Heart Disease Prediction Using Various Algorithms of Machine Learning

Rati Goel

Computer Science and Engineering
Inderprastha Engineering College
rati.python@gmail.com

Abstract—
Heart acts a major role in corporeal organism. The diseases of heart wants more perfection and exactness for diagnose and analyses. Heart disease is dangerous disease. This disease occurs due to various problems such as over pressure, blood sugar, high blood pressure, Cholesterol etc. in human body By using Python and machine learning, this paper is analyzed and predicted of the heart disease.. We can predict this disease by using various attributes in the data set. We have collected a data set consists of 13 elements and 383 individual value to analyze the patients performance. The main aim of the paper is to get a better accuracy to detect the heart disease using ML algorithm.

Keywords—; Python programming, SVM, KNN, Naïve Bayes, confusion Matrix, Jupyter-Notebook, supervised, unsupervised

1. Introduction

The most vital and essential part (organ) of human body is Heart. There are many disease that are linked to heart so The analysis of prediction of heart must be accurate. To resolve this, virtual study about this field obligatory. Normally these diseases predicted at end stage and this is the main reason of death of heart patient due to deficiency of correctness because of this there is requisite to identify about proficient algorithms for diseases prediction [1].
One of the proficient capable and effective technologies is Machine Learning that is established on specifically training and testing with the support of python and python libraries. Method acquires training directly form data and skill, based on this training, testing should be done on various type of need as per requisite algorithms. For Testing and Training, Machine learning can be used as an effective technology. It is belongs to AI (Artificial Intelligence). AI has one of its branches as machine learning. The work which is done by human using human intelligence that work can be done using machine learning technology. To enable human intelligence features in ML, The ML technology is equipped with different process to make use of data. As ML description, it absorbs from the ordinary phenomenon. By using python libraries with the algorithms of machine learning, this prediction has to be done [2]. In this detection, elements of biological is used such as chest pain (cp), sex, blood pressure(bp), cholesterol(chol). By using of these elements six algorithms of ML such as SVM, KNN, Decision Tree, Random Forest, Naïve Bayes, Ligistic Regression is applied for prediction of analysis and conclude that which technique is best on the basis of confusion matrix [3].

Section I contains the introduction of heart diseases and about the machine learning Section II explained the related work about prediction of heart diseases. Section III provides about the methodology of prediction system. Section IV illustrated various algorithms of ML. Section V provides experimental result and analysis of this project. And the last Section VI presents conclusion.

2. RELATED WORK

Kohali et al.[4] effort on extrapolation of many medical diseases such as cardiovascular diseases, breast cancer diabetes prediction by using various machine learning algorithms that creates different range of accuracy.

A.Mishra, et al.[5] proposed the strategy and concept that how the furthermost appropriate problems of MIP (medical image processing) will be acknowledged and evaluated in the framework of improving assessment of MIP solutions.

Sneha A, Mane et al.[6] implemented the two neural network algorithms such as learning vector quantization and function of radial basis that are used for diagnosis purpose. The comparison is made also between the two algorithms using MATLAB software. The major purpose is to find the best tool for medical analysis to shrink analysis time and increase efficiency with accuracy.

Tijjani et al.[7] shared the brief of the ANN based approaches to predicting problem of kidney through comparing mental behavior of the patient using MATLAB software..

Gavhane et al.[8] proposed a technique by using multilayer perceptron model for heart diseases prediction and create exactness using CAD technology.

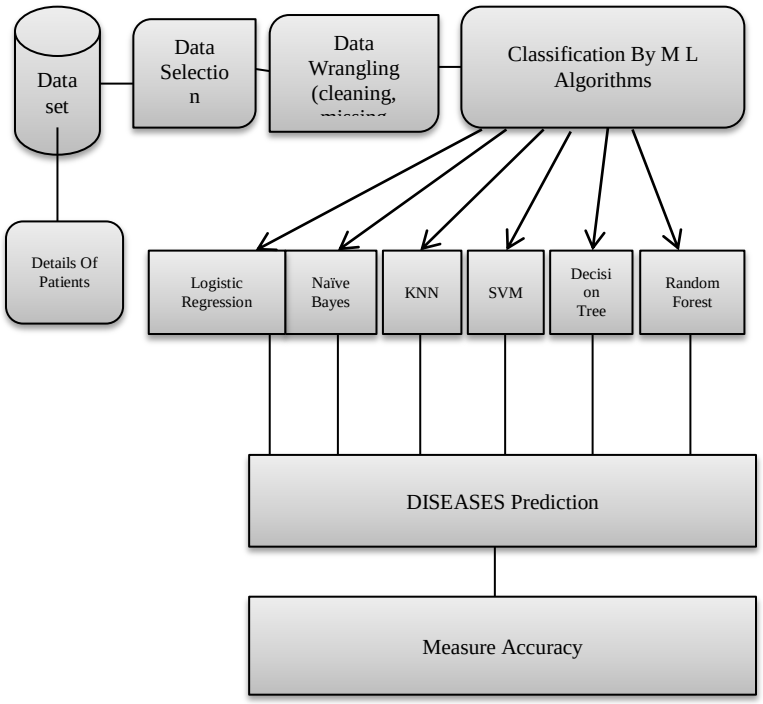
Kaur et al.[9] performed on large data set of heart disease. They compared on data mining approach and various algorithm of machine learning and finding perform accuracy comparison on various machine learning and data mining approaches for getting best accuracy.

Mohan et al.[10] explained that heart disease prediction need to be done very wisely and worked on different algorithms such as naïve bayes,generic algorithm, decision tree and knn. They also proposed hybrid algorithm and get 88% accuracy by applying hybrid algorithm.

Himanshu st al.[11] briefly discussed about large data set and small data set of heart diseases prediction. They shared that small data set take minimum time for traing as well as testing and performed prediction using SVM and knn algorithm. Discussed about prediction of heart diseases and prove that the some algorithms of machine learning does not better perform for accurateness predication though create good accuracy by hybridization [12]

3. Proposed Mehodology

Processed methodology start with the collection of data for this download tha data from kaggle that is well verified by researchers. In This methodology, There are many steps as shown in block diagram Fig 1.



3.1 Data Collection

First step for predication system is data collection, data collect from net [12] after that used data wrangling for data cleaning and then determining about the training and testing dataset. In this project we have used 80% training dataset and 20% dataset used as testing dataset the system. After Cleaning, In this data set There are 14 columns and 383 rows created by code (data.info). 5 row of table is created when use code (data.head()) as shown in Fig. 2

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Fig 2. Dataset Info

3.2. Element Selection

Dataset Elements are property of dataset that are used for analysis and prediction. There are many elements such as sex, age, slope and many more that are shown in TABLE.1 for system analysis.. .Block Diagram of Prediction System is .displayed in Fig.2.

Table 1.Elements of dataset

SNO	EXPLANATION	ELEMENTS
1	PATIENT's AGE	age
2	MALE, FEMALE	sex
3	CHEST PAIN	Cp
4	REST BLOOD PRESSURE	trtbps
5	COLESTROL	chol
6	FASTING BLOOD SUGAR	fbs
7	REST ELECTROCARDIOGRAPH	Restecg
8	MAX HEART RATE	thalachh
9	EXERCISE_INDUCED ANGINA	Exng
10	ST. DEPRESSION	Oldpeak
11	SLOPE	Slp
12	NO. OF VESSELS	Caa
13	THALASSEMIA	Thall
14	OUTPUT(Heart disease Patient	Output

3.3. Preprocessing of data

Preprocessing needed for achieving prestigious result from the machine learning algorithms.

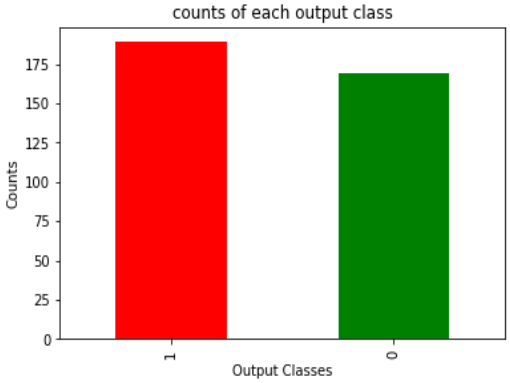


Fig.3 Target class view

Some ML algorithm does not support missing values for this we have to manage null values from original raw data. Some attribute of data set have been detected that is not useful for prediction such as education city etc.. Fig 3 shows green color [0] bar represents non disease patient and red color bar [1] presents heart disease patient. This diagram has created by the code ()

3.4 *Histogram of Elements*

Elements shows the range of dataset attributes. By using code [dataset.hist()], there are various histogram created in a frame

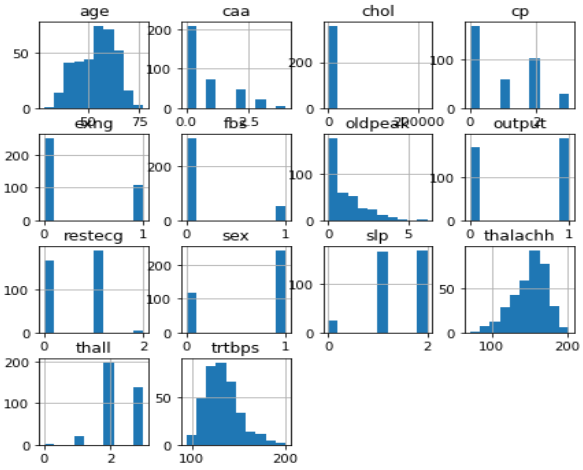


Fig.4 Histograms of elements

4. MACHINE LEARNING ALGORITHMS

4.1. Naïve bayes[NB]

NB is a supervise classification algorithm. It is a simple technique using Bayes theorem. To get the probability, mathematical concept is used with the support of bayes theorem. The correlation is neither related to each other nor predictor to one another. All parameters work autonomously for getting the maximum probability.

$$P(x/y) = \frac{p(y/x) \times p(x)}{p(y)} \qquad (1)$$

Where p(x)=Class predictor probability,
p(y)= Predictor Probability, P(x/y)= Posterior probability,
P(y/x)=possibility, probability of predictor

4.2. Decision Tree[DT]

DT is an algorithm that classifies parameters in categorical form in spite of arithmetic data. Tree like structure is created by DT. Many large data set related to medical have analyzed by DT due to its simple nature. It works on tree node for analysis.
Leaf Node: Signify the solution of every Test Interior Node:
Handle numerous element Main Node[Root Node]: Other nodes work based on main node
Data is to be divided into two or more parallel set by applying this algorithm. Then entropy of each parameter is calculated. After that divide the data with predictor having extreme information gain that means minimum entropy

$$\text{Entropy} = -\sum_{i,j=0}^{Ng-1} p(i,j)\log(p(i,j)) \qquad (2)$$

4.3. Random forest [RF]

RF algorithm is supervised primarily based learning. It is used as classifier in numerous fields. By using this more trees makes a forest. If we have more number of trees then it create higher accuracy. It is also used for regression task. but it accomplish well when classify the task. And may overwhelmed misplaced values. There are three approach of RF:
Forest RC(Random Blend)
Forest RI(Random input)
And combination of RC and RI

4.4. Logistic regression [LR]

LR is the supervised ML learning method. It is established on the association between dependent and independent variable as seen in Fig.5 variable "a" and "b" are dependent variable and independent variable and relation between them is shown by equation of line which is linear in nature that why this approach is called linear regression.

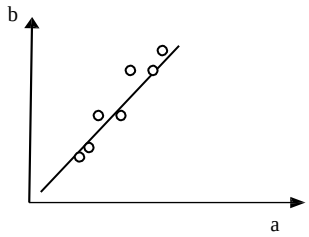


Fig.5 Relation between a and b

It gives a relation equation to predict a dependent variable value "b" based on a independent variable value "a" as we can see in the Fig.5 so it is concluded that linear regression technique give the linear relationship between a(input) and b(output).

4.5. Support Vector Machine

SVM is one type of ML method that work on the conception of hyper plan. It is used to find a hyper plan in n dimensional space, using this data point can be classified specifically [13].

(Xa, Ya) is training sample of data set where a=1,2,3,…….n and Ya is the target vector and Xa is the ith vector. Hyper plan quantity select the variety of support vector such as example if a line is used as hyper plan then method is called linear support vector.

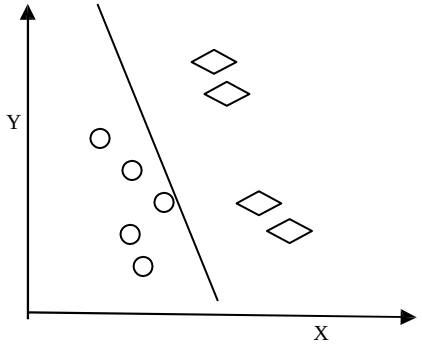


Fig.6 Linear Regression

4.6. K-nearest Neighbor [KNN]

KNN is a classification algorithm that belongs to supervise learning. It categorizes the entity that reliant on nearest neighbor. KNN could be a wide applied methodology used as a classifier and regression in numerous field like image process, data processing, pattern recognition and different applications. The output result of the algorithmic program depends on K-nearest neighbor class that enforced by finding K- variety of coaching points nearest to the specified character and contemplate the votes among the K object. The algorithmic program is incredibly easy. However, is capable of learning highly-complex non-linear call boundaries and regression functions [14]. The intuition of KNN that similar instances ought to have similar category labels (in classification) or similar target values (regression). On the drawback, the algorithmic program is computationally high-priced, and is vulnerable to over fitting.

5. Result Analysis

5.1. About Jupytor Notebook

Jupyter notebook is the tool for simulation of programming and open source network application. By using this, we can work with python programming.

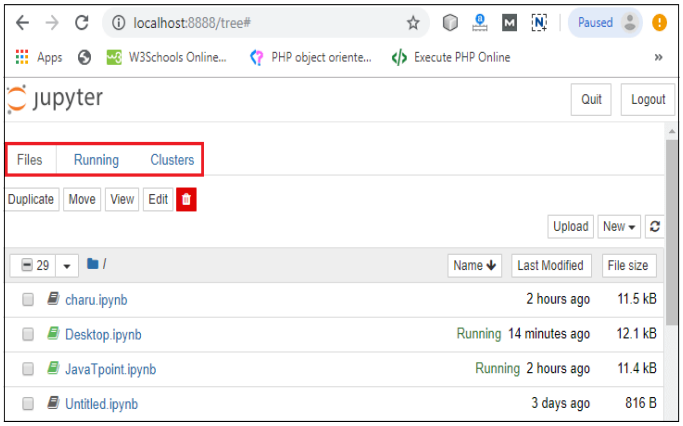


Fig.8.Display Of Jupyter Notebook

It is very comfortable tool .It contains coding, scripting language elements that have links, equations, figures, plots and many more. By importing various libraries of python programming, work on large dataset and analysis and visualize with different graphs of data in real time. With jupyter notebook, data cleaning , numerical imitation, statistical modeling and many work have to be done.

5.2. Confusion Matrix

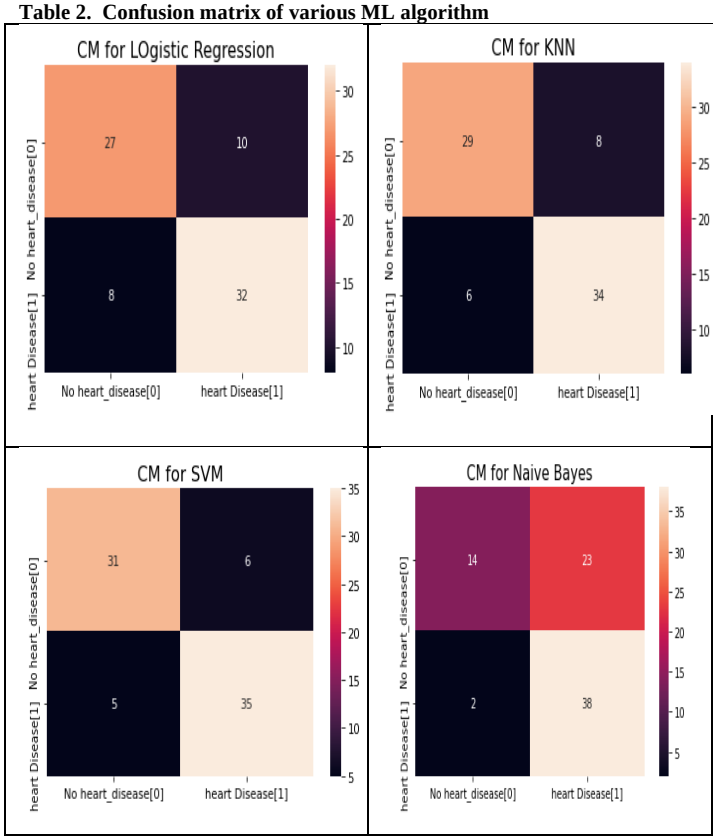
The confusion matrix is created by various classifier and it includes expected and real classifications information. The confusion matrix provides analysis to judge the effectiveness of proposed methodology [15].
Where,
The number of actual negative cases in the data = Condition Negative (N)
Condition Negative (N) = Total number of negative cases
Condition Positive (P) = Total number of positive cases
True Positive (TP) = number of correct positive prediction
True Negative (TN) = number of correct negative prediction
False Positive (FP) = Type I Error, No. of incorrect positive prediction
False Negative (FN) = Type II Error, No. of incorrect negative prediction

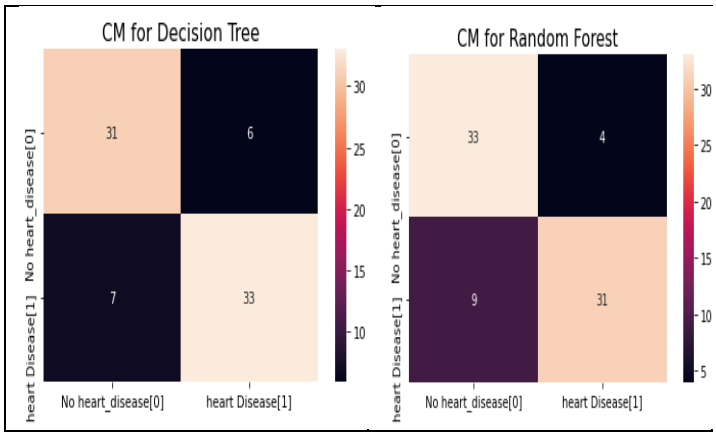
Accuracy: The accuracy of classification process is based on correct and incorrect predictions. Accuracy of the classification can be calculated by equation (4.6).

Accuracy (ACC)

$$\mathbf{ACC} = \frac{\mathbf{TP+TN}}{\mathbf{P+N}} = \frac{\mathbf{TP+TN}}{\mathbf{TP+TN+FP+FN}} \qquad (3)$$

Table 2. Confusion matrix of various ML algorithm





5.3. Features Selection

Features selection acts a indeed significant role particularly when functioning with large data set in machine learning. It can lead to better classification. The main features are shown in Fig. 9.using SVM. Green features shows the negative factors and blue corresponding to positive ones.

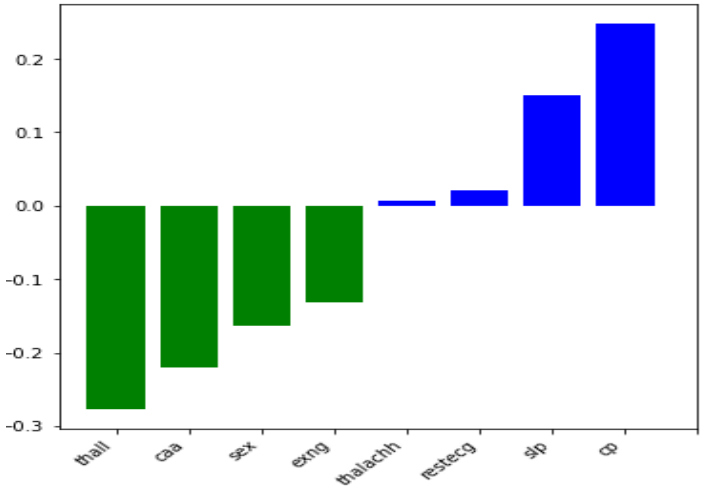


Fig.9. Features Selection

Since the graph above, it shows that the top 4 significant features are chest pain type (cp), slope peak depression (slp), rest electrocardio graphic result(restecg), and maximum heart rate (thalachh), these are used to speculate the heart diseases.

After Training and Testing by using various ML approach we get that accurateness of the SVM is far proficient as relate to other algorithms. To Find the accuracy of each algorithm we use confusion matrix as shown in Table 2. And it is conclude that SVM is best among them with 86% accuracy and the comparison is shown in Table. 3

TABLE.3 Accuracy of Various Models

S.NO.	Model	Accuracy
1	Logistic Regression	77
2	KNN	82
3	SVM	86
4	Naïve Bayes	68
5	Decision Tree	83
6	Random Forest	83

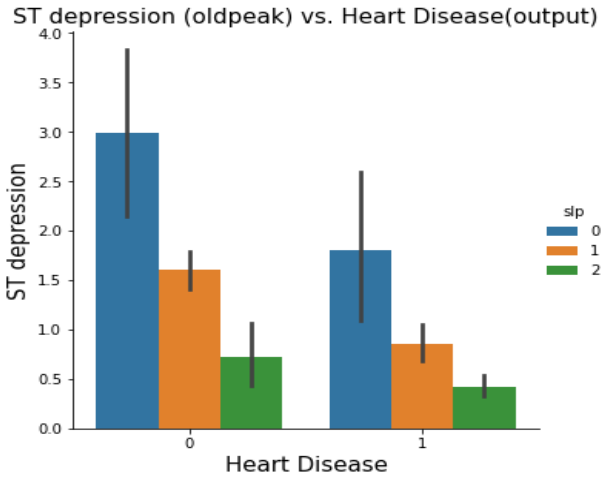


Fig.10. Cat Plot For Distribution Between oldpeak and heart disease patient

There are some plots such as cat plot, box plot violin plot that equal distribution of categories data. The basic Statics of data is easily shown by these plots. In cat plot st depression (slope) is irregularly short under the starting point this can lead to heart Disease. This is supports, that above plot the high ST depression is considered healthy and normal whereas low ST Depression consider at greater risk for heart disease. While a. The Slope (slp) values: Down sloping: 2 1 Flat: 1, Up sloping: 0.. Both negative and Positive heart disease patient distribute equally as shown in Fig. 10. By using these plots we can easily detect outliers. By using box plot, negative people have lower whereas positive patients demonstrate higher than negative for depression level. Female and male output have not more difference , and male patient have marginally superior ranges of ST Depression as display in Fig.11.

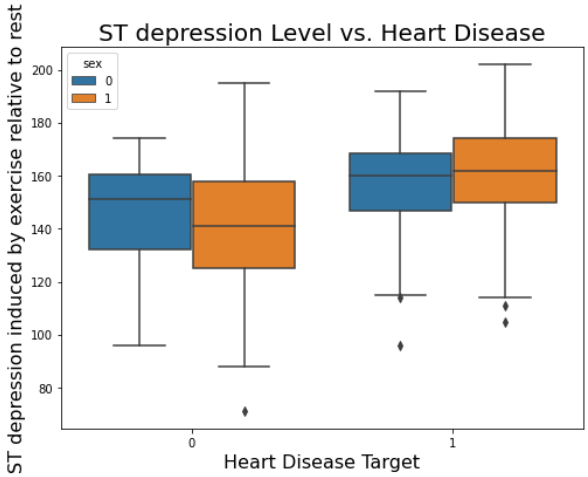


Fig. 11 Box Plot For Distribution between Depression and Heart Disease Patients

6. CONCLUSION

Heart acts a major role in corporeal organism. The diseases of heart wants more perfection and exactness for diagnose and analyses.In real time heart diseases may not be detect in early stage. This need further analysis. In proposed work, an accurate and early heart diseases prediction is presented by using data set of heart diseases .The presented methodology requires various ML algorithms. The analysis is carried out based on Confusion matrix and comparing accuracy among them and get SVM is finest algorithm. Thus the efficacy of presented work has been verified. This technique may be used as an support for early and accurate prediction of heart disease. There are many more ML algorithms that can be used for finest exploration and for earlier prediction of heart diseases for the upcoming possibility. This needs further diagnosis.

References

[1] E.Taylor,P.s.Ezekiel,F.B.Deedam. (2019). "A Model to Detect Heart Disease using Machine Learning algorithm" International journal of Computer Science and engineering.vol-7,issue-11

[2] R. Goel and A. Jain. (2018) "The Implementation of Image Enhancement Techniques on Color n Gray Scale IMAGEs," *2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, , pp. 204-209, doi: 10.1109/PDGC.2018.8745782

[3] Archana Singh, Rakesh k. (2020). "Heart disease Prediction Using machine Learning Algorithms" International Conferences On Electrical and electronics Engineering(ICE3)

[4] Pahulpreet Singh Kohli and Shriya Arora. (2018). "Application of Machine Learning in Diseases Prediction", 4th International Conference on Computing Communication And Automation(ICCCA)

[5]. A. Mishra, Abhishek Rai andAkhilesh Yadav, (2014). Medical ImageProcessing: A ChallengingAnalysis"International Journal of Bio-Science and Bio-Technology Vol.6, No.2.

[6] .Sneha A Mane, S R Chougule, (2016). Neural network of Kidney Stone Detection" International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Volume 5 Issue 4.

[7]. Adam, Tijjani, U. Hashim, and U. S. Sani, (2012). Designing an Artificial Neural Network model for the prediction of kidney problems symptom through patient's metal behavior for pre-clinical medical diagnostic. Biomedical Engineering (ICoBE), 2012 International Conference on. IEEE.

[8] Aditi Gavhane, Gouthami Kokkula, Isha Panday, Prof. Kailash Devadkar. (2018) "Prediction of Heart Disease using Machine Learning", Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology(ICECA), 2018.

[9] Amandeep Kaur and Jyoti Arora .(2019)."Heart Diseases Prediction using Data Mining Techniques: A survey" International Journal of Advanced Research in Computer Science , IJARCS.

[10] M. Nikhil Kumar, K. V. S. Koushik, K. Deepak.(2019). "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.

[11] Himanshu Sharma and M A Rizvi. (2017). "Prediction of Heart Disease using Machine Learning Algorithms: A Survey" International Journal on Recent and Innovation Trends in Computing and Communication Volume: 5 Issue: 8 , IJRITCC August 2017.

[12] https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[13] Hazra, A., Mandal, S., Gupta, A. and Mukherjee, (2017). " A Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review" Advances in Computational Sciences and Technology .

[14]Devansh Shahet.al.(2020) " HeartDiseasePredictionusingMachineLearningTechniques" © Springer Nature Singapore Pte Ltd

[15] Goel R., Jain A. (2020) Improved Detection of Kidney Stone in Ultrasound Images Using Segmentation Techniques. In: Kolhe M., Tiwari S., Trivedi M., Mishra K. (eds) Advances in Data and Information Sciences. Lecture Notes in Networks and Systems, vol 94. Springer, Singapore. https://doi.org/10.1007/978-981-15-0694-9_58