ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 8, August 2022

Heart Disease Prediction Using Ensemble Stacking Technique

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

[1] Research Scholar, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, India.
[2] Professor, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, India.
Corresponding Author Email: [1] sadiya.pa@gmail.com, [2] smanjupr@gmail.com

Abstract*— Heart disease is one of the critical reasons behind the majority of the human loss.Heart failure has proven as the major health issue in both men and women.This causes human life very dreadful.Diagnosing heart issues in advance is a tedious task as it requires enormous amount of clinical tests.Data mining techniques like machine learning and deep learning have proven to be fruitful in making decisions and diagnose various diseases in advance.In this paper,various machine learning techniques have been used along with stacking ensemble method that focus to improve the prediction of heart failure.The accuracy of diagnosis is very important in the case of heart disease.Due to the inadequacy of prediction and diagnosis, traditional approaches fail to discover various heart failures.Health care organizations collect heart data sets which can be used to apply machine learning models for prognosis.

Index Terms*— Machine Learning, XGBoost ,Stacking,Ensemble.

I. INTRODUCTION

As per World Health Organization[4]'s report, around 17 million people lose their lives due to Cardio Vascular Diseases. Heart is the primary part of human body and its main purpose is to pump blood.If anything happens to the transportation of blood, the entire body including the brain will damage.This is why heart disease is treated as the most deadliest disease of the world.17.8% of total deaths in India is due to heart failures.The reasons of heart failures may be due to high level of alcohol consumption,smoking,high intakes of fast foods,high blood pressure, increased sugar and cholesterol level and sometimes hereditary[1]. In order to identify whether a person has heart disease or not,several invasive tests like Electro CardioGram, Blood pressure, cholesterol, chest X-ray,tred mill test etc have to be done.But it is not bearable to everyone especially people in developing and undeveloped countries.Since all these test requires complex scientific equipments, heart disease diagnosis is still continue to be an unattainable for many people.This makes heart disease prediction a major concern for researchers in the past few years.Heart diseases are different types [10] like Coronary heart disease,rheumatic heart diseases etc.

Several contributory risk factors like blood pressure,insulin level, cholesterol,pulse rate etc require detailed clinical tests and analysis of patient's history by the medical pratitioners.Increase in the body mass index is also one of the reasons of failure of heart functions[16].Since this is a time taking process researchers have now turned their concerns toward modern scientific approaches like Machine Learning and Deep Learning for the prognosis of heart disease.Machine learning,deep learning techniques,genetic algorithms[17] etc can be used for predicting various disease like cancer[2][3].Data mining methods can be applied to existing heart dataset along with different software tools: researchers are now able to predict various cardio vascular disease in advance with improved accuracy.Machine learning aids computers to learn and make decisions accordingly.Predict cardiac diseases of persons in advance is the main aim of the proposed research work. Machine learning algorithms like SVM,NB,KNN etc can be used as a base classifier for building an ensemble classifier in order to improve performance[17].The proposed design is sectioned into different parts.Section 1 describes the introduction, section 2 describes various research studies in heart disease prediction field, section 3,4 and 5 descibe the method applied in this work, the proposed method,results and discussion respectively and at last section 6 describes the conclusion.

II. RELATED WORKS

Medical diagnosis is an important application of Machine Learning.In Mohan et al.[5] Hybrid Random Forest with Linear Model(HRFLM) is the combination of RF and Linear Method has used. Pre-processing has done on Clevland dataset and then decision tree entropy based feature selection applied.Accuracy obtained is 88.7%.According to Sumit Sharma[6] et.al. UCI dataset is applied by Taloz optimizer and Talos DNN model has 90.76% accuracy.In V.Sharma[7] the authors have created synthetic dataset from Clevland to avoid the problems of small dataset.Later they compare the performance of both Clevland and synthetic dataset on the basis of accuracy,recall and precision.synthetic dataset performs well.

In B.Fredrick[8] NB,Decision Tree and RF have applied on UCI Statlog dataset.For results validation,different number of experiments are done using cross validation and

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 8, August 2022

percentage split.Only precision,recall,F-measure,ROC area are measured.R.Indrakumari [9] et.al considered the main risk factors that affect heart disease and un supervised algorithm K-means clustering is applied on Clevland dataset and four types of chest pain are predicted with the help of Tableau visualization tool.B.Saleh[10]et.al reviewed various researches based on Data mining and identified major risk factors that affect heart disease.WEKA tool is applied here.D.Krishnani[11]et.al used Framing ham Heart study dataset .Missing values are replaced by mean and for balancing the class random sampling applied. Random forest performed well with 96.8% than Decision tree and K nearest neighbour.

III. MATERIALS AND METHODS

This section describes data and methodologies used in this research.First part describes the online dataset and follows various pre-processing done on the data.

A. Datasets

In the proposed study,the dataset is taken from KaggleStatlog-Clevland-Hungary dataset and is available in the repository[19].The dataset consists information of patients from US, UK, Switzerland and Hungary.It has 1190 records with 11 features and one target variable.Various features are age,sex,chestpaintype,resting blood pressure, cholesterol,fasting blood pressure, max heart rate , rest ecg, exercise induced angina, st_depression,st_slope from ECG readings and target.

The target variable shows the existence or non-existence of heart disease.In order to analyse, medical practitioners examine the historical data given by the patients[14].The research[15] has explained the featuresvery well.TheKaggle dataset is described in Table 1.

Table1. Description of Clevland ,Hungary, Statlog dataset

#	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	St segment's the slope of the peak exercise ST segment.
11	Thal	3 = normal; 6 = fixed defect; 7 = reversible defect
12	Target	Heart disease or not

B. Data Preprocessing

Among 12 features,dataset contains 5 numeric variables and 6 nominal variables.Four different types of chest pain are recorded in the dataset such as typical angina,atypical angina,non angina pain and asymptomatic.Here normal value can be asymptomatic and it is changed to 0 and all other categorical values are changed to 1.Likewise all nominal values are changed accordingly.The dataset is found to be null value free.Distribution of target variable is shown in Figure 1 and numerical features are shown in Figure 2.

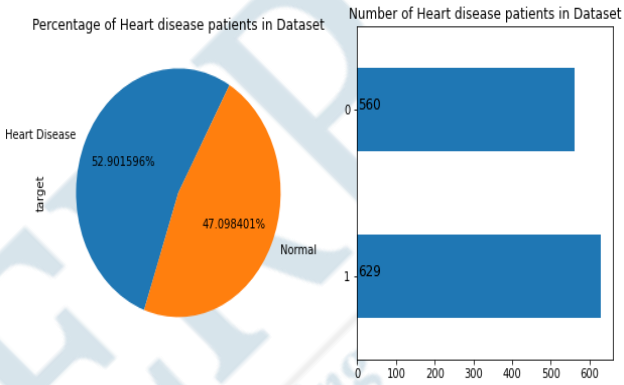

Fig 1. Distribution of target variable

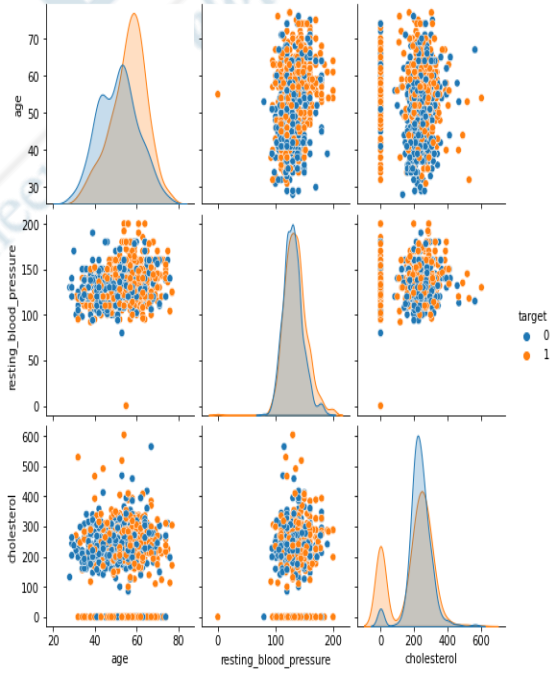

Fig 2. Distribution of numerical features

An outlier is a very high or very low data value with respect to other data in the list.It may be due to data entry errors.Identifying outliers in a dataset is an important part of data preprocessing.In this research, outlier is identified using Z-score method.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 8, August 2022

Z-Score is signed representation of standard deviations by which the data exceeds the mean value of what is being noticed.Then a threshold of 3 has been defined and a total of 17 data has obtained as outliers.After removing all outliers,the dataset contains 1173 records of 12 features.

For selecting feature selection ,correlation of different variable is checked.Then it is found that cholesterol and maximum heart rate achieved are negatively correlated towards the target variable.So these two features are removed for better performance.Later the ensemble stacking model is compared with full features and with reduced nine features which has got after removing the two negatively correlated variables.

After that the dataset is split into training and testing set.Training set contains 938 records and testing set contains 235 records.Then all numeric values are normalized in the range 0 to 1.

Normalized in the range 0 to 1.

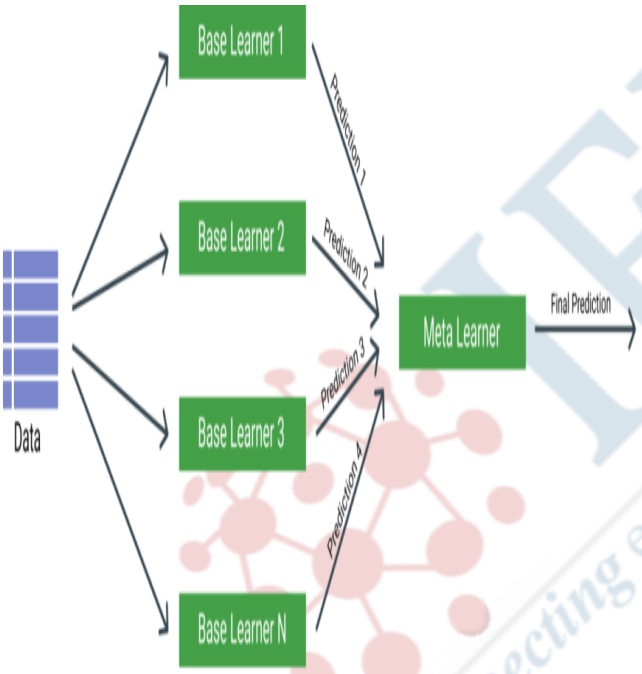

Fig 3. Single level stacking model

IV. ENSEMBLED STACKING MODEL

Ensemble learning[12] combines a number of weak learners.Inensembling, different models are combined on the same learning data.There are different ways to perform ensembling; majority voting,bagging,boosting,stackingetc.In this process,research stacking method is applied.In stacking individual learners may be heterogeneous whereas in bagging and boosting individual learners should be homogeneous.A single level stacking model [13]can be shown in the figure 3.The detailed diagram of proposed ensemble stacking model is shown in the figure 4.

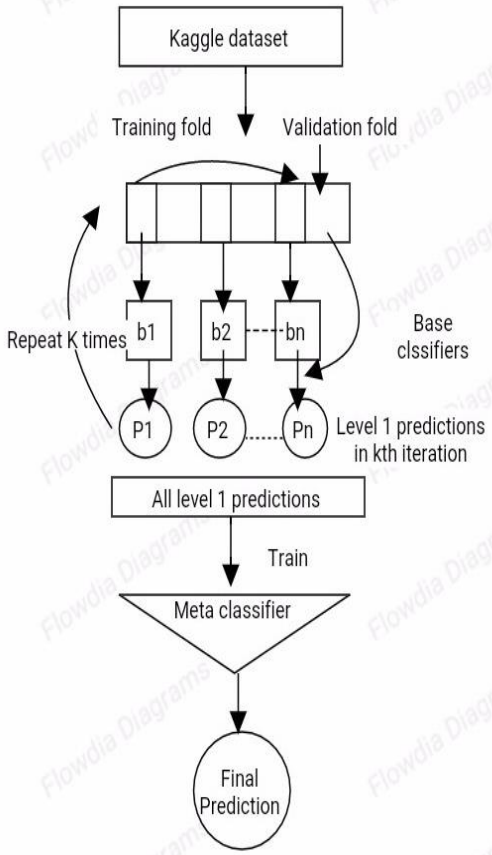

Fig 4:Ensembled stacking model

In this proposed work,the ensemblestacking is used where the performance of individual learners can be increased by using the combining different algorithms.Here,the training data is divided into 10 fold.Each fold divides train and test set.Seven different machine learning classifiers Random Forest,Extra tree classifier, MLP, KNN, XGB, SGD and Adaboost are fed to train the data.Each of these classifier generates different predictions.These predictions are fed into the meta learner to generate the ensemble technique.Here Random Forest is used as the meta learner.In order to improve performance, out of eight base learners,the best performing five models are selected from the comparisons.The final prediction accuracy obtained is 90.21%.

V. RESULTS AND DISCUSSION

The aim of the proposed work is to select top five performing base models and combine these model using stacking ensemble method with k-fold cross validation to achieve improved results.The proposed model is evaluated using various evaluation metrics.In some researches , researchers focused on only some evaluation metrics.But the proposed model is evaluated using sensitivity, accuracy, precision ,specificity,F1 score ,ROC and MCC.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 8, August 2022

Table 2. Model Performance

Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	mathew_ corrcoef
Stacked Classifier	0.902128	0.884615	0.934959	0.866071	0.9090	0.9005	0.804719
Random Forest	0.897872	0.883721	0.926829	0.866071	0.904762	0.89645	0.795852
EXtra tree classifier	0.893617	0.882812	0.918699	0.866071	0.900398	0.892385	3.674389
MLP	0.825532	0.820312	0.853659	0.794643	0.836653	0.82415	0.650192
KNN	0.838298	0.829457	0.869919	0.803571	0.849206	0.83675	0.675997
XGB	0.868085	0.859375	0.894309	0.839286	0.876494	0.8668	0.735734
SGD	0.817021	0.87037	0.764228	0.875	0.8138	0.81961	0.640624
Adaboost	0.817021	0.81746	0.837398	0.794643	0.827309	0.81602	0.633007

From the table 2, it is clear that the proposed stacking ensemble method has the highest performance in all measures.Area under curve and precision recall curve are shown in the figure 3 and figure 4 respectively.Confusionmatrix is given in the figure 5. The highest average area under curve is obtained by Extra tree classifier.When the proposed method is applied on full feature set the accuracy obtained is 89.7 Whereas the proposed ensemble model is applied on reduced feature set by removing two negatively correlated features, the accuracy is increased to 90.21

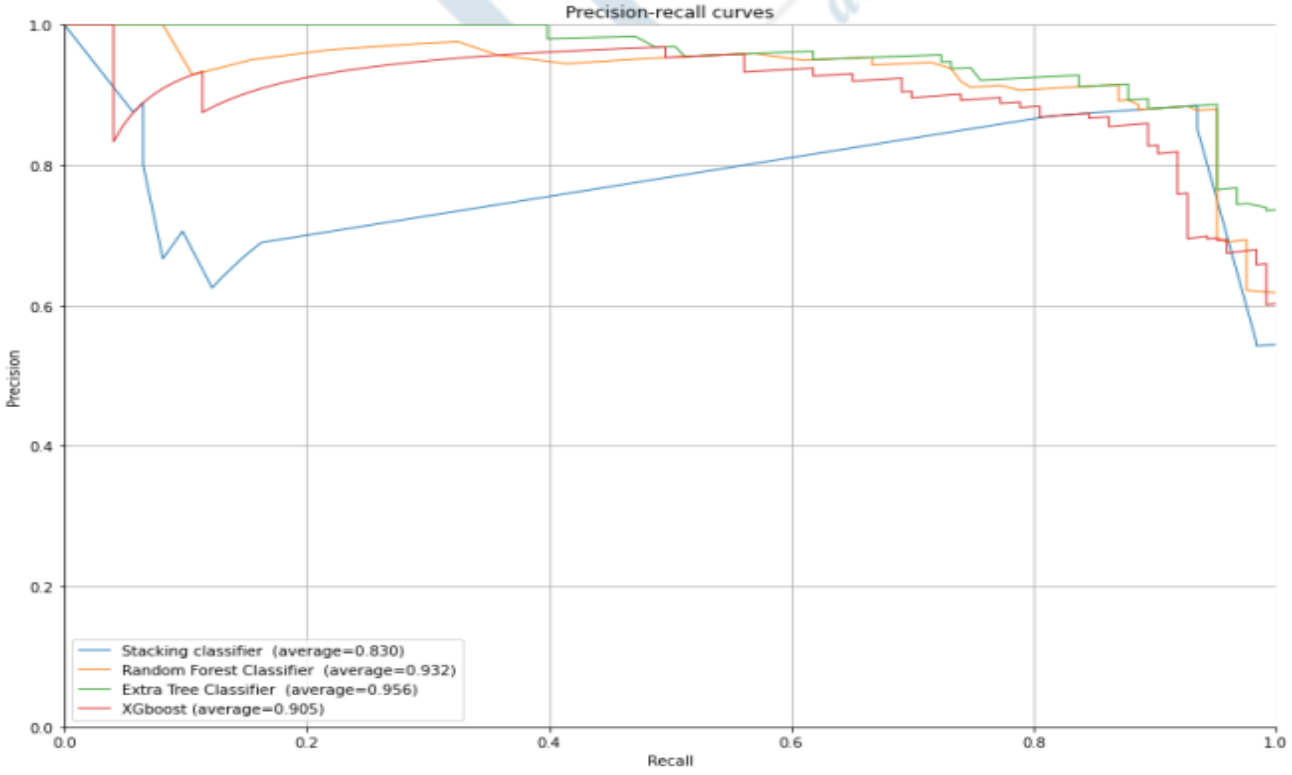

Fig 3. Precision-Recall Curve

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 8, August 2022

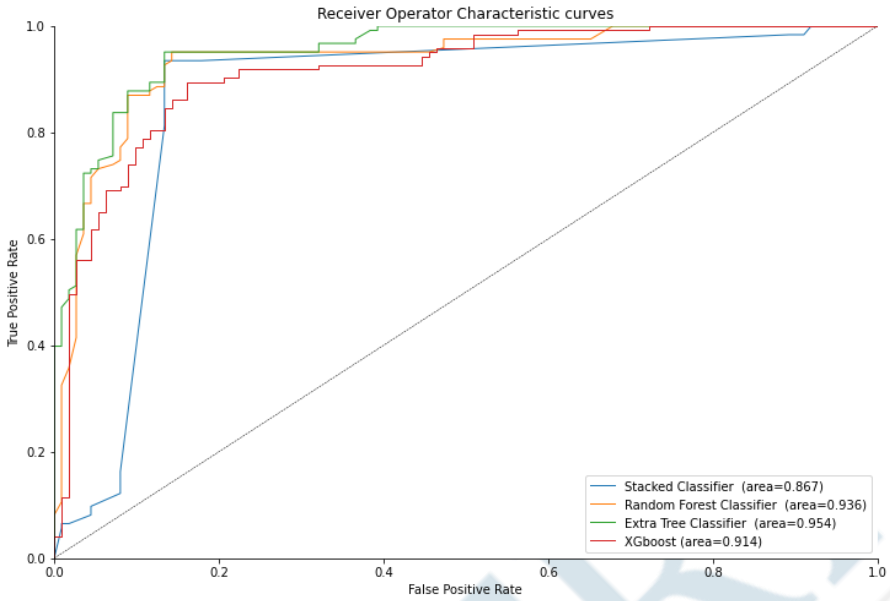

Fig 4:AUC curve

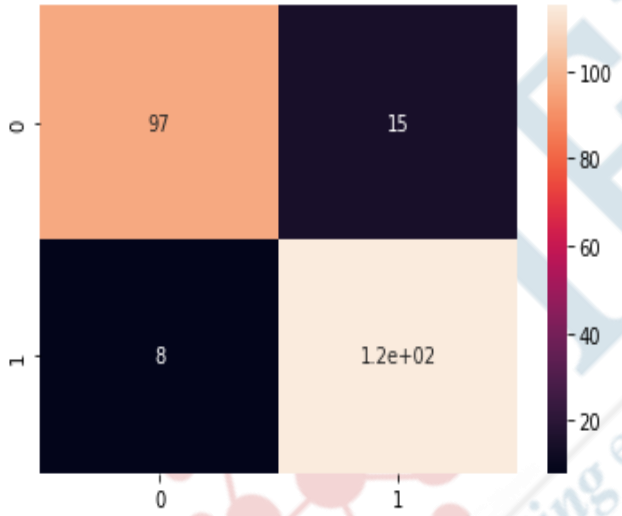

Fig 5: Confusion matrix

VI. CONCLUSION

Large real health care data is collected from various medical organization. Many researchers have used these health care data and apply many machine learning methods for the early diagnosis of heart disease . Now a days smart wearables are a common thing. Different types of IoT sensors can be used to diagnose heart failure in advance.In this work, different weak classifiers are applied on KaggleStatlog-Clevland-Hungary dataset and the predictions are combined to the meta learner to produce the improved result. This stack ensemble method is an effective technique for predicting heart failures in advance.The next step is to improve performance by using various ensemble neural network.

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, (2018), pp. 233–239.

[2] 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

[3] 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:V7-P 160-166(2018).

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

[5] Mohan, SenthilkumarThirumalai, Chandrasegar Srivastava, Gautam-Effective heart disease prediction using hybrid machine learning techniques-IEEE Access Volume 7-Pages-81542-81554 (2019)

[6] Sharma, Sumit Parmar, Mahesh-Heart Diseases Prediction using Deep Learning Neural Network Model-International Journal of Innovative Technology and Exploring Engineering-Volume 9,Issue 3,Pages- 2244-2248(2020)

[7] Sharma, Vijeta Yadav, Shrinkhala Gupta, Manjari-Heart Disease Prediction using Machine Learning Techniques-Proceedings - IEEE 2nd International Conference on Advances in Computing, Communication Control and Networking, ICACCCN 2020-Volume 29,Issue 3 pages-177-181

[8] 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 View project HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES- Journal on Soft Computing,Issue November,Pages-1824-1831(2018).

[9] R. Indrakumari, T. Poongodi, Soumya Ranjan Jena- Heart Disease Prediction using Exploratory Data Analysis- Procedia Computer Science,Volume 173,Issue C,Pages-130-139(2020)

[10] Basma Saleh *, Ahmed Saeidi, Ali al-Aqbi, Lamees Salman - MEDICAL REVIEWS Analysis of Weka Data Mining Techniques for Heart Disease Prediction System- Int J Med Rev,Volume 7,Issue 1,Pages 15-24(2020)

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

[12] Wolpert, David H.- Stacked generalization-Neural Networks,Volume 5-Issue 2-Pages-241-259.

[13] https://towardsdatascience.com/stacking-made-easy-with-sklearn-e27a0793c92b

[14] Detrano R, Salcedo EE, Hobbs RE, Yiannikas J. Cardiac cinefluoroscopy as an inexpensive aid in the diagnosis of coronary artery disease. Am J Cardiol 1986;57 (13):1041–6. https://doi.org/10.1016/0002-9149(86)90671-5

[15] Detrano R, Janosi A, Steinbrunn W, Pfisterer M, Schmid JJ, Sandhu S, et al. International application of a new probability algorithm for the diagnosis of coronary artery disease. Am J Cardiol 1989;64(5):304–10. https://doi.org/ 10.1016/0002-9149(89)90524-9.

[16] Gokulnath Chandra Babu1 , S. P. Shantharajah, Optimal body mass index cutoff point for cardiovascular disease and high blood pressure, Neural Computing and Applications-https: //doi.org/10.1007/s00521-018-3484-3.

[17] Fida, Benish Nazir, Muhammad Naveed, Nawazish Akram, Sheeraz-Proceedings of the 14th IEEE International Multitopic Conference 2011, INMIC 2011-Pages19-24.

[18] Singh, Mayank Gupta, P.K. Tyagi, Vipin Flusser, Jan Oren, Tuncer-Advances in Computing and Data Sciences: Third International Conference-Volume 2, Pages752(2019)

[19] https://www.kaggle.com/datasets/sid321axn/heart-statlog-cleveland-hungary-final.