Investigation of Classifiers Algorithms of ML for Heart Disease Prediction

Ved Prakash Singh[1], Ravi Ray Chaoudhari[2], Krishna Kumar Joshi[3]

[1, 2, 3]Department of Computer Science & Engineering, ITM University, Gwalior, India

Abstract: A ML computer plays an important role in predicting the presence or absence of movement disorders and heart disease. The resting part of the body as compared to the Heart s, is the largest and most concentrated organ in the human body. Data analysis helps in predicting heart disease in the medical field is an important task. Machine learning is recycled in the medical industry throughout the world. The presence or absence of movement disorders and cardiac diseases is a key factor in machine learning. Data analysis helps predict more information and prevents various diseases in medical centers. The main impartial of the research paper is toward predict a patient cardiac disease using an algorithm for machine learning as a random forest is most predictable. A large number of patient data are kept every month. The data stored can be used to predict future diseases. Certain data mining and machine learning technologies are used to forecast heart disease, including artificial neural networks (ANN), decision trees, fuzzy logic, K-Nearest neighbors (KNN), naive bays and vector supporting equipment (SVM). The ultimate objective of this paper is to inspect the best logistic regression which signifies the machine's python learning. The UCI machine learning depot used the data sets of heart disease.

Keywords: K-Neighbor Classifier, Vector Classifier, Tree Decision Classifier, Random Forest Classifier, ML, Heart Disease Prediction.

I. INTRODUCTION

 Heart disease investigation based on the patient data base is usually cover under data mining and analysis. Recently various scientist work to predict the disease using ML techniques. On the collective diagnosis report of large number of patients, it is possible to predict their behaviors. Medical centers around the world collect information about various health problems. It can use various methods of ML to enter this data for useful information. However, the data collected is very large and in many cases the data can be very noisy. These confusing details can be easily explored using different techniques of ML. Therefore, these algorithms have recently been very helpful in accurately predicting the presence or absence of Heart -related diseases.

A. Prediction of Heart disease
Deep computing algorithms primarily use the dimension reduction method to reduce the data set. It is the first step and the most basic procedure used to filter important data sets that have the greatest impact on disease. This prediction includes the next step in which the data is already made for prediction.
1) To summarize the most important data
2) Missing value treatment (Replace the shapes or median values of the blank spaces)
3) Divide data sets into two parts
4) One as a test data and the second as a train data set
5) Use good order over data sets
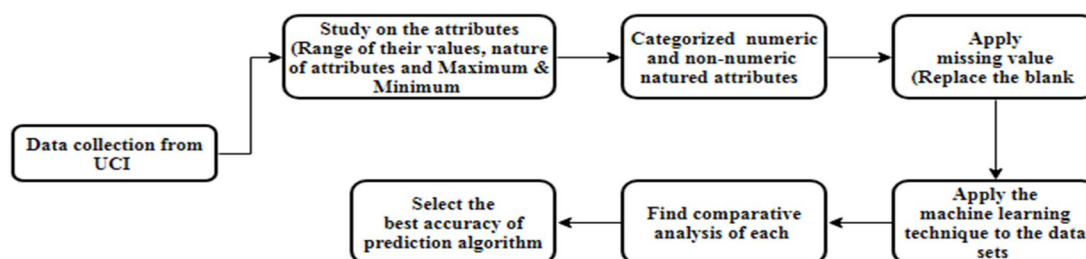6) Get more precision
7) Find the best accuracy algorithm.

Fig. 1 Process of prediction and analysis

B. Depreciation

Data analysis of Heart involves selecting a mathematical representation so that most but not all the distribution of specific data is combined to include the most important information. Information that is deemed to be a work or problem may contain many or many dimensions, but not all of the relevant aspects. A large number of symptoms or functions may affect the complexity of the complex or cause excess, with serious consequences. Therefore, reducing the size of given Heart data is a very important step that needs to be considered when building a model.

C. Release Factor

In this case, the new feature set is based on the original feature set. Feature extraction involves feature conversion. This change is often unrepentant. This change no useful information is lost in the process. Principal component analysis (PCA) was used for feature extraction. The principal component analysis is a general transformation algorithm. In the feature field, find the index with the largest variance and find the corresponding coordinates. This is a global algorithm that provides the best multiplication.

D. Feature Selection

In this case, the bottom set of the original feature is selected. The features are selected using a combination of CFS (Correlation based Feature Selection) and a good size reduction method. The chi-square test is used to select the most important factors.

II. RESEARCH BACKGROUND

1) Abdar et al. (2019), focused on the revolutionary new machine learning tool for identifying patients with coronary artery disease. CAD is one of the most frequent fatal heart diseases. This method of early detection might impact clinical decision-making. Machine-based CAD detection is a major addition to this research since it works better than standard machine learning approaches. They have used eleven classic machine learning techniques on the Z-Alizadeh Sani heart disease dataset to see whether any can successfully diagnose the condition. throughout their tests, they used a normalization strategy for data pre-processing Furthermore, the GA and PSO procedures are now implemented in two ways: First, in one configuration, the GA and PSO are combined with 10-fold cross-validation.

a) First, only classifiers parameters were optimized with GA and PSO, and

b) Parallel GA and PSO-based feature selection and parameter optimization were carried out to choose the most important features in the dataset and the classifier parameters.

2) Diwakar et al. (2020), new developments in the use of machine learning and image fusion for predicting cardiac illness Without a proper diagnosis, it is almost impossible to effectively treat the disease. Preventing disease when you should be is very important since people's lives may be saved. To aid the medical field, machine learning may be used to provide speedy illness diagnoses. Because heart disease is one of the world's most serious and hazardous illnesses today, the appropriate moment for both patients and doctors is when it is found. Also in this work they provide an overview of machine learning and image fusion classifiers that healthcare providers may use to determine if someone has cardiac disease.

3) O'Connor et al. (2010), The safety and efficacy of sertraline for the treatment of depression in patients with heart failure were assessed. People with heart failure are at an increased risk of death due to depression. In a clinical trial on people who had remained depressed even when using nonpharmacologic treatments, Sertraline was found to have no effect on cardiovascular outcomes. It may therefore be an appropriate treatment strategy for patients who meet these two additional conditions:

a) They remain depressed even with the use of nonpharmacologic treatments, and

b) With or without other conditions, the individual has an indication for Sertraline. people with depression and/or heart failure (HF) and coronary heart disease (CHD) who have CHD alone may be candidates for therapy for depression and/or depression and HF

4) Torres et al. (2015), Those who are at greater risk of contracting pneumococcal illness include: People with chronic health issues are more likely to be affected by the burdens associated with pneumococcal illness year-round, particularly invasive pneumococcal illness. It has been shown that those with COPD, asthma, or who smoke, and those with diabetes mellitus, or chronic heart disease, are at greater risk of pneumococcal illness. Short-term and long-term death rates after pneumonia may also be affected by these illnesses and smoking. A patient's quality of life is negatively impacted by community acquired pneumonia, and in particular pneumococcal pneumonia.

5) Tekale et al. (2018). In this paper they have studied different machine learning algorithms. They have analyzed 14 attributes related to Heart patients and predicted the accuracy of different machine learning algorithms such as Decision of the graduate and support machine. From the analysis of the results, it can be seen that the decision tree algorithms provide 91.75% accuracy and the SVM gives 96.75% accuracy. To create a machine learning model that identifies heart disease with an overall accuracy of 99.99%, it would require millions of records with zero non-zero.

6) Alshebly & Ahmed (2019) the purpose of this study is to compare the performance of Artificial Neural Networks (ANNs) and Logistic Regression (LR) categories on the basis of the following criteria: Accuracy, Sensitivity, Specification, Positioning, and Areas under the curve (ROC) in the prediction of Heart. From the experimental results, it is considered that the performance of the ANNs classifier is better than the Logistic Regression model. With an accuracy of 84.44%, sensitivity of 84.21%, specificity of 84.61% and AUCROC of 84.41%. Also, with the final integrated models used, the most important factors that have a clear impact on heart disease patients are creatinine and urea.

7) Norouzi et al. (2016), proposed the prediction analysis for Heart failure. This model may accurately describe more than 95% accuracy of prediction. This paper actually limitation work considering the GFR value at much extent in time interval.

8) Arasu & Thirumalaiselvi (2017) They proposed one dimensional predictive analysis. Basically, this model based on the regression and random forest. For big data application there are already RF, Classified and Regression Tree and C4.5 works better. In the exploration of Heart prediction, they proposed a new hybrid algorithm which enhance the prediction graph. This model assigns the importance of each attribute to a dataset and to carry the priorities of the classification process.

9) Ravindra et al. (2018) in this paper, the effectiveness of the proposed scheme is evaluated in terms of the sensitivity, specificity and accuracy of the phases. The results reveal the accuracy of the total separation of 94.44% obtained by combining the 6 attributes. It can be concluded that the SVM-based method found may be a member of kidney.

10) Rahman et al. (2019) this paper introduces the concept of detecting the presence of Heart disease using a computerized systematic study model, considering the patient's ECG signal. In their research, they found an accuracy rate of 97.6% which is an average using both the QT and RR intervals, comparing the accuracy obtained when using one of the signals.

11) Hele et al. (2018). In this article, it has been proposed that in the medical network a neural-trained algorithm to detect early Heart disease. The results revealed that NN-GA was more active than others in the past and was able to detect heart more effectively. Future research could focus on studying some such teaching methods for training NNs to improve the performance of NNs in real applications.

12) Khamparia et al. (2019). This research paper presents a framework for the number of deep neural pathology of Heart disease accident using focal auto encoder installation. The model is built using embedded auto codes that contain two auto codes arranged in a fashionable form with one deep max component.

13) Slovakia and Verma (2017). In this paper, the comparative performance of specific and integrative students is analyzed in the kidney set of the UCI reading library. The performance of CKD using the predictive analysis need data collection, filtering of data, missing value treatment. They model give accuracy, recall, f-rate, and the ROC-Curve as metrics to compare classroom performance. They are also developing the Wrapper and hybrid method to select the desire feature.

14) Subasi et al. (2017). In this study, different machine learning methods are used to diagnose kidney This study has shown that RF isolation results in significantly higher performance during the partition operation.

15) Zhang et al. (2018) in this work, they build two virtual neural networks, one of which is classical MLPs, and the other combines the selection of the LASSO feature and is highly functional with a small but deep structure. The results indicate that the performance of each model is the same, both of which obtain high accuracy.

16) Leyva et al. (1997), The level of serum uric acid may be used as an indicator of poor oxidative metabolism in patients with chronic heart failure. In individuals with heart failure, blood uric acid levels are inversely correlated with markers of functional capability. According to the link found between blood uric acid levels and MVO2, chronic heart failure seems to be accompanied with a decline in oxidative metabolism. In clinical circumstances in which hypoxia is present, elevated blood uric acid concentrations have been seen. They wanted to see whether blood uric acid concentrations linked to functional ability and disease severity in heart failure patients.

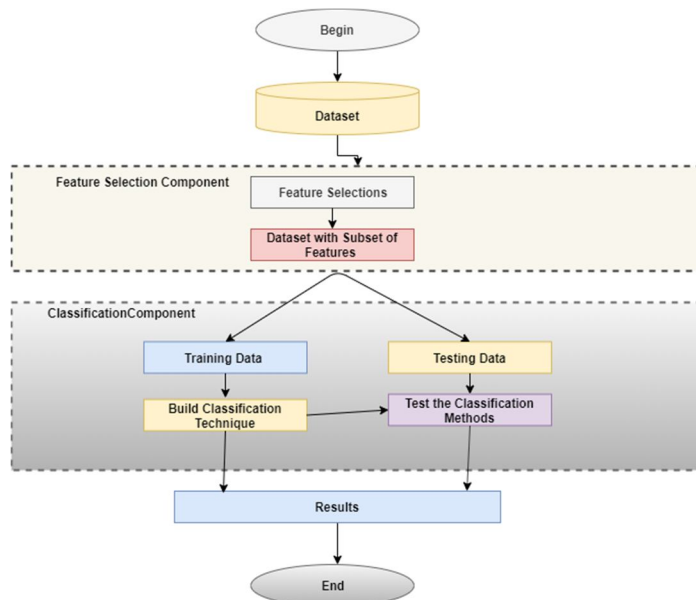### III. ARCHITECTURE OF THE EXPERIMENTS

Fig. 2 Architecture of the ML experiments

IV. ML ALGORITHMS FOR INVESTIGATION OF DISEASE

A. Curvilinear Regression

Name retrieval can be defined as measuring and analyzing the relationships between independent or dependent variables. Dissolution can be divided into two categories: Direct undo and direct undo. The normalization of normal thinking is reproduced with a straight line. It is primarily used to measure the relative or multidimensional variables.

The translation of rational insight can begin with the interpretation of general cognitive functions. A logical function is a sigmoid function that takes a real value between 0 and 1. It is defined as

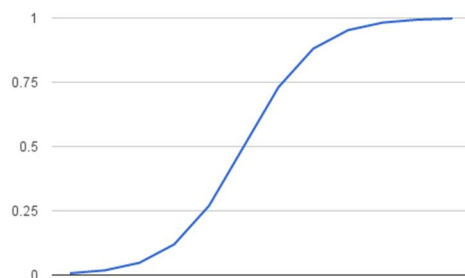$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

Fig. 3 Representing the non-linear curvilinear regression

Let's take a look at the functionality of a line in an unequal return model.

$$t = \beta_0 + \beta_1 x$$

So the non-linear curvilinear regression Equation becomes

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

B. Support Vector machine

Support vector machines are a very popular way of studying the monitored machine (using pre-programmed target objects) that can be used as processors and predictors.
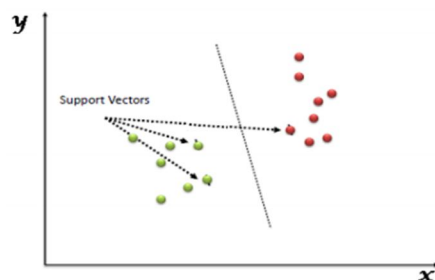

Fig. 4 A presentation of SVM

It is the extension of the linear regression with involvement of clusters. A main trend line iterates and continue till the optimization reach.

C. K - Nearest Neighbor

K-Near method is one of the best ways to differentiate data. It is used for non-swearing classification functions about data and little or no information about the distribution of data. The algorithm finds the nearest k points closest to the data points not found in the training set and the amount of data points obtained from it. In classification settings, the nearest neighbor algorithm basically creates the most votes among the most similar K conditions for some "intangible" recognition.

$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + \ldots + (x_n - x'_n)^2}$$

However, there are some metrics that are appropriate for specific settings, such as Manhattan, Chebyshev, and the Hamming distance.

D. Decision Tree

This method is used mainly for separation problems. Make it easy with continuous signs and paragraphs. This algorithm divides the population into two or more similar sets based on the most important predictors. The algorithm tree first lists the entries for each attribute. The data set is segmented by a variable or predictor with the largest or shortest gain. These two steps are performed in duplicate with the remaining structures

$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$

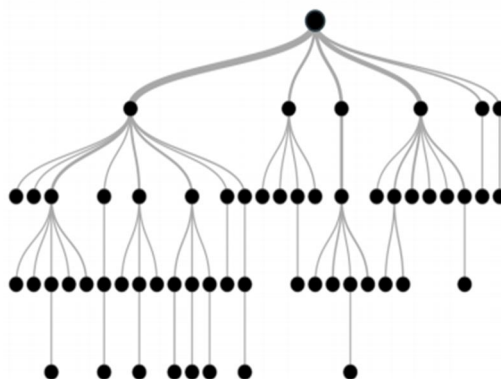$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$

Fig. 5 Shows the Decision Tree

E. Random Forests

Random Forest is also a supervised learning algorithm. This method can be used for recording and classification functions, but generally improves the performance of classification functions. As the name implies, a random forest path looks at many decision trees before making a result.
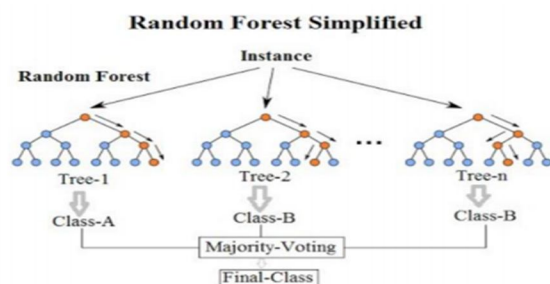

Fig. 6 Random Forest

Random Forest (RF) creates many decision trees during training. Summarize all tree predictions with the final prediction; Partition model or average forecast when making final decisions using a set of results, a combination of techniques is used.

F. Gaussian_NB

Gaussian Naive Bayes is simply understanding as simple distribution with control curve. This algorithm based on the conditional probability based on the occurrence of data under Gaussian plane.

G. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) (Duda et al., 2001) is the most widely used methods for reducing size and separation. With given set of data their feature has been estimated and centroid and covariance extract the reducible sets of data.

LDA makes simple assumptions about your data:

1) That your data is Gaussian, that each variable is formed as a bell curve when plotted.
2) For each attribute having the same value, those values of each variable differ by the mean of the same value on average.

H. Ada_Boost_Classifier

AdaBoost is short for Adaptive Boosting. Basically, Ada Boosting was the first successful add-on algorithm designed for binary classification.

AdaBoost could be apply even in an unequal group. It has good properties for general practice that can be proven to increase accuracy of prediction.

I. Gradient Boosting Classifier

Gradient intensification is one of the most competitive algorithms that works with the goal of strengthening weak learners by shifting focus from problematic observations that were difficult to predict in the past and to meeting weak readers, general decision trees. It builds the model in a way that is almost identical to other lifting methods, but proves them by allowing good performance for mysterious losses. For starters, we fit the model in the model that produces 75% accuracy and the remaining undisclosed difference is taken in the error name. After that we fit another model in the error set to pull an additional descriptive component and add it to the original model, which should improve the overall accuracy:

Error = G(x) + Error2

$$f_0(x) = \frac{\arg\min}{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$$

J. H. Quadratic Discrimination Analysis

Quadratic discriminant analysis is performed directly as in the direct discrimination analysis except that we use the following functions based on the matrices of each category.

$Di(X) = -1/2\ LN(|Si|) -1/2\ (X-Y)^{T} Si^{-1} (X-Y)$

$Si\ (X) = di\ (X) + LN(\pi)$

K. MLP Classifier

The MLP can be viewed as a classical regression classifier where the input is first modified using a non-linear regression learned \ Phi. This modification puts the input data into a space where it is directly split. This middle layer is called the hidden layer. One hidden layer is enough to make MLPs a universal standard. An MLP (or Artificial Neural Network - ANN) with a single hidden layer can be represented in bold as follows:

Fig. 7 Neural network

V. PROPOSED DATA MEASUREMENT MEASURES

1) STEP 1: find the essential attributes of Heart data sets. An attribute with low and high data is selected for statistical analysis.
2) STEP 2: Estimate the accuracy of the data in the statistical analysis.
3) STEP 3: Measure the methods and methods of treatment that are missing.
4) STEP 4: Fill in the missing values by means of and between data sets.
5) STEP 5: Divide train test data and analysis by 70:30.
6) STEP 6: Perform a ML algorithm on train data sets
7) STEP 7: Measure accuracy with test data sets.

VI. METHODOLOGY

1) *Step 1: Data recovery*
{
Outline of data
Identify and complete the output
Identify and manage missing information
Appropriate strategies are used
Replace the adjective and spelling
}

2) *Step 2: Model Selection*
{
Checking the amount of data (classes)
An in-depth study of Algorithm Sorting's
}

3) *Step 3: Implementation Example using Python*
{
Import Data Entry all models together using Python
}

4) *Step 4: Performance classification*
{
Calculate Accuracy using the "Operation" operator to analyze the result by the calibration process
}

5) *Step 5: Comparison Result*
{
Comparing accuracy between all models by comparing the result with all the proposed algorithms for learning
Calculate the final output of each proposed algorithm
Find the best of all.
}

VII. PSEUDO CODE

Let a1 = {a1, q2, a3,… an} be the given data

A = {}, a set of Algorithms programmers

M = mean and Median {c1, c2, c3, … cn}, a set of

Z = Means, the mediator of M.

For (i = vacant, i = 0, i ++);

{

For (j = vacant, j = 0, j ++);

}

Apply in-depth Algorithm

f = Depth (Mod: Data);

Let D = {d1, d2, d3,… Dn} be the given data

E = {E1, E2, E3,… En}, a set of ensemble classifiers

C = {c1, c2, c3, … cn}, a set of arrays

X = training set, X D

Y = test set, Y D

K = class of Meta levelifier

L = n (D)

Because I = 1 to L do

M (i) = A model trained to use E (i) in X

The following i

M = M K

Result = Y divided by M

VIII. CONCLUSION & FUTURE SCOPE

This paper mainly explores major ML algorithm for the prediction of the heart analysis. The prediction in terms of accuracy and precision give the better understanding of the ML algorithms which could be further use in the real time data analysis and prediction using the graphical user interface. The main purpose of this research paper is to analyze the efficiency of various classification algorithms for the diagnosis of Heart disease. Databases can be used to forecast the incidence of a possible illness. So, in future a graphical use interface will have do the same work as doctor does for a specific disease.

REFERENCES

[1] Abdar, M., & Ksi ¸a ˙zek, W. (2019). A new machine learning technique for an accurate diagnosis of coronary artery disease. Computer Methods and Programs in Biomedicine, 179, 104992.

[2] Alshebly, O. Q., & Ahmed, R. M. (2019). Prediction and Factors Affecting of Heart Disease Diagnosis using Artificial Neural Networks Model and Logistic Regression Model. Iraqi Journal of Statistical Science, 28, 1-19.

[3] Arasu, S. D., & Thirumalaiselvi, R. (2017, February). A novel imputation method for effective prediction of coronary disease. In 2017 2nd International Conference on Computing and Communications Technologies (ICCCT) (pp. 127-136). IEEE.

[4] Diwakar, M., Tripathi, A., & Joshi, K. (2020). Latest trends on heart disease prediction using machine learning and image fusion. Materials Today.

[5] Hore, S., Chatterjee, S., Shaw, R. K., Dey, N., & Virmani, J. (2018). Detection of disease: A NN-GA-based approach. In Nature Inspired Computing (pp. 109-115). Springer, Singapore.

[6] Khamparia, A., Saini, G., Pandey, B., Tiwari, S., Gupta, D., & Khanna, A. (2019). KDSAE: Heart disease classification with multimedia data learning using deep stacked autoencoder network. Multimedia Tools and Applications, 1-16.

[7] Leyva, F., Anker, S., & Swan, J. W. (1997). Serum uric acid as an index of impaired oxidative metabolism in chronic heart failure. European Heart Journals, 1(8), , 858-865.

[8] Liden, R. C., Wayne, S. J., Zhao, H., & Henderson, D. (2008). Servant leadership: Development of a multidimensional measure and multi-level assessment. The leadership quarterly, 19(2), 161-177.

[9] Maurya, A., Wable, R., Shinde, R., John, S., Jadhav, R., & Dakshayani, R. (2019, January). Heart Disease Prediction and Recommendation of Suitable Diet Plan by using Machine Learning. In 2019 International Conference on Nascent Technologies in Engineering (ICNTE) (pp. 1-4). IEEE.

[10] Norouzi, J., Yadollahpour, A., Mirbagheri, S. A., Mazdeh, M. M., & Hosseini, S. A. (2016). Predicting renal failure progression in Heart disease using integrated intelligent fuzzy expert system. Computational and mathematical methods in medicine, 2016.

[11] O'Connor, C. M., & Jiang, W. (2010). Safety and Efficacy of Sertraline for Depression in Patients with Heart Failure. Journal of the American College of Cardiology, 59(9), 0735-1097.

[12] Rahman, T. M., Siddiqua, S., Rabby, S. E., Hasan, N., & Imam, M. H. (2019, January). Early Detection of Heart Disease Using ECG Signals Through Machine Learning Based Modelling. In 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST) (pp. 319-323). IEEE.

[13] Ravindra, B. V., Sriraam, N., & Geetha, M. (2018). Classification of non- and heart disease using SVM neural networks. International Journal of Engineering & Technology, 7(1.3), 191-194.

[14] Salekin, A., & Stankovic, J. (2016, October). Detection of Heart disease and selecting important predictive attributes. In 2016 IEEE International Conference on Healthcare Informatics (ICHI) (pp. 262-270). IEEE.

[15] Sisodia, D. S., & Verma, A. (2017, November). Prediction performance of individual and ensemble learners for heart disease. In 2017 International Conference on Inventive Computing and Informatics (ICICI) (pp. 1027-1031). IEEE.

[16] Subasi, A., Alickovic, E., & Kevric, J. (2017). Diagnosis of Heart disease by using random forest. In CMBEBIH 2017 (pp. 589-594). Springer, Singapore.

[17] Tekale, S., Shingavi, P., Wandhekar, S., & Chatorikar, A. (2018). Prediction of Heart disease using machine learning algorithm. Int. J. Adv. Res. Comput. Commun. Eng, 7, 92-96.

[18] Torres, A., Blasi, F., & Dartois, N. (2015). Which individuals are at increased risk of pneumococcal disease?

[19] Wickramasinghe, M. P. N. M., Perera, D. M., & Kahandawaarachchi, K. A. D. C. P. (2017, December). Dietary prediction for patients with heart disease by considering blood potassium level using machine learning algorithms. In 2017 IEEE Life Sciences Conference (LSC) (pp. 300-303). IEEE.

[20] Zhang, H., Hung, C. L., Chu, W. C. C., Chiu, P. F., & Tang, C. Y. (2018, December). Heart Disease Survival Prediction with Artificial Neural Networks. In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 1351-1356). IEEE.

[21] Ordonez .C.,(2006).” Association Rule Discover with the Train and Test Approach for the Heart Disease Prediction”, IEEE Transactions on Information Technology in Biomedicine, Vol. 10, No. 2, PP. 334-343.

[22] Ozcift.A and Gulten.A.,(2011). “Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms”, Journal of Computer Methods and Programs in Biomedicine, Vol.104, PP.443-451.

[23] Park.Y.J,Chun.S.H and Kim.B.C.,(2011). “Cost-sensitive case-based reasoning using a genetic algorithm: Application to medical diagnosis”, Journal of Artificial Intelligence in Medicine, Vol.51, PP.133-145.

[24] Pfister.R ,Barnes.D,Luben.R.N,Khaw..K.T,Warehm.N.Jand,Langenberg.C.,(2011).“Individual and cumulative effect of type 2 diabetes genetic susceptibility variants on risk of coronary heart disease”, Journal of Diabetologia, Vol.54, PP.2283-2287.

[25] Polat.K and Gunes.S.,(2009).” A new feature selection method on classification of medical datasets: Kernel F-score feature selection, Journal of Expert Systems with Applications, Vol. 36, PP.10367–10373.

[26] Setiawan.N.A,Venkatachalam.P.A and Hani.A.F.M..,(2008).” A Comparative Study of Imputation Methods to Predict Missing Attribute Values in Coronary Heart Disease Data Set”, Journal in Department of Electrical and Electronic Engineering, Vol.21, PP. 266–269.

[27] Tan. K.C, TeohE.J,Yu.QandGoh.K.C.,(2009). ”A hybrid evolutionary algorithm for attribute selection in data mining , Journal of Expert system with application Vol.36, PP.8616-8630.

[28] Tartarisco, G., Tonacci, A., Minciullo, P.L. et al. ClinMol Allergy (2017) 3 “The soft computing-based approach to investigate allergic diseases: a systematic review”, 15: 10. https://doi.org/10.1186/s12948-017-0066-

[29] Uyara.K and İlhana.A.,(2017). “Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks”, 9th International Conference, Procedia Computer Science 120, Pp:588–593.

INTERNATIONAL JOURNAL
FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 ⓦ (24*7 Support on Whatsapp)