International Journal of Engineering Applied Sciences and Technology, 2023
Vol. 8, Issue 07, ISSN No. 2455-2143, Pages 01-06
Published Online November 2023 in IJEAST (http://www.ijeast.com)

HEART DISEASE PREDICTION BY EMPLOYING BEST RISK FACTORS USING MACHINE LEARNING

Annette Nellyet
Department of IT
Britts Imperial University College
Sharjah, UAE

Salmodin Ansari
Department of IT
Nest Academy of Management Education
Dubai, UAE

Abstract— **Heart disease is the major global cause of death in the modern era. Heart illness might be difficult to diagnose. It requires a number of expensive diagnostic tests to be found. The scientific community is undergoing a transformation thanks to machine learning (ML), a subset of artificial intelligence. In this paper, exploratory data analysis (EDA) utilizing multivariate analysis is utilized to find correlations and outliers in the data. The study's dataset consists of 270 records with 14 variables, including age, chest pain type, blood pressure, blood glucose level, resting ECG, heart rate, and chest pain. Logistic Regression algorithm is employed on the dataset to predict heart disease on the basis of chosen risk factors that are obtained using recursive feature elimination incorporating random forest classifier method. The paper entails pre-processing methods, proposed algorithms, evaluation metrics and an accurate prediction based on the methodology.**

Keywords— **Exploratory Data Analysis, Logistic Regression, Heart Disease, Recursive Feature Elimination, Random Forest classifier**

I. INTRODUCTION

Cardiovascular disease has now become the main threat for death of people in this era. Cardiac problem affects the main blood vessels that supply the heart muscles. Blood flow to the heart and other vital organs is impeded by plaque that forms in these blood channels and is made up of cholesterol deposits. It can result in a heart attack, stroke, or heart failure if this is not treated. To identify and detect heart disease different diagnoses are done such as MRIs, CT scan, various blood tests, video x-rays, ECGs or Holter monitoring. The data collected from these tests are called medical big data which are kept across numerous databases, none of which offer any particular value by themselves. However, if these data are combined with and analyzed employing machine learning and AI, then it is possible to produce diagnostic data that can save lives and keep expenses down.

Machine learning is a subset of Artificial Intelligence that helps in data mining from large datasets and exact valuable information from them. The purpose of this analysis is to look at the potential of machine learning and how it may be used to anticipate cardiac disease at an early stage which will save the time, money and life of people. But, to analyze and predict, the most recent dataset is required which will be a reliable source to train the model to predict heart disease. In this study, a Dataset from Kaggle containing 270 records and 14 attributes was considered for analysis. The main objectives of this study are to learn how an optimal model can be used to forecast heart illness and how machine learning algorithms can be utilized to detect heart disease. The proposed methodology for predicting heart disease includes pre-processing, EDA, features selection method that is performed with the help of Recursive Elimination Method incorporating Random Forest Classifiers and logistic regression model.

II. LITERATURE REVIEW

Numerous studies have been conducted on this subject by various researchers, and various approaches to the prediction of cardiac disorders using machine learning algorithms have been developed. This study can help those whose medical histories suggest they have a higher chance of developing a heart problem. By identifying specific heart disease characteristics or symptoms, such as chest pain, high blood pressure, etc., to diagnose the condition with fewer medical tests and effective therapies, so that it may be treated appropriately. A method to improve the precision of heart disease prediction using machine learning has been put forth by Saboor et al [5]. Several methods were used in the study, including logistic regression, k-nearest neighbors, and support vector machines, to analyze a dataset of patient medical information. The outcomes showed that the suggested method outperformed existing approaches by a substantial margin, achieving an amazing prediction accuracy of 92.5%. The WEKA tool was used by Khourdifi & Bahaj et al. [17] to implement Particle Swarm Optimization, Ant Colony Optimization, and Fast Correlation-Based Feature Selection algorithms on the UCI heart disease dataset. Dimension reduction is the ideal preprocessing strategy, and redundant data can be eliminated with the use of K-Nearest Neighbors and Random Forest classifiers. It was observed that RF classifiers obtained the highest accuracy of 99.7%. Su et al. [15] investigation into the effectiveness of random forest algorithms for predicting heart disease based on genetic and clinical characteristics. With a sensitivity of 75% and a specificity of 82%, the random forest model successfully predicts heart disease with an overall accuracy of 80%. For better prediction performance, the study underlines the significance of including genetic data along with clinical factors. The outcomes illustrate the value of random forest algorithms in determining cardiovascular risk and show how machine learning can be used to improve patient care. Amarbayasgalan T et al. [18] proposed a Deep Auto encoder-based neural network for the prediction of heart disease. For dimension reduction, an artificial neural network (ANN) with auto encoders is used. This auto encoder has three hidden layers made up of 4, 1, and 4 neurons each. The author use this encoder for feature learning and Relu activation for model tuning. The model's accuracy was 83.53%, which was adequate but not sufficient. Ali et al. [6] examined a number of machine learning classifiers for heart disease prediction. He built a powerful model using a variety of methods for prediction. In comparison to KNN, SVM, and other algorithms, the accuracy attained by the random forest algorithm is 72.59% for 12 features. In order to forecast cardiac disease using machine learning methods, K.G. Dinesh et al.[19] employed a Cleveland Clinic dataset that is available on the UCI Machine Learning repository. He applied some machine learning methods, including SVM, gradient boost, naive bayes, logistic regression, and random forest. Handling missing values and scaling the data with common scalars were both engaged in the preprocessing of the data. It was discovered after employing ML models that the logistic regression had the highest accuracy when compared to other models, with a score of 86.51%. Similar to this, accuracy values for other models such as random forest, gradient boost, Naive Bayes, and SVM were 80.89, 84.26, 84.26, and 79.77, respectively.

As per the review of other papers, It is observed that many of the researchers have used different models to predict heart disease; however, their accuracy was found to be comparatively low as all the factors were considered for the study. Selecting the appropriate risk variables will help healthcare workers quickly diagnose the disease and will improve the prediction model's effectiveness. In this study, a logistic regression model is used to assess the dataset's maximum accuracy using the RFE approach in order to determine the optimal risk variables for disease prediction. The relationship between the attributes has also been illustrated through data visualization.

III. RESEARCH METHODOLOGY

In this paper, the machine learning algorithms such as Random Forest Classifier and Logistic Regression are used to predict the disease. This research can be useful for the medical analysts to identify heart disease. The proposed methodology comprise certain steps that are given in the figure below:

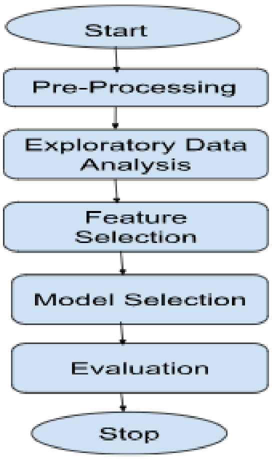


Fig. 1. Proposed Model for Heart disease prediction

The dataset contains categorical variables and numeric variables. The categorical variables used are Sex, EKG results, Exercise angina, Slope of ST, Thallium and Heart disease. Whereas the Numeric variables used are Age, BP, Cholesterol, FBS, Max HR, ST depression, Number of vessels fluro.

Pre-processing: This is the stage where the data is explored. It employs identifying missing values, dealing with inconsistencies in data and performing normalization to fit into the model. It is observed that the dataset has no missing values. For the categorical variables, the get dummy function is used to convert it into dummy or indicator variables. With the help of this function the data validity is improved. For Numeric value, standard scaling factor is employed. It was also observed that there were outliers in the data making it more inefficient for the study.

Exploratory Data Analysis: In this stage various attributes are analyzed inorder to determine the relationships with dependent and independent factors. Heat map and count plot are used to justify the multivariate analysis.

Feature Selection: Recursive Feature Elimination (RFE), a common feature selection technique that is based on the Random Forest Classifier, is taken into consideration for the study in order to increase the suggested model's accuracy [16]. In the analysis, a random forest classifier model is built on all of the predictors variables and determines the relevance of each predictor. In every iteration two less significant features are removed and the predictors that are more important are kept to re-create the model. In this way the best risk factors are obtained from this method.

Model Selection: After obtaining the best risk factors, four popular machine learning models namely Support Vector Machine, Logistic Regression, XGBoost and Decision Tree were employed and the accuracies were compared.

It is observed from the study that the Logistic regression model is an appropriate model for further analysis.

Logistic regression model can be expressed as:

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

Where, $\rho(x)$ is interpreted as the probability of the dependent variable Y, $\beta_0$is intercept and $\beta_1$is slope of the function x .

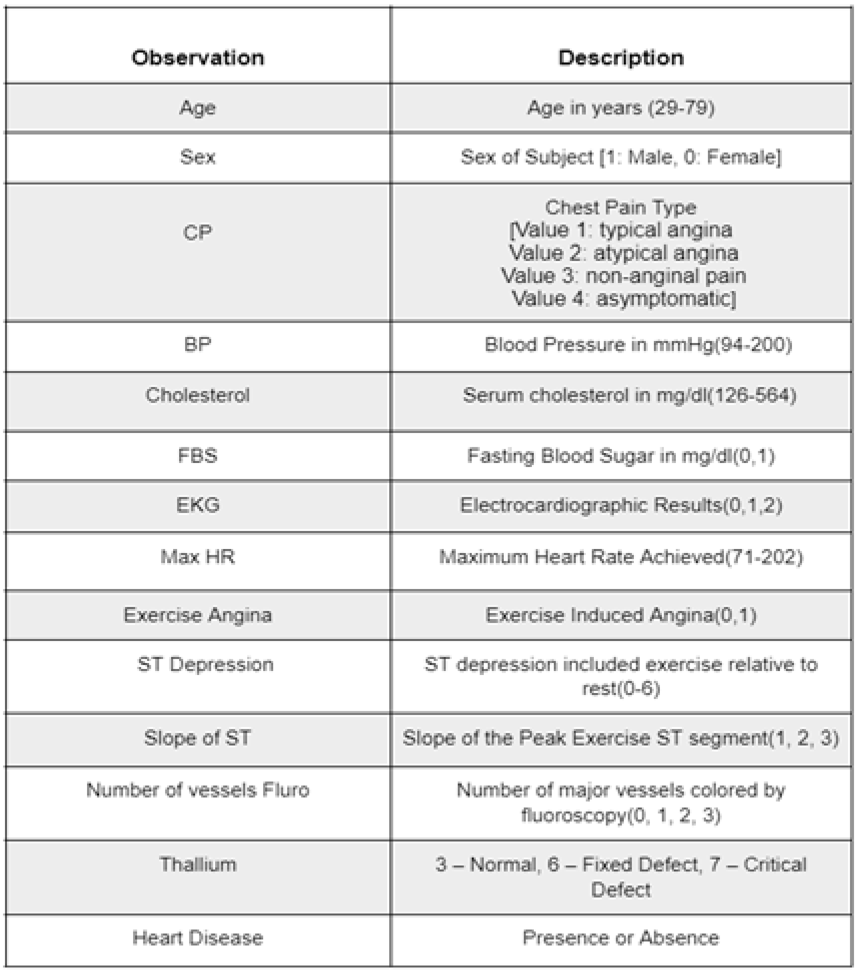
Evaluation: To determine the efficiency of the proposed model, this paper employs classification matrix and ROC-AUC Curve report.

The accuracy of classification model is evaluated using performance metrics and is expressed as% Accuracy = $\frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \times 100$, (2)

Where TP stands for true positive, TN stands for true negative, FP stands for false positive, and FN stands for false negative.

IV. EXPERIMENT AND RESULT

The dataset used for the study is publicly available on UCI repository and on KAGGLE website [1]. Which contains 270 patient records with 14 attributes as shown in tabular column below:



Observation	Description
Age	Age in years (29-79)
Sex	Sex of Subject [1: Male, 0: Female]
CP	Chest Pain Type [Value 1: typical angina Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptomatic]
BP	Blood Pressure in mmHg(94-200)
Cholesterol	Serum cholesterol in mg/dl(126-564)
FBS	Fasting Blood Sugar in mg/dl(0,1)
EKG	Electrocardiographic Results(0,1,2)
Max HR	Maximum Heart Rate Achieved(71-202)
Exercise Angina	Exercise Induced Angina(0,1)
ST Depression	ST depression included exercise relative to rest(0-6)
Slope of ST	Slope of the Peak Exercise ST segment(1, 2, 3)
Number of vessels Fluro	Number of major vessels colored by fluoroscopy(0, 1, 2, 3)
Thallium	3 – Normal, 6 – Fixed Defect, 7 – Critical Defect
Heart Disease	Presence or Absence

Table -1 Table list for attributes from the dataset [1]

The inconsistent data which are classified as outliers were removed from the features to improve the efficiency and are depicted as shown in Figure 2.

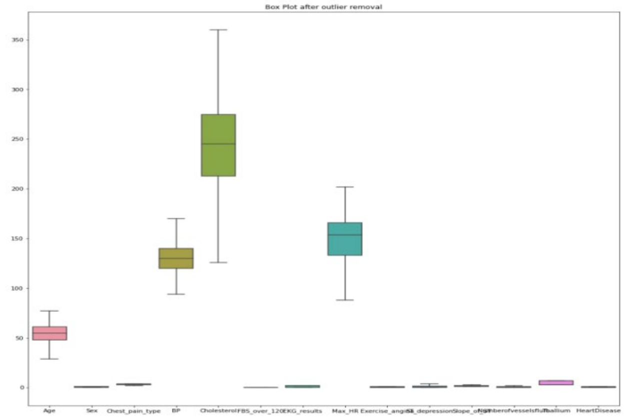


Fig. 2. Output from Box plot after removing the outliers

Figure 3 depicts the correlation between all the attributes associated with the data. The heat map clearly indicates that Chest pain type, Max_HR, Exercise angina, ST depression and Thallium has correlation with the target attribute - Heart Disease with a threshold of ±0.4.

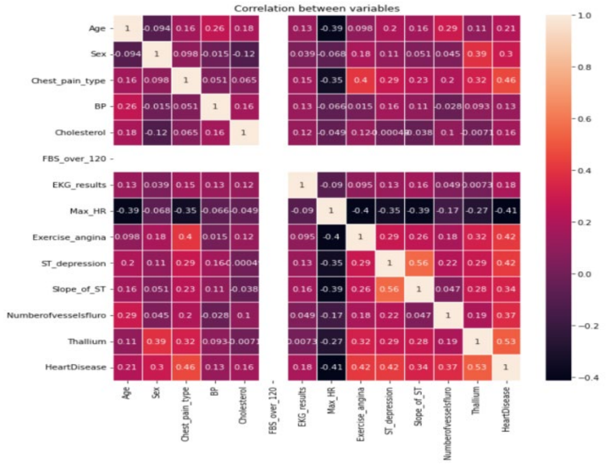


Fig. 3. Heatmap for Correlation between variables

In addition to correlation, Multivariate analysis is conducted on BP, Cholesterol and Age factors to determine the relationship with heart disease.

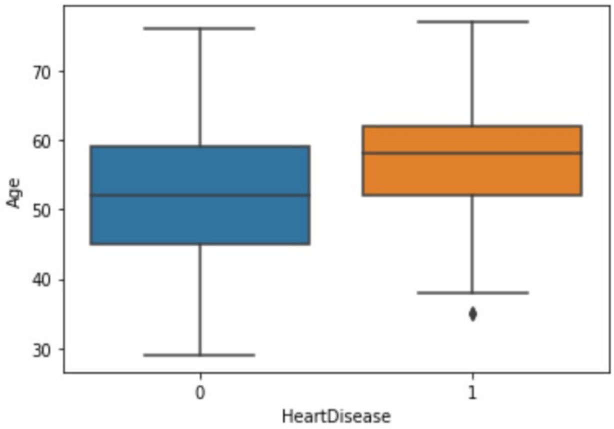


Fig. 4. BoxPlot of Age vs Heart Disease

Figure 4, clearly shows that heart disease is dependent on age. The probability of people of age above 50 years are having more chances of suffering from heart disease than that of the people below it.

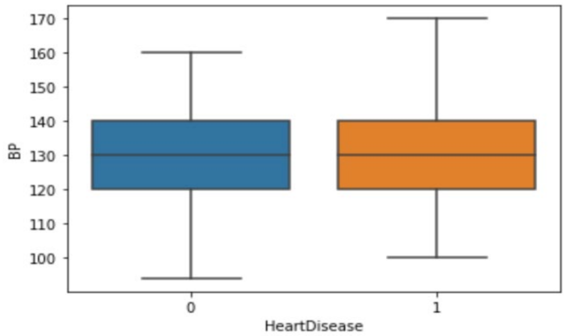


Fig. 5. BoxPlot of BP vs Heart Disease

Figure 5 clearly indicates that there is no significant relationship between target variable and BP.

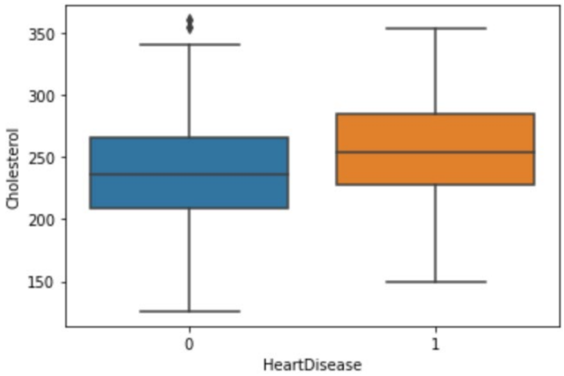


Fig. 6. BoxPlot of Cholesterol vs Heart Disease

Figure 6 indicates that the higher the cholesterol level greater is the risk of being prone to heart disease. Hence there is a relationship between cholesterol and target variable.

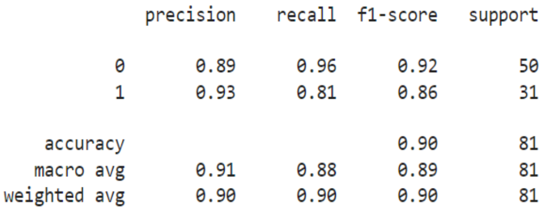
Age, Sex, Chest_pain_type, BP, Cholesterol, Max_HR, ST_depression, Number of vesselsfluro, and Thallium are the best features identified by the RFE feature selection model utilizing the Random Forest classifier.

With a ratio of 70:30, the complete dataset is divided into training and testing sets. The machine learning models were trained on the dataset and the accuracies were compared.

Table -2 Comparison of accuracies of different models

Model	Accuracy
SVM	81.4%
Decision Tree	76.5%
Logistic Regression	90.0%
XGBoost	81.4%

From the above table, the accuracy of the logistic regression model was observed to be 90% which is also depicted as shown in Table 2. It is also observed that f1-score is found to be 0.92 which clearly proves that model is accurate and highly efficient for the dataset considered for study.



```
              precision    recall  f1-score   support

           0       0.89      0.96      0.92        50
           1       0.93      0.81      0.86        31

    accuracy                           0.90        81
   macro avg       0.91      0.88      0.89        81
weighted avg       0.90      0.90      0.90        81
```

Fig. 7. Classification Report for the proposed model

The outcome of the result showed that there are 48 True Positive values, 25 True Negative values, 2 False Positive values and 6 False Negative values from the test data. The results clearly indicate that false negative and false positive rates are comparatively low. Hence the efficient model to predict heart disease. The model's performance could be enhanced by adding other attributes corresponding to heart disease as well as by populating the dataset.

Figure 8 shows the receiver operating characteristics (ROC) curves for the proposed model. The test set prediction accuracy of the logistic regression model is 90% with ROC of 0.86 for the selected 9 attributes of the 270 patients dataset. As observed, the model's AUC is greater and hence the proposed model is more accurate in discriminating between patients with heart disease and without.

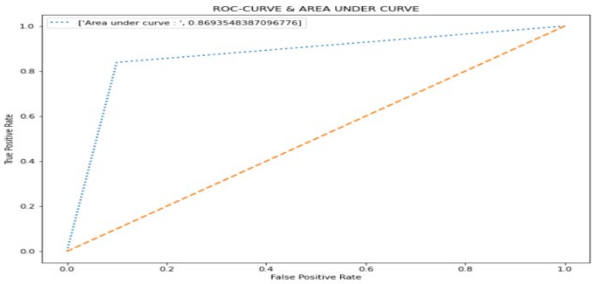


Fig. 8. ROC Curve Report for the proposed model

V. CONCLUSION

In this paper, study is emphasized on building a prediction model to detect heart disease with higher accuracy. To obtain an efficient model, the best feature selection method is chosen which results in obtaining nine optimal features for the dataset considered for the study. The training process for the predictive model requires substantially less time and computing power if it is trained on just these features as opposed to all 26 features of the scaled dataset. Thus, there is a lower possibility that the model will over fit. In addition, it is observed that the male population are suffering more from heart disease as compared to females. From the analysis, it can be concluded that patients suffering from nine features are more prone to heart disease.

VI. REFERENCE

[1] Hoyt, R. (n.d.). Dataset. Kaggle. Retrieved September 5, 2023, from https://www.kaggle.com/datasets/thedevastator/predicting-heart-disease-risk-using-clinical-var?select=Heart_Disease_Prediction.csv

[2] Stojanov, D., Lazarova, E., Veljkova, E., Paolo Rubartelli, & Giacomini, M. (2023). Predicting the outcome of heart failure against chronic-ischemic heart disease in elderly population – Machine learning approach based on logistic regression, case to Villa Scassi hospital Genoa, Italy. Journal of King Saud University - Science, 35(3), 102573–102573. https://doi.org/10.1016/j.jksus.2023.102573

[3] Hassan, Ch. A. ul, Iqbal, J., Irfan, R., Hussain, S., Algarni, A. D., Bukhari, S. S. H., Alturki, N., & Ullah, S. S. (2022). Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers. Sensors, 22(19), 7227. https://doi.org/10.3390/s22197227

[4] Ahmed, & Intisar. (2022). A study of heart disease diagnosis using machine learning and data mining. Electronic Theses, Projects, and Dissertations. 1591.

[5] Saboor, A., Usman, M., Ali, S., Samad, A., Abrar, M. F., & Ullah, N. (2022). A Method for Improving Prediction of Human Heart Disease Using Machine

Learning Algorithms. Mobile Information Systems, 2022, 1–9. https://doi.org/10.1155/2022/1410169

[6] Md. Jubier Ali, Badhan Chandra Das, Suman Kumar Saha, Al Amin Biswas, & Chakraborty, P. (2022). A Comparative Study of Machine Learning Algorithms to Detect Cardiovascular Disease with Feature Selection Method. Springer, 573–586. https://doi.org/10.1007/978-981-19-2347-0_45

[7] Karthick, K., Aruna, S. K., Samikannu, R., Kuppusamy, R., Teekaraman, Y., & Thelkar, A. R. (2022). Implementation of a Heart Disease Risk Prediction Model Using Machine Learning. Computational and Mathematical Methods in Medicine, 2022, 1–14. https://doi.org/10.1155/2022/6517716

[8] Nandal, N., Goel, L., & TANWAR, R. (2022). Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis. F1000Research, 11, 1126. https://doi.org/10.12688/f1000research.123776.1

[9] Pal, M., Parija, S., Panda, G., Dhama, K., & Mohapatra, R. K. (2022). Risk prediction of cardiovascular disease using machine learning classifiers. Open Medicine, 17(1), 1100–1113. https://doi.org/10.1515/med-2022-0508

[10] Nitant, & Priya, Dr. R. (2021). Predicting Heart disease using Machine Learning. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(13), 370–376. https://doi.org/10.17762/turcomat.v12i13.8299

[11] Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. IOP Conference Series: Materials Science and Engineering, 1022, 012072. https://doi.org/10.1088/1757-899x/1022/1/012072

[12] Taqdees, S., Dawood, K., & Akhtar, N. (2021). Heart Disease Prediction.

[13] Indrakumari, R., Poongodi, T., & Jena, S. R. (2020). Heart Disease Prediction using Exploratory Data Analysis. Procedia Computer Science, 173, 130–139. https://doi.org/10.1016/j.procs.2020.06.017

[14] Sen, S. K. (2017). Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms. International Journal of Engineering and Computer Science. https://doi.org/10.18535/ijecs/v6i6.14

[15] Su, X., Xu, Y., Tan, Z., Wang, X., Yang, P., Su, Y., Jiang, Y., Qin, S., & Shang, L. (2020). Prediction for cardiovascular diseases based on laboratory data: An analysis of random forest model. Journal of Clinical Laboratory Analysis, 34(9). https://doi.org/10.1002/jcla.23421

[16] Misra, P., Yadav, A., Yadav, M., & Singh, A. (2020). Improving the Classification Accuracy using Recursive Feature Elimination with Cross-Validation Improving the Classification Accuracy using Recursive Feature Elimination with Cross-Validation. International Journal on Emerging Technologies, 11(3), 659–665.

[17] Khourdifi, Y., & Bahaj, M. (2019). Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization. International Journal of Intelligent Engineering and Systems, 12(1), 242–252. https://doi.org/10.22266/ijies2019.0228.24

[18] Amarbayasgalan, T., Lee, J. Y., Kim, K. R., & Ryu, K. H. (2019). Deep Autoencoder Based Neural Networks for Coronary Heart Disease Risk Prediction. Heterogeneous Data Management, Polystores, and Analytics for Healthcare, 237–248. https://doi.org/10.1007/978-3-030-33752-0_17

[19] Dinesh, K. G., Arumugaraj, K., Santhosh, K. D., & Mareeswari, V. (2018, March 1). Prediction of Cardiovascular Disease Using Machine Learning Algorithms. IEEE Xplore. https://doi.org/10.1109/ICCTCT.2018.8550857