ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

Analysis of Supervised Machine Learning Models on Heart Disease Prediction

Sadiyamole P A[1], Dr.S Manju Priya [2]

1Research Scholar, Department of Computer Science,Karpagam Academy of Higher Education,Coimbatore 21,India

sadiya.pa@gmail.com

2Professor, Department of CS,Karpagam Academy of Higher Education,Coimbatore 21,India

smanjupr@gmail.com

Abstract— **Different types of heart diseases are the major cause of reduction of human being on earth.Health sector around the globe find out many obstacles inorder to predict the disease.The number of death increases just because the patient who has heart disease can not be diagnosed very fastly.For this different types of tests like blood glucose,ecg,cholesterol etc have to be done .For the past couple of decades ,health care people collect various details of patients undergone heart problems.These data are stored on online databases.By using this data , the status of a patient's heart health can be predicted. These stand-in data can be collected from different repositories.For this various machine learning techniques are available.The main objective of this paper is to compare different ML algorithms like Logistic Regression, K Nearest Neighbour (KNN), Random Forest, Support Vector Machine (SVM) and XGBoost etc which have been applied in state of the art research for heart disease prediction. .**

Keywords— **Machine Learning,Random Forest,Support Vector machine**

I. INTRODUCTION

Nowadays technology is getting involved more and more in our day to day life. Just like in another field, technological innovations affect very deeply in the medical sector. New devices have been developed and disease diagnosis has been very much easier. New technologies like Big data implementation in healthcare sector[12] is one of the promising field in recent years.Heart is the most important organ of human body whose function is to pumb blood to different parts of the body. It has been working continuously since our birth without much problem.But due to lack of exercise and some unhealthy food habit ,the heart sometimes shows some abnormalities. These assymetrical problems of heart will affect the whole body and sometimes it cause the death of a person. The major symptoms of heart failure includes sweating, fatigue, chest pain etc[1].According to World Health Organization[14] around 17 million people die of CVDs, particularly heart attacks. In United States one person dies in every 37 seconds due to CVDs. The rate of deaths can be controlled to a great extent if heart failures can be detected at an early stage.In order to make sure whether a patient has heart disease or not, several body inavasive tests like BP,ECG,blood sugar,cholesterol etc have to be done. With the help of technology like machine learning, the time taken for heart disease diagnosis can be reduced to a large scale. Researchers have brought forward different machine learning models with improved accuracy for the heart disease prediction.The work suggested in this paper mainly focuses on different machine learning methods that can be applied in heart disease prediction. According to WHO ,CVDs are a group of disorders of the heart and blood vessels and it includes coronary heart disease – a disease of the blood vessels supplying the heart muscle;cerebrovascular disease –a disease of the blood vessels supplying the brain; peripheral

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

arterial disease – a disease of blood vessels supplying the arms and legs; rheumatic heart disease – damage to the heart muscle and heart valves from rheumatic fever, caused by streptococcal bacteria; congenital heart disease – malformations of heart structure existing at birth; deep vein thrombosis and pulmonary embolism – blood clots in the leg veins, which can dislodge and move to the heart and lungs.

Heart disease diagnosis with the help of machine learning has been a fast-growing area in the field of research. Likewise Internet of Things has been using in HD prediction[15]. In IoMT also ML techniques can be applied for proper diagonisis. In this paper some machine learning algorithms like Support Vector Machine(SVM), K Nearest Neighbour(KNN), Logistic Regression, Decision Tree,Random Forest and Naïve Bayes, etc. have been used by various authors to predict the heart disease. This paper is divided into different sections. Section 1 gives an introduction. Section 2 provides various studies related to heart disease prediction, section 3 contains materials and methods, then discuss the results and finally provides the conclusion.

II. LITERATURE REVIEW

Dr.S.V.Kogilvani[2] et.al predicted heart disease by using Clevland dataset.Since the online dataset contains small number of references, they create some synthetic dataset.After proper data preprocessing , performance of both Clevland and synthetic data are compared on the basis of accuracy,recall and precision and synthetic data found to be performed well.

In [3] Rishabet.al also took data from Clevland and developed a web based application form which will be helpful for the patients to know whether their heart has some problem by entering their health data in the form.The interactive web application was developed using HTML,CSS Django framework along with ML techniques.They have compared Logistic Regression,SVM,DT and NB and Logistic Regression got highest accuracy of 82.89.The advantage of the research is the end user can use this web application for preliminary prediction about the condition of their heart.

G.Dinesh Kumar[4] analyzed patient's heart health by taking data from Clevland,Hungarian,Switzerland and Longbeach databases.After proper data preprocessing different algorithms like LR,NB,RF,SVM and GB are used for modeling and R programming language is used for statistical computation and visualization.

B.fredrik[5] has used several number of experiments using cross validation and percentage split on UCI statlog dataset.Here NB,DT and RF are compared to analyse heart data.Three algorithms are evaluated a number of ties by using different evauation strategies and Random Forest performed well.In this research only precision ,recall,f-measure,ROC are measured but no accuracy.

R.Indrakumari[6] et.al suggested a research work in which the main risk features that affect heart is considered and K-means algorithm is used for analysis along with talaeu tool for isualization.By using this unsupervised algorithm the research predicts four types of chest pain.

Divya Krishnani[7] et.al proposed a model that use various ML algorithms RF,DT,KNN on Framing ham Heart Study dataset. The dataset contains 4240 record of which only 15.2% has heart problems and the other 84.8% has no heart disease.Missing values are replaced by mean.In order to balance the class, random sampling is applied.Different ML algorithms RF,DT and KNN are used for analysis and RF got highest accuracy.10 fold cross validation is also applied in the model for better accuracy.

A.Gonsalves[8]et.al built a prediction model by using South African Heart disease dataset from KEEL.For pre-processing and other data

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

analysis,they used WEKA tool.DT,SVM and NB are used for analysis with10 fold cross validation.Since the data set is unbalanced,the researchers measure other metrics such as specificity and sensitivity apart from accuracy.But sensitivity and specificity rates are very low.

Terrada[9] et.al has suggested supervised ML based medical diagnosis system that uses three different databases(Clevland,Hungarian,Z-alisadeh) to predict atherosclerosis that causes heart disease.Three ML techniques ANN,adaboost,DT are used in three datasets for comparison.For ANN,they tried different number of simulations for best parameters.ANN outperformed well with 94% accuracy when used with Z-alisadeh dataset. than others.Haq et al. [10] suggested a hybrid method for HD prediction by using LR,KNN, ANN, SVM, DT and NB on Clevland dataset.The researchers first used these classifiers on full dataset.Then they applied feature selection methods Relief,mRMR and Lasso with K-fold cross-validation.Then both results are compared and various performance measures are used for evaluation. Relief FS performed well than other two feature selection methods.

In[11]M.Thyagragen brought forward PSO and RS with TSVM,that uses data from UCI database.Z-score is used for data normalization along with PSO and RS based attribute selection methods.The proposed PSO RBF TSVM method is compared with existing IT2FLS and MFA and RBF-SVM based methods and the proposed method performed.

III. MATERIALS AND METHODS

Since medical data are freely available on internet researchers make use of it for the easy diagnosis.Using these dataset it is possible to analyze and reveal various diseases. Machine learning methods can be applied to the heart disease dataset which is available on the internet for predicting heart failure. These dataset may contain some missing values, or some redundancies. So instead of just applying the Machine Learning technique, some type of feature engineering can be done on these datasets so that the efficiency of these models will improve.[11].

A.Datasets

Various heart disease datasets are available on internet of which the "Cleveland heart disease dataset"[12] has been used by most of the researchers in this review. This dataset has a total of 303 records of heart patients with 76 attributes. There are so many missing values, errors and some redundancies in the records. So some rows are removed and selected. 297 records with 13 important features and one target variable. The target variable shows whether a patient has heart disease or not. The UCI Clevland dataset is shown in Table 1.Most of the heart datasets contain these types of parameters.

TABLE1. DESCRIPTION OF UCI CLEVLAND HEART DISEASE DATASET

Sl No	Attributes	Explanation
0	Age	Age of the patient
1	sex	Gender of the patient
2	Cp	Value of Chest pain - 1: typical angina - 2: atypical angina - 3: non-anginal pain - 4: asymptomatic
3	Trestbps	Resting hyper tension
4	Chol	Cholesterol
5	Fbs	Blood Sugar Level in Fasting
6	Restecg	Resting ECG
7	Thalach	Peak heart rate attained
8	Exang	exercise induced angina (1 =yes; 0 = no)
9	Oldpeak	ST depression due to exercise corresponding to rest
10	Slope	Maximum exercise ST segment slope - 1: up sloping - 2: flat -3: down sloping
11	Ca	Count of main vessels (0-3) colored by fluoroscopy
12	Thal	3 = normal; 6 = fixed defect; 7 = reversible defect
13	Target	Whether the patient has Heart

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

failure or not

B. Methodology

Machine Learning is a type of Artificial Intelligence that helps to identify patterns from data. Nowadays data is the new oil. By using this precious data, Machine Learning can do different analysis and operations then finally takes decision. There are different types of Machine Learning tools to predict heart disease.ML is a set of techniques to make computers better at doing things than human beings can do.ML involves makimg machines learn things like human do.ML uses a set of techniques to extract knowledge from existing data and with the help of these data decisions are made.The main steps in ML process are

Exploratory data analysis-Understand and analyse data.

Feature Engineering-Handling missing values, handling outliers, normalization and standardization.

Feature selection-Removing those features which are not contributing the output thus by including only strong features.

Model Selection-Selecting the right algorithm from the available ML models. Some models may be suitable for numerical processing; others may be suitable for image processing etc. To select the appropriate model for our problem is one of the major tasks.

Model Training-The aim of this step is to make predictions correctly. Training means learning good values for all the weights and bias.

Model Evaluation-Once the model training is complete, the model's quality should be evaluated. Hereby using some metrics, measure the performance of model. Test the model against some new data.

Parameter Tuning-After evaluation, it is possible to see if it is possible to improve the training by 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.

IV. RESULTS AND DISCUSSION

There are different metrics used to evaluate Machine Learning models. They are confusion matrix, accuracy, precision, recall, ROC curve, etc. Using the correct kind of metric to find out how good our model is one of the important factors. This is used to get an idea of our model's performance. Various Machine Learning models used in the above papers are listed in Table 2.

There are different classification matix in Machine learning models .They are confusion matrix, Accuracy, precision, recall,f1 score,ROC curve,AUC score.Selection of right metric is important in describing the model.If the wrong metric is selected, then the performance of the model will be very poor.The most important metric is confusion matrix ,which is shown below(Fig 1).Here FP is known as Type 1 Error and FN is called Type 2 error.TP and TN are the most accurate results and in any classification problem the aim should be to reduce type1 and type2 errors.From the confusion matrix ,accuracy can be measured as

Accuracy=(TP+TN) / (TP+FP+FN+TN).

But if the dataset is imbalanced,then accuracy will not be a good metric.In that case,other metrics like Recall,Precision,F score,MCC,AUC score etc will be helpful. In this paper, researchers have taken different metrics for different datasets.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

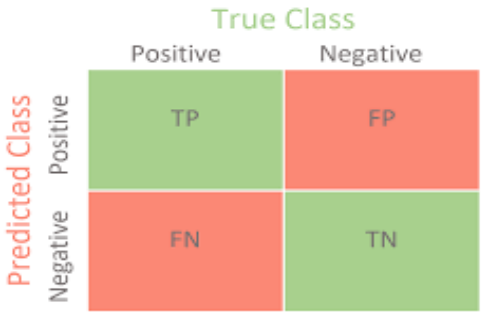

Fig 1 Confusion Matrix

Comparisons of different research works is shown in Table 2.Four researchers have taken Clevland dataset only and three of them have taken Clevland along with other datasets.Three authors have taken different datasets.Out of the ten papers , seven authors used Decision tree,four authors used Logistic Regression,five authors used Support Vector Machine,five of them used Naïve Bayes, four used Random Forest,three papers applied K nearest neighbours and adaboost one and only one author used unsupervised algorithm

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

References	Dataset	Data preprocessing	Algorithms compared	Evaluation	Prediction performance and remarks
Dr.S.Kogilvani	Clevland and synthetic data	missing value treatment,standard scalar,minmax scalar	DT,KNN,RF,SVM,LR	Not mentioned	Accuracy,precision,recall,f1score.-synthetic dataset performs well
Rishabh Magar	Clevland	Null value treatment,feature scaling	LR,SVM,DT,NB	Not mentioned	Only accuracy is calculated.LR has highest accuracy of 82.89
G.Dinesh Kumar	Clevland,Hungarian, Switzerland,Longbeach	Data cleaning,data integration,missing value treatment	LR,SVM,NB,RF,GB	Not mentioned	Only accuracy is calculated.LR has highest accuracy of 91.61
B.Fredrick David	UCI Statlog	not mentioned	RF,DT,NB	K-fold cross validation& percentage split	TP rate,FP rate,precision,recall,f1 score,ROC,MCC,PRC.RF performs well but accuracy is not defined
R.Indrakumari	Clevland	missing value treatment,Tableau for data visualization	Unsupervised algorithm-K means clustering	Not mentioned	Four types of chest pain are predicted.
Krishnani	Framing ham Heart Study	missing value treatment,random sampling for dataset balancing	RF,DT,KNN	10-fold coross validation	confusion matrix, accuracy, precision,recall,f1score,,ROC RF performs well with 96.8% accuracy
H Amanda	South African HD dataset from KEEL	WEKA tool used	DT,NB,SVM	10-fold coross validation	accuracy,sensitivity and specificity.NB performed well with 71.5.Dataset is imbalanced.
Terrada	Clevland,Hungarian, Z-Alisadeh	missing value treatment,changed categorical values to binary values	ANN,Adaboost,DT	10-fold coross validation	Accuracy,precision,recall,f1score,MCC.ANN performed well with 94% in Z-Alisadeh dataset.
A.Haq	Clevland	missing value treatment,standard scalar,minmax.FS methods .Relief,mRMR and Lasso	LR,KNN,ANN,SVM,NB,DT	10-fold coross validation	Accuracy,sensitivity,specificity, error rate,MCC,AUC and ROC.Relief FS method performed well.
M.Thyagaragen	UCI	Z-score for data normalization and PSO and RS for FS	RBF TSVM, IT2FLS and MFA and RBF-SVM	Not mentioned	Accuracy,sensitivity,specificity. RBF TSVM performed well.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

K means clustering.In many papers data processing as well as validation methods are not used. Accuracy is very important in measuring ML models, but some papers fail to predict the accuracy. By considering these disadvantages, to design machine learning model with proper feature engineering and validation methods is the main aim of the future work.

V. CONCLUSIONS.

Many industries are successfully using Machine Learning. Healthcare sector is the most important beneficiaries of Machine Learning that help them in identifying different diseases in advance. Since datasets about various diseases are freely available on internet, disease prediction becomes much easier without doing any invasive tests on human body. In the case of heart disease prediction, Machine Learning methods can be applied to heart disease dataset available on internet, we can easily predict whether a patient has heart disease or not. The state-of-the-heart of the research related to heart disease prediction using Machine Learning has been discussed and compared in this paper. The next priority is to improve the accuracy of the models by using neural networks with proper feature engineering methods.

References

[1].C.A.Devi,S.P.Rajamhoana, K.Umamaheswari , R. Kiruba, K. Karunya, and R. Deepika, ''Analysis of neural networks based heart disease prediction system,'' in Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI), Gdansk, Poland, Jul. 2018, pp. 233–239.

[2] Dr.S.V.Kogilavani, K.Harsitha, P.Jayapratha and S.G.Mirtthula- Heart Disease Prediction using Machine Learning Techniques- International Journal of Advanced Science and Technology Vol. 29, No. 3s, (2020), pp. 78-87.

[3] Rishabh Magar, Rohan Memane, Suraj Rau- HEART DISEASE PREDICTION USING MACHINE LEARNING. Journal of Emerging Technologies and Innovative Research (JETIR) www.jetir.org- June 2020, Volume 7, Issue 6.

[4] G, Dinesh Kumar-Prediction of Cardiovascular Disease Using Machine Learning Algorithms-Proceeding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India

[5] Fredrick David, Benjamin H,Benjamin Fredrick David, H,Antony Belcy, S-HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES Content Based Image Retrieval View project Medical Data Mining -Journal on Soft Computing;Year:2018;November;Pages1824-1831

[6] R. Indrakumari, T. Poongodi, Soumya Ranjan Jen-Heart Disease Prediction using Exploratory Data Analysis-Procedia Computer Science-2020;Volue173-Issue C-Pages130-139.

[7] Krishnani, Divya,Kumari, Anjali,Dewangan, Akash,Singh, Aditya,Naik, Nenavath Srinivas-Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms-IEEE Region 10 Annual International Conference, Proceedings/TENCON
2019 October;Pages: 367-372.

[8] Gonsalves, Amanda H.,Thabtah, Fadi,Mohammad, Rami Mustafa A.Singh, Gurpreet-Prediction of coronary heart disease using machine learning: An experimental analysis-ACM International Conference Proceeding Series-January 2021-Pages51-56

[9] Terrada, Oumaima,Hamida, Soufiane,Cherradi, Bouchaib,Raihani, Abdelhadi,Bouattane, Omar- Supervised machine learning based medical diagnosis support system for prediction of patients with heart disease-Advances in Science, Technology and Engineering Systems-2020-Volume5-Issue5-Pages 269-277.

[10]]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.

[11] M. Thiyagaraj and G. Suseendran-"Enhanced Prediction of Heart Disease Using Particle Swarm Optimization and Rough Sets with Transductive Support Vector Machines Classifier"from-https://www.researchgate.net/publication/336024368-P217-230

[12] Nkundimana, Gakwaya,Manju, S.- Big Data Analytics in Healthcare and Delve Bioinformatics Data Space for Health Amelioration-International Journal of Computer Applications-2018-V180-I 9-P 43-50.

[13] https://archive.ics.uci.edu/ml/datasets/heart+disease.

[14] https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[15] Mendez, Fernando,Jabba, Daladier- IoT Connected Health Architecture for Heart Rate Monitoring based on Interoperability Standards- 2018 IEEE 2nd Colombian Conference on Robotics and Automation, CCRA 2018-Pages1-6.