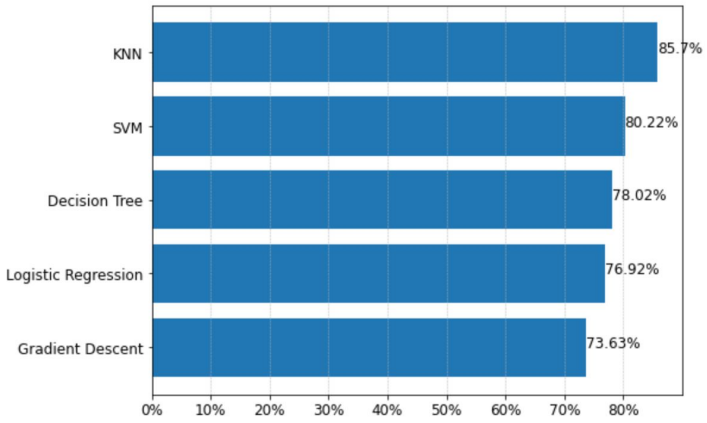


Fig. 8. Accuracy Performance of Machine Learning Algorithms

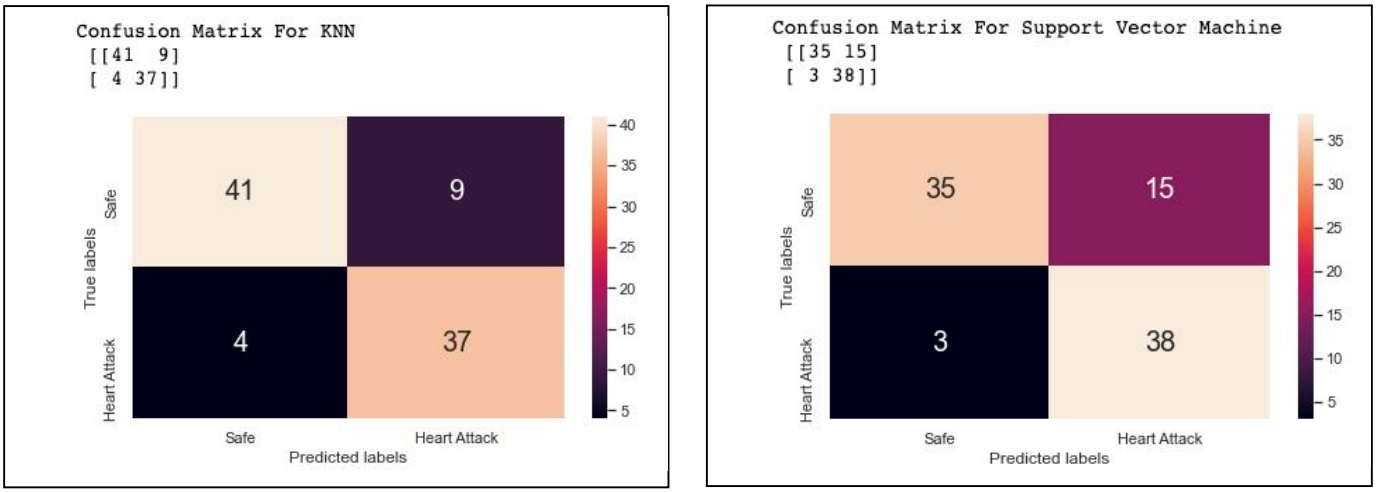


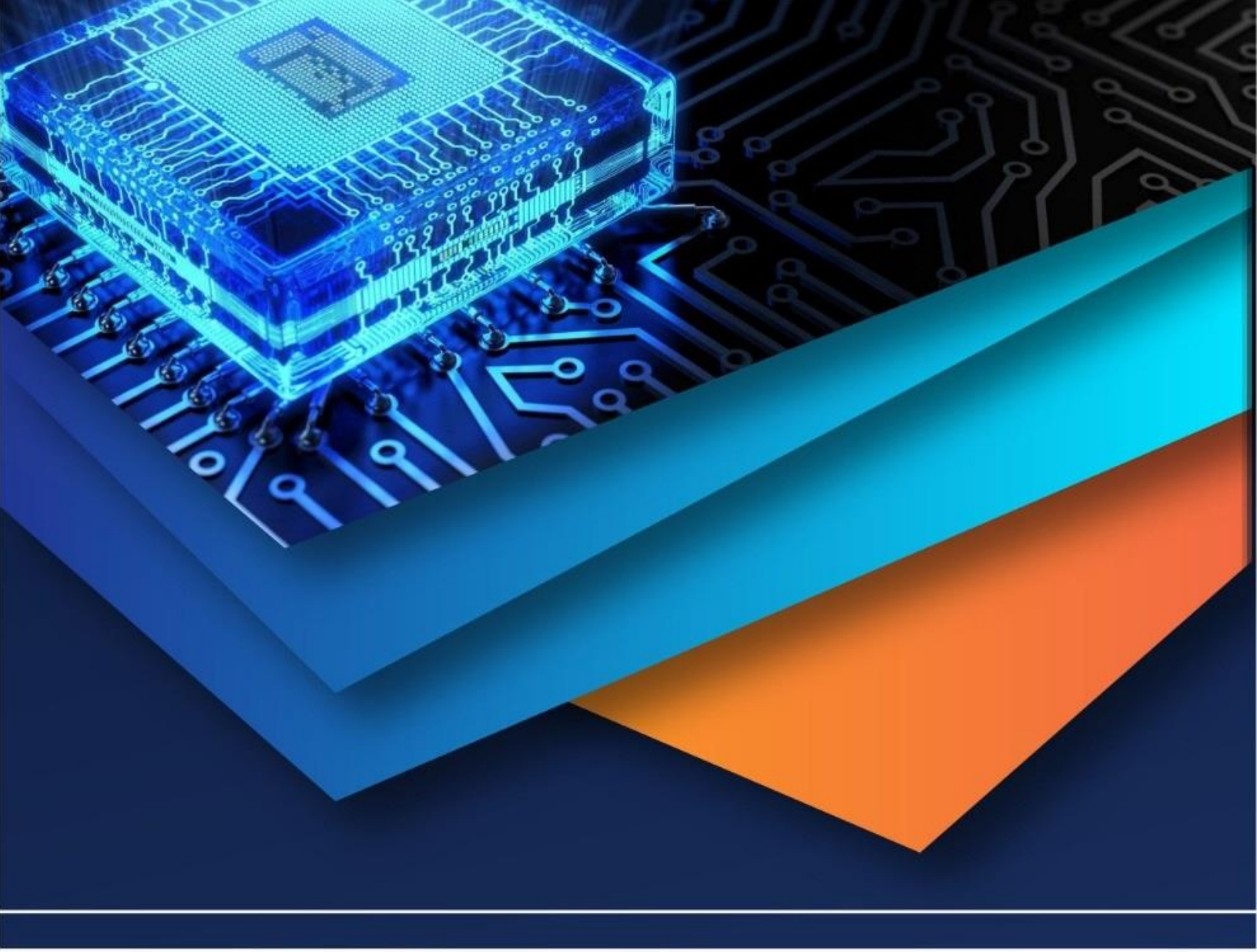
Fig. 9. KNN Algorithm and SVM Algorithm Confusion Matrix

VI. CONCLUSION AND FUTURE SCOPE

In this study, we did a comparative study on several machine learning algorithms to predict heart disease patients efficiently in an early stage. Some algorithms show good accuracy whereas some others poor accuracy. We see that the highest accuracy was achieved by k-nearest neighbor. Our model can be used by hospitals and it will help them to predict heart disease patients from their datasets. The model can be further trained on different hospital data's and get high-quality outcome.

REFERENCES

[1] K.Sudhakar, Dr. M. Manimekalai, "Study of Heart Disease Prediction using Data Mining", IJARCSSE, 2016.
[2] KaanUyar Ahmet Ilhan. "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. "9th international conference on theory and application of soft computing, computing with words and perception. Budapest, Hungary: ICSCCW; 2017. 24-25 Aug 2017.
[3] Vanisree K, JyothiSingaraju. "Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks. "Int J Comput Appl April 2011;19(6). (0975 8887).
[4] C. Beulah Christalin Latha, S. Carolin Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," Informatics in Medicine 16 (2019) 100203.
[5] Divya Krishnani, Anjali Kumari, Akash Dewangan, Aditya Singh, Nenavath Srinivas Naik, "Prediction of Chronic Kidney Disease Using Data MiningPrediction of Coronary Heart Disease using Supervised Machine Learning Algorithms", IEEE Region 10 TENCON Conference, 2019.
[6] Mr.Santhana Krishnan.J, Dr.Geetha.S, "Survey on Current Trends and Techniques of Data Mining Research,", London Journal of Research in Computer Science and Technology, Volume 17, Issue1,pages: 7-15, 2017.
[7] Chu-Hsing Lin, Po-Kai Yang, Yu-Chiao Lin, and Pin-Kuei Fu, "On Machine Learning Models for Heart Disease Diagnosis", ISBN: 978-1-7281-8712-9, Second IEEE Eurasia Conference on Biomedical Engineering Healthcare and Sustainability, 2020.
[8] B.Keerthi Samhitha, Sarika Priya.M.R, Sanjana.C, Suja Cherukullapurath Mana and Jithina Jose, "Improving the Accuracy in Prediction of Heart Disease using Machine Learning Algorithms", International Conference on Communication and Signal Processing, 28-30 July 2020.
[9] Elzhan Zeinulla, Karina Bekbayeva, and Adnan Yazici, "Effective diagnosis of heart disease imposed by incomplete data based on fuzzy random forest", 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 19-24 July 2020.
[10] Riddhi Kasabe and Prof. Dr. Geetika Narang, "Heart Disease Prediction using Machine Learning", International Journal of Engineering Research & Technology (IJERT), Volume: 9 Issue: 8, Aug. 2020.
[11] Montu Saw, Tarun Saxena, Sanjana Kaithwas, Rahul Yadav, Nidhi Lal,"Estimating of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning", 2020 International Conference on Computer Communication and Informatics, 22-24 Jan. 2020.
[12] Noor Basha, Ashok Kumar P S, Gopal Krishna C and Venkatesh, "Early Detection of Heart Syndrome Using Machine Learning Technique" 4th International Conference on Electrical, Eletronics, Communication, Computer Technologies and Optimization Techniques, 13-14 Dec. 2019.
[13] Rahul Katarya. and Polipireddy Srinivas.,"Predicting Heart Disease at Early Stages using Machine Learning: A Survey", 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2-4 July 2020.
[14] Lanxin Miao, Xuezhou Guo, Hasan T Abbas, Khalid A Qaraqe, and Qammer H Abbasi, "Using Machine Learning to Predict the Future Development of Disease", 2020 International Conference on UK-China Emerging Technologies (UCET), 20-21 Aug. 2020.
[15] Halima EL HAMDAOUI, Saïd BOUJRAF1, Nour El Houda CHAOUI, Mustapha MAAROUFI, "A Clinical support System for Prediction of Heart Disease using Machine Learning Techniques", 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), 2-5 Sept. 2020.
[16] Tülay Karayilan and Özkan Kiliç, 2017 International Conference on Computer Science and Engineering (UBMK), 5-8 Oct. 2017.
[17] M. Akhil, B. L. Deekshatulu, and P. Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," Procedia Technol., vol. 10, pp. 85–94, 2013.
[18] Murat KORKMAZ, Selami GÜNEY, Şule Yüksel YİĞİTER, "The Importance Of Logistic Regression Implementations In The Turkish Livestock Sector And Logistic Regression Implementations/Fields", HR.Ü.Z.F. Dergisi, 16(2), Page:25-36, 2012.
[19] Divya Krishnani, Anjali Kumari, Akash Dewangan, Aditya Singh, Nenavath Srinivas Naik, "Predicition of Chronic Kidnewy Disase Using Data Mining Predicitons of Coronary Heart Disease Using Supervised Machine Learning Algorithms", IEEE Region 10 TENCON Conference, 2019.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY