INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD ISSN: 2455-0620 Volume - 7, Issue - 3, Mar – 2021
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value: 86.87 Impact Factor: 6.719
Received Date: 01/03/2021 Acceptance Date: 18/03/2021 Publication Date: 31/03/2021

A Comparison of Different Machine Learning Models on Heart Disease Prediction

[1]Sadiyamole P. A., [2]Dr.S Manju Priya
[1]Research Scholar, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore 21, India
[2] Professor, Department of CS,CA and IT, Karpagam Academy of Higher Education, Coimbatore 21.
Email – [1]sadiya.pa@gmail.com, [2]smanjupr@gmail.com,

Abstract: *The healthcare sector has many obstacles in finding various diseases. Huge amount of patients data are collected by different healthcare organizations. With the help of these data along with data mining techniques, it is possible to predict various diseases. This is a great gift to mankind as there is no need of any invasive tests to be done on human body. Cardio Vascular Diseases (CVD) take around 17.9 million lives every year. Differentiating different ML models like Logistic Regression, K Nearest Neighbor , Random Forest, Support Vector Machine, XGBoost, etc. and find out the best technique for predicting heart disease.*

Key Words: *Machine Learning, Support Vector Machine, K Nearest Neighbor.*

1. INTRODUCTION:

Heart is the main part of human body. Basically, the main function of heart is to regulate blood flow in our body. If any asymmetry occurs in the heart it will affect entire body. According to World Health Organization [13], around 17 million of total population loss their lives due to heart failures. In United States one person dies in every 37 seconds due to CVDs. In India heart diseases contribute to 17.8% of total deaths. There are several reasons for heart diseases and some of them are modern lifestyle, smoking, alcohol and high intake of fat, lack of proper exercise etc. If the problems of heart can be recognized in advance and maintain a proper healthy diet, we can control the rate of deaths. Quality diagnosis and providing the best service are the major issues in the healthcare area nowadays. Several clinical tests like BP, cholesterol, ECG, fasting blood sugar etc. have to be done to check the victim's heart condition. This is a time taking process. In this digital world most of the processes are converted into digital form especially in healthcare field [].In this world of technological innovations, there are a lot of techniques like data mining[16] and neural networks those can be effectively used to predict the chances of occurrence of different diseases like cancer [11] and heart diseases in a person. The work suggested in this paper mainly focuses on different ML methods in heart disorder prediction.

Diagnosing heart disease by applying machine learning is one of the fast-growing areas in research. In this paper some ML algorithms like SVM, KNN, Logistic Regression, RF and XGBoost, etc. have been compared and analyzed to predict the heart disease. This paper is divided into different units. Division 1 provides an introduction. Division 2 provides various studies related to heart disease prediction, division 3 contains different methods used in this study, then discuss the experiments and results and finally provides the conclusion.

2. LITERATURE REVIEW:

Haq et al. [1] suggested various feature selection algorithms like Relief, mRMP and Lasso with K-fold cross-validation to get the crucial features from the dataset. Logistic Regression (LR), KNN, ANN, SVM, Decision Trees(DT) and Naïve Bayes are used as the models. Quality were checked on the selected features.LR with 10 fold cross validation displayed the prime accuracy when Relief was used as the FS. According to Mohan et al.[2] Hybrid Random Forest with Linear Model(HRFLM) which combines the characteristic of Random Forest and Linear Method has used. The data has been taken from Cleveland UCI repository. They started the work from pre-processing then applied feature selection based on decision tree entropy. Feature selection and modeling repeated many times for different attribute to get the best accuracy**.** Fatma Zahra Abdeldjouad et al.[3] proposed a version for anticipating Heart Disease. By using two different tools, Weka and Keel different algorithms were selected. Under Weka tool some methods like Multi-Objective Evolutionary Fuzzy Classifier (MOEFC) were selected where as Genetic Fuzzy System-LogitBoost (GFS-LB) and some other methods were applied in Keel tool .By using Weka and Keel software, **t**he quality was assessed by using Specificity, Accuracy, Sensitivity and Error rate . According to Shahadat Uddin1 et al. [4], two databases (scopus and PubMed)were searched for different types of search items. By focusing on different research articles on disease risk prediction that use machine learning models, they have selected 48 articles for comparisons. Analyzing all these articles they found that SVM and RF show the superior accuracy. Different authors have selected different variables for their study. If a new variable is added, an underperformed algorithm might have

improved. This is the limitation of this study. The accuracy of algorithm can be improved by hybridization or combination of different algorithms. In Monther Tarawneh et al.[5] the dataset is first preprocessed to eliminate any irrelevant features. Different feature preparation methods have applied on the dataset to avoid non-optimal parameters. After the first step, a number of classification techniques have applied .Accuracy, Precision, Recall and F-measures are compared. Naïve Bayes and SVM were always found to be performed better. Then hybridization applied on the selected classification. According to Saleh et al. [6] WEKA is a toolkit for machine learning .WEKA supports different data mining processes like preprocessing, grouping, classification, correlation etc. By using WEKA tool different algorithms K-Star, J48, SMO, NB, Random Forest, etc. can be applied for more prediction accuracy. In M. Thiyagaraj et al. [7] data selected from the UCI repository is normalized with the help of Zero-Score. Then Particle Swarm Optimization algorithm and Rough Sets based feature cutting method is used for choosing the ideal subset of attributes. Then RBF-SVM classifier is applied for predicting heart disease. PSO and RBF-SVM methods and other methods like IT2FLS and improved FA and RBF-SVM are compared and the proposed method proved better performance in Sensitivity, Specificity, and Accuracy than the other two.

Atul Kumar Ramotra et al. [8] compared various machine learning methods in WEKA and SPSS tools. In the WEKA tool Naïve Bayes has highest accuracy(85.39) and an accuracy of 85.87% is obtained by using SVM in SPSS tool. They considered DT, NB, SVM, ANN, and KNN for comparison. Devarajan et al.[9] used J48Graft decision tree classifier for personalized healthcare support system is good with low implementation time and higher accuracy but it is mainly focus on diabetic patient's healthcare. The authors focused on fog based health support and it has some disadvantages like data breach of patient's sensitive information and other privacy issues. Latha et al.[10] used different ensemble algorithms bagging, boosting, stacking and majority voting in their experiment and analyse the accuracy of these ensemble techniques. The result showed that majority voting has the highest accuracy among all. Fredrick David et al.[12] compared Naïve Bayes, Decision Trees and Random Forest on UCI data and they experimentally proved that Random Forest has the best performance.
The following section provides a view about the materials and methods used in this paper.

3. MATERIALS:

Most of the researchers use the "Cleveland dataset", which is available on internet. The database has a total of 303 specimen records of different heart patients with 76 features. Since there are some values are missing in the database, some rows are removed and selected only 297 records with 13 input features, and one target variable. The target variable describes whether a patient is affected with heart problem or not.
The UCI dataset is shown in Table 1

#	Column	Description
0	Age	Age of the patient.
1	sex	Patient's sex
2	Cp	Chest pain value
3	Trestbps	Blood pressure while resting
4	Chol	Patient's Cholesterol
5	Fbs	Fasting Glucose level
6	Restecg	Resting ECG
7	Thalach	Achieved peak heart rate
8	Exang	exercise induced angina pain.(1 =positive; 0 = negative)
9	Oldpeak	ST depression induced by exercise relative to rest
10	Slope	the slope of the peak exercise ST segment
11	Ca	number of main vessels coloured by fluoroscopy
12	Thal	3 = normal; 6 = fixed defect; 7 = reversible defect
13	Target	Heart disease or not

Table1. Description of UCI heart disease dataset

The suggested system has been developed to recognize patients who have heart disease or not. According to WHO, CVDs are a collection of chaos of the heart and its vessels like

- coronary heart disease – happened when the blood supply to the heart got blocked;

- cerebrovascular disease –a disease that affects blood flow to the brain;
- peripheral arterial disease – It happens when some fatty substances deposit in the veins of the arms and legs;
- rheumatic heart disease – Destruction to heart caused by rheumatic fever,
- congenital heart disease – Harm to heart at birth itself;

3.1. METHODS:

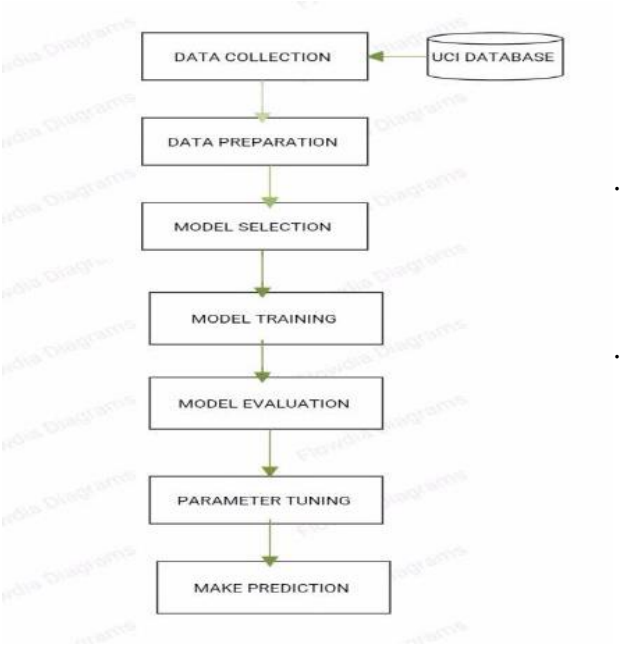

Figure 1 Steps in ML

Several kinds of research have been made in the area of forecasting heart disease using various Data Mining and Machine Learning tools. The famous ML techniques like Naïve Bayes, Decision Tree, SVM, KNN, Logistic Regression, Random Forest etc. have been used by various researchers. Different steps in ML process are depicted in fig .1.All the above models are implemented in Python3.

- Data Collection-Data can be collected from various datasets like UCI, Kaggle etc.
- Data Preparation:-This step contains removing of errors and duplicates, treating missing values, normalization etc.
- Model selection-Selecting the right algorithm from the available ML models.
- Model training-The aim of this step is to make predictions correctly. Each iteration of the process is a training step.
- Model evaluation-Is done by using some metrics measure the performance of model. Test the model against some previously unseen data.
- Parameter tuning-Tuning of model parameters for improved performance is done in this step
- Make prediction-Using further test set which has been withheld from the model are used to test the model.

Heat map of the correlation matric of Cleveland dataset drawn in python is exhibited in fig 2

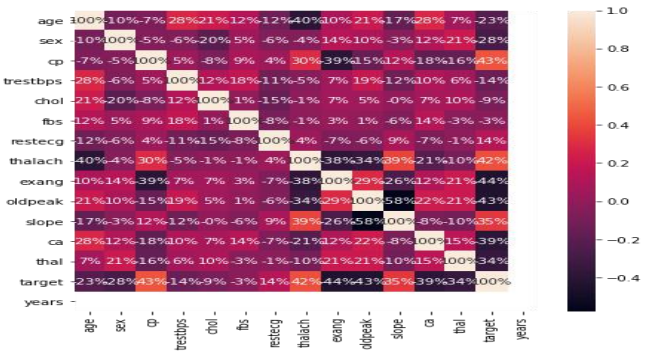

Figure 2 Heatmap of the Cleveland dataset

The following algorithms are taken for experiments and all these are implemented in Python.

1 Logistic regression

It is a type of classification problem. The cost function in logistic regression is known as Sigmoid Function. The hypothesis expectation of logistic regression limit the cost function between 0 and 1.ie.$0 \leq h\theta(x) \leq 1$.Logistic Regression gives an accuracy of 84.46

2 KNN

KNN is a classifier problem even though it is used in the regression field also. KNN assumes that similar things stand close to each other. KNN calculates distance between points on a graph as it relies on closeness. In order to select the particular K value for our problem, we have to run the program with different values of K and select K that has fewer errors. Here KNN shows an accuracy of 72.13 when k=1 and gives the best result of 73.77 when K=5.

3 SVM

A supervised ML algorithm called SVM can be applied in classification and regression fields. The purpose of SVM is to obtain a hyper plane in n-dimensional space, where n is the total attributes. Then, carryout categorization by calculating the hyper-plane that alter the two classes' .That conveys to find a plane that has the extreme margin. SVM shows an accuracy of 75.41

4 Random forest

Random forest is an ensemble model that contains many decision trees. Individual tree in the random forest gives a forecasting and the maximum voted class is selected as our model. It's training time is low as compared to other model and it gives the accuracy of 69.73.

5 XGBoost

XGBoost is an ensemble ML method based on decision-tree that uses a gradient boosting framework. XGBoost and Gradient Boosting methods are ensemble techniques. However, XGBoost performs well over GBM framework by system boosting and algorithmic intensification. It has the highest accuracy of 88.52% of all other models. The Fig 3 shows the implementation of XGBoost[14] .

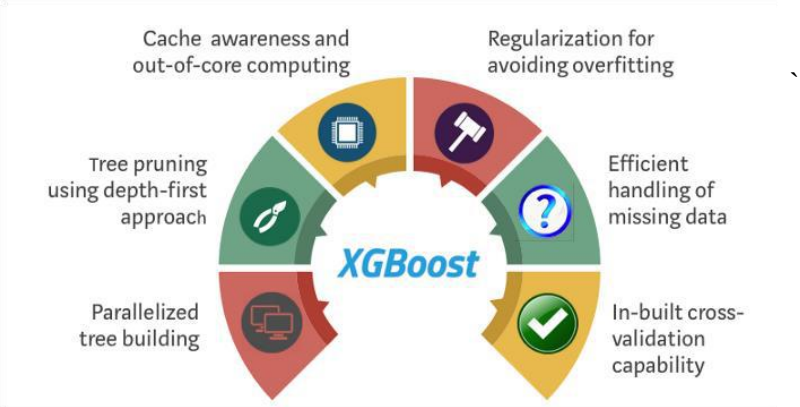

Figure 3 XGBoost implementation

4. EXPERIMENTS AND RESULTS:

The confusion matrix is a method of providing a summary of prediction results. The confusion matrix shows how much your model is confused when predicting results. Confusion matrixes are nowadays widely used to measure the performance of our models. In a two-class problem confusion matrix can be displayed in fig 4and fig 5 depict confusion matrices of all models used in this study.

		Actual Value		
	TP-	**TP**	**FP**	True Positive
	FP-			False Positive
Predicted Value	FN-	**FN**	**TN**	False Negative
	TN-			True Negative

Figure 4. Confusion Matrix

Accuracy can be calculated as

Accuracy = [TP+TN]/[TP+FP+FN+TN]

24	4
5	28

Logistic Regression

20	10
5	26

SVM

17	10
6	28

KNN

25	10
13	28

Random Forest

24	3
4	30

XGBoost

Figure 5.Confusion matrix of all models.

Table 2 shows the comparisons between different ML models and Fig 6 shows a bar graph drawn by Matplotlib that depicts model evaluation used in this study .

Table 2. Different models' comparison.

Model name	Accuracy	Precision	F1-Score	Recall
Logistic Regression	84.46	82.07	84.21	85.71
KNN	73.77	74.28	78.78	83.87
Random Forest	69.73	65.78	68.48	71.42
SVM	75.41	80	73	66.66
XGBoost	88.52	85.71	87.27	88

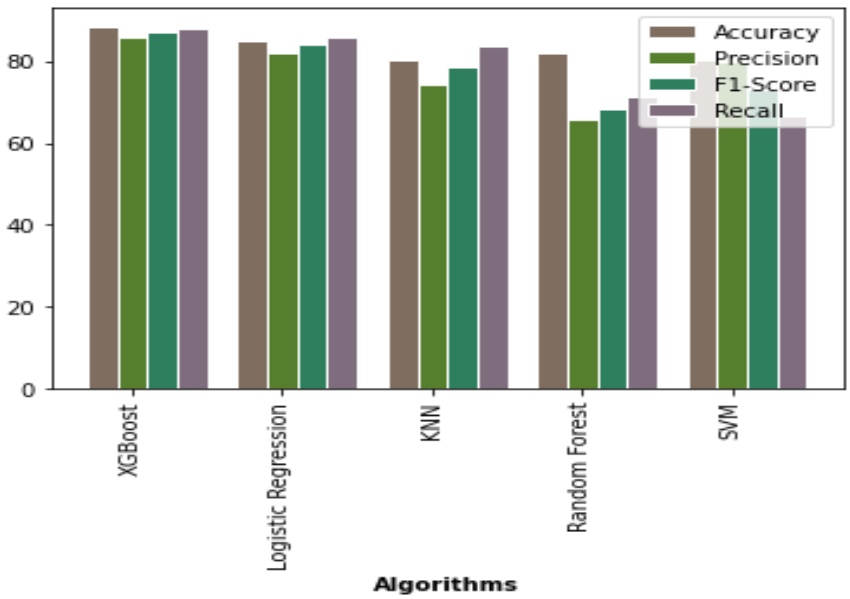

Figure 6 Model Evaluation

5. CONCLUSION:

Healthcare data mining has a prominent role in identifying different diseases in advance which saves human lives. There is no need to perform different invasive tests as different clinical data related to different diseases are freely available on internet and other sources. This data can be effectively applied in different machine learning methods to anticipate the possibilities of happening those diseases. In this paper, an attempt has been made to apply heart disease datasets on different ML techniques and the prediction result of this paper showed that XGBoost outperformed well with the best accuracy of 88.52%.

REFERENCES:

JOURNALS

1.]Haq, Amin Ul,Li, Jian Ping Memon, Muhammad Hammad, Nazir, Shah Sun, Ruinan", A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms"- Mobile Information Systems.-2018.
2. Mohan, Senthilkumar, Thirumalai, Chandrasegar, Srivastava, Gautam- "Effective heart disease prediction using hybrid machine learning techniques"-2019-7-P 81542-81554.
3. Fatma Zahra Abdeldjouad, Menaouer Brahami,,Nada Matta-" A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques"- ICOST 2020: *The Impact of Digital Technologies on Public Health in Developed and Developing Countries*-2020-V 12157-P 277-286.
4. Shahadat Uddin1,, Arif Khan1, Md Ekramul Hossain and Mohammad Ali Moni-"Comparing different supervised machine learning algorithms for disease prediction"-*BMC Medical Informatics and Decision Making*-2019-8-p1-16.
5. Monther Tarawneh and Ossama Embarak –"Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques"-2019- *Springer International Publishing*- P 447-454
6. Saleh, Basma,Saeidi, Ahmed Al-Aqbi, Ali Salman, Lamees- "Analysis of Weka Data Mining Techniques for Heart Disease Prediction System"- *International Journal of Medical Reviews*.2020-7-1-P15-24.
7. M. Thiyagaraj and G. Suseendran-"Enhanced Prediction of Heart Disease Using Particle Swarm Optimization and Rough Sets with Transductive Support Vector Machines Classifier"- *https://www.researchgate.net/publication/336024368-P217-230*.
8. Atul Kumar Ramotra, Amit Mahajan, Rakesh Kumar and Vibhakar Mansotra-"Comparative Analysis of Data Mining Classification Techniques for Prediction of Heart Disease Using the Weka and SPSS Modeler Tools"-*Smart Innovation, Systems and Technologies*.-2020-V165-P393-404
9. Devarajan, Malathi,Subramaniyaswamy, V.Vijayakumar, V.Ravi, Logesh-"Fog-assisted personalized healthcare-support system for remote patients with diabetes*"-Journal of Ambient Intelligence and Humanized Computing*.-2019-V10-P 3747-3760
10. Latha, C. Beulah Christalin Jeeva, S. Carolin-"Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques"-*Informatics in Medicine Unlocked*-2019-V16-P- 100203.
11. Abed Mohammed, Mazin Khanapi Abd Ghani, Mohd Mostafa, Salama Taha Al-Dhief, FahadIbrahim Obaid, Omar Mostafa, Salama A Taha AL-Dhief, Fahad-"Evaluating the Performance of Machine Learning Techniques in the Classification of Wisconsin Breast Cancer" -*Article in International Journal of Engineering and Technology*-2018:V7-P 160-166.
12. Fredrick David, Benjamin H,Benjamin Fredrick David, H Antony Belcy, S- Heart Disease Prediction Using Data Mining Techniques- *Journal on Soft Computing*-2018-p- 1824-1831.
13. M.Roopa, Dr.S.Manju Priya - A Review of Big Data Analytics in Healthcare-IJSRD - *International Journal for Scientific Research & Development| Sp. Issue – Data Mining 2015* | ISSN (online): 2321-0613.
14. V. Kirubha1 , S. Manju Priya-Comparison of Classification Algorithms in Lung Cancer Risk Factor Analysis-*International Journal of Science and Research (IJSR)* ISSN (Online): 2319-7064.

Web References:

- https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
- http://dataanalyticsedge.com/2019/11/23/xgboost-using-python/