2023-10

A Comparative Study of Machine Learning classifiers to analyze the precision of Myocardial Infraction prediction

Khan, Razib Hayat

Independent University, Bangladesh

A Comparative Study of Machine Learning classifiers to analyze the precision of Myocardial Infraction prediction

Razib Hayat Khan
Department of Computer Science and Engineering
Independent University, Bangladesh
Dhaka, Bangladesh
rkhan@iub.edu.bd

Jonayet Miah
Department of Computer Science
University of South Dakota
South Dakota, USA
Jonayet.miah@coyotes.usd.edu

Shah Ashisul Abed Nipun
Department of Electrical and Computer Engineering (ECE)
North South University
Dhaka, Bangladesh
shah.aa.nipun@northsouth.edu

Majharul Islam
Department of Electrical and Computer Engineering (ECE)
North South University
Dhaka, Bangladesh
majharul.islam02@northsouth.edu

Abstract: In the modern world, heart disease ranks among the main causes of death. Smoking, high blood pressure, and cholesterol are the three key risk factors for getting one heart disease, and 47% of all US people have at least one of these risk factors. Prediction of myocardial illness is a significant problem in the field of medical research methodology. coronary infarction heart disease prediction is a hard issue that hospitals and clinicians must deal with. The precision of the heart disease plays a crucial influence in this prediction. In response to this worry, the authors used a myocardial dataset and a well-known machine-learning method to predict myocardial infarction. The system for detecting cardiac illness utilizing artificial intelligence and machine learning algorithms is the main topic of the study. Here, we demonstrate how machine learning can be used to determine a person's risk of developing heart disease and we are also trying to exactly predict which factors are important to cause Myocardial disease. We are Comparing Six different machine learning models where we got a satisfying result to predict myocardial disease. These are Six different machine learning models LightGBM, XGBoost, Logistic Regression, Bagging, Support Vector Machine, and Decision Tree. The accuracy We got these six models, Logistic Regression, Support Vector Machine, XGBoost, LightGBM, Decision Tree, and Bagging get the results 79.06%, 72.90%, 83.85%, 84.60%, 72.80%, and 82.01% respectively. So, from that, we can take the decision that LightGBM performs best among these six models Our findings suggested a promising future for the treatment of myocardial infarction, but further study and investigation are required before it can be employed commercially, particularly in the healthcare sector.

Keywords: Machine Learning, Myocardial Infarction, Heart Disease, LightGBM.

I. INTRODUCTION

The heart is the most important organ in the human body. It supplies blood to every organ and bone in our body. The brain and several other organs will stop functioning if it isn't functioning properly, and the person would die in a matter of minutes. Cardiovascular disease has garnered a great deal of attention in medical studies because it is one of several ailments that can be fatal. Cardiovascular diagnosis and treatment are challenging tasks that can generate automated predictions about the patient's cardiac condition to increase the efficacy of subsequent treatments. treatment. Cardiovascular disease is normally determined by the likelihood of developing cardiovascular disease and can be influenced by a variety of factors, including smoking, high cholesterol levels, a family history of the condition, being overweight, having high blood pressure, and a sedentary lifestyle. These can be determined through an examination of the patient's signs and symptoms. Numerous heart-related disorders are becoming more common due to changes in lifestyle, anxiety from the workplace, and poor eating habits. Heart attacks, also known as myocardial infarctions (MI), are a prevalent condition worldwide. Many different types of cardiac problems affect a sizable population. The number of patient deaths from heart disorders (Myocardial Infarction-MI) is significantly rising every day. Using patient records, it is a challenging assignment to detect myocardial infarction (heart attack) [16]. In the past, several researchers have employed various techniques to gather and evaluate information to detect heart

failure. These data include cardiovascular disease datasets, biomedical science datasets from UCI, data from patients' electronic health records (EHRs) with cardiovascular disease in different hospitals throughout the world, etc. Although some research has been conducted using machine learning, neither thorough analysis of various models nor attempts at new models like Deep Forest have been made. These gaps can be filled by the work in this paper because nobody offers a sustainable solution for the myocardial disease. In our study, we collect real-time data from hospital to hospital from patients' prescriptions. It was hard, but we are willing to work on an authentic dataset that can give us exact results for potential future analysis. In this study, we propose a comparative study and try to give the best model to predict heart failure. The use of machine learning to uncover hidden patterns in vast amounts of data that can be used for clinical diagnosis has considerable potential.

II. LITERATURE REVIEW

Dwivedi, et al. [1] The author worked on the effectiveness assessment of various machine learning models for the diagnosis of cardiovascular disease. Using six machine learning models, the author of this paper assessed the precision of forecasting heart disease. Additionally, eight distinct classification performance indices were used to assess these algorithms' performance. The receiver operating characteristic curve was also utilized to assess these methods. The most accurate classification method demonstrated to have an 85% accuracy rate, as logistic regression, which had a sensitivity and specificity of 89 and 81%, respectively.

Harshit, et al. [2] The author discussed the development of a by analyzing a patient's medical history, an algorithm can predict if they are likely to have a heart condition. Different machine learning techniques, such as KNN and logistic regression, were used to classify patients with heart issues. This approach was found to be effective in determining the accuracy of predicting myocardial infarction in individuals. The proposed model using KNN, and logistic regression showed good performance and high accuracy in identifying heart disease indicators for a specific individual. ln contrast to the classifiers that were previously used, like Naive Bayes, etc. Given's technique for predicting cardiac disease improves patient treatment while costing less. This initiative has provided us with a wealth of information that can be used to predict who will get heart disease. Their research's major goal was to identify the fewest attributes that can most accurately predict the presence of heart disease. Initially, thirteen factors were considered for predicting heart disease. The researchers used genetic algorithms to identify the most useful characteristics for diagnosing heart conditions, which reduced the number of tests required for a patient. The genetic search technique was able to condense the thirteen attributes into six. The authors then used three different classifiers (Naive Bayes, classification by clustering, and decision trees) to predict patient diagnoses with the same level of accuracy as before, reducing the number of characteristics. The results indicate that, by incorporating feature subset selection, the decision tree data mining technique was found to be the most accurate, with a 99.2% accuracy rate and a mean absolute error of 0.00016. Before the model was built, contradictions and missing values were fixed; in real-time, this is not the case. Additionally, depending on the outcomes, the severity of the illness was unpredictable. It is their goal to continue using fuzzy learning models to gauge the severity of cardiac illness. Masethe, et al. [4] In this study author uses data mining techniques to predict cardiac attacks, including J48, Naive Bayes, REPTREE, SIMPLE CART, and Bayes Net algorithms. The research got a 99% prediction accuracy rate. Having to compare the best way of prediction, the researchers tested the use of multiple data mining algorithms to predict heart attacks. The experiment can aid doctors in recognizing risky circumstances in their practice and provide suitable recommendations. The categorization model will be able to address more complex queries about the conditions that lead to heart attacks. The algorithms' predicted accuracy results imply that the characteristics utilized are trustworthy predictors of the occurrence of heart diseases.

Latha, et al. [5] In this study, the author used an ensemble of classifiers to evaluate the accuracy of predicting heart disease. The Cleveland Heart dataset from the UCI machine learning library was used for training and testing. They used ensemble algorithms such as bagging, boosting, stacking, and majority voting to improve accuracy. Bagging increased by 6.92%, boosting an increase of 5.94%, majority voting an increase of 7.26%, and stacking an increase of 6.93%. The results indicate that majority voting was the most effective in improving accuracy. Feature selection approaches were used to further improve performance. The ensemble algorithms' accuracy was boosted by the feature selection strategies. With the FS2 feature set, majority voting produced the highest accuracy.
Yazdani, et al. [6] The author the researchers used the weighted association rule mining algorithm to predict heart disease based on the scores of key features. They also sought the validation of a set of critical feature scores and rules for diagnosing heart disease from cardiologists. The highest confidence score for predicting heart disease, 98%, was obtained from studies using the UCI open dataset, which is commonly used in research on cardiovascular disease

Hassan, et al. [7] In this paper, the author predicted the presence of heart issues using data from the UCI collection. For predicting heart disease, they evaluated ML methods. The researchers used Gradient Boosted Tree (GBT) and Multilayer Perceptron (MLP) for early prediction of heart disease, which is a departure from previous research that did not use the UCI cardiovascular disease dataset. The results showed that GBT and MLP achieved 95% accuracy in predicting the presence of coronary heart disease among the applied machine learning classifiers. Random Forest (RF) was found to have the highest classification accuracy of 96.28% with a specificity and sensitivity of 0.9628 and 0.9537 respectively.

Groepenhoff, et al. [8] In this research during routine care follow-up, the study used a sample of 24 women (5%) and 75

men (15%) to diagnose coronary artery disease using invasive coronary angiography or CT angiography. Elastic net regression was used to determine the diagnostic value of various chest pain characteristics and risk factors. The overall model accurately predicted outcomes for both sexes (area under the curve (AUC) of 0.76 (95% CI 0.68 to 0.85) in women and 0.83 (95% CI 0.78 to 0.88) in men) by considering factors such as age, provocation by temperature or stress, relief at rest, and functional class. Both the sex-specific models considered similar factors such as age, pressure type, radiation, duration, frequency, progression, provocation, and relief at rest. However, the male model also considered functional class and diabetes, while the female model also considered dyspnea, body mass index, hypertension, and smoking. The results showed that the sex-specific algorithms performed better in females than in males when compared to the overall model (AUC: 0.89, 95% CI: 0.81 to 0.96; 0.84, 95% CI: 0.73 to 0.90)

III. METHODOLOGY

A. Data Collection

All the data have been collected manually from different clinics and hospitals in Dhaka, Bangladesh. The authors gave the most concentration to the Heart failure-related patients' recent conditions during the collection of the data from them. Table I is the list of the clinics and hospitals from where the team collected the data. 13 traits and 600 total instances, with are shown in Table II have been collected in which the dataset contains information about a specific patient and includes a class attribute that is separated into two categories: Distinctive and Non-Distinctive. Where 12 attributes are distinctive and only one attribute is non-distinctive. The dataset includes the treatment records of 600 heart failure patients, with each patient's profile having 13 different clinical characteristics. These records were collected over the course of the patients' treatment. The authors showed the correlation matrix in figure 2.

1.	Holy Family Red Crescent medical college Hospital
2.	National Heart Foundation Hospital
3.	Ibrahim Cardiac Hospital & Research Institute
4.	Uttara Adhunik Medical College and Hospital
5.	Farida Clinic

Table I: The Hospitals which helped to collect data

B. Data Preprocessing & Filter

We employed two unsupervised filters in the preprocessing stage in the well-known machine-learning Waikato Knowledge Analysis Environment (WEKA 3.8.3). We initially eliminated the Absent Items from the dataset, after which we updated those. The mean, median, and modes were used in this filtering method to substitute the values that were missing from qualitative and quantitative properties. Second, we employed Randomly select filtering, which fills in the gaps in the data without significantly reducing speed. Another one is the median (), which determines the dataset's middle value.

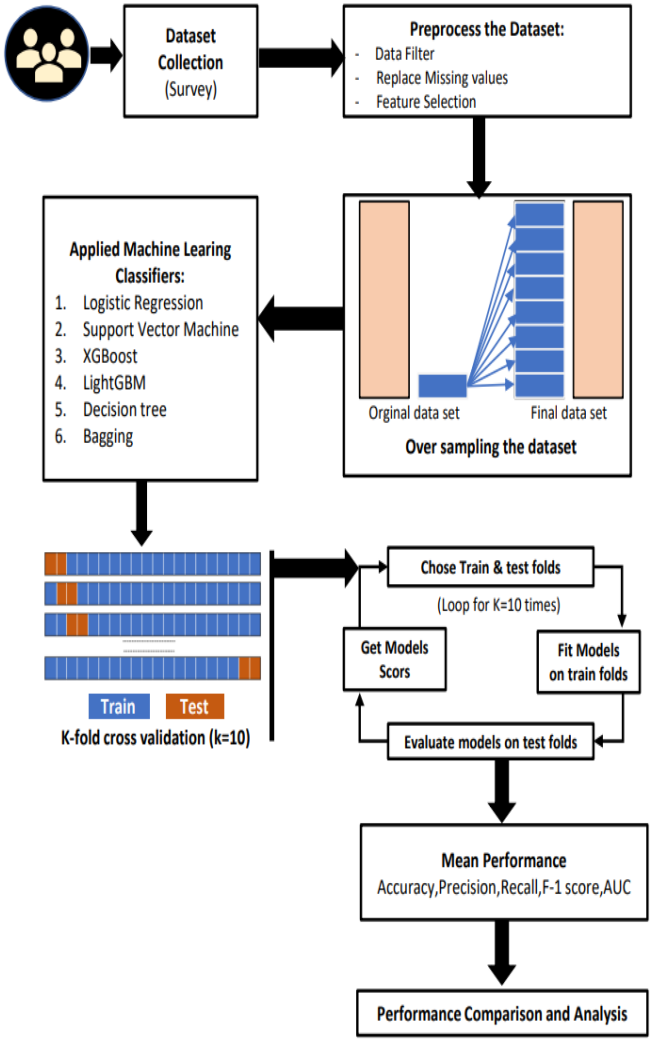

Fig. 1: The overview of our study

C. Feature Selection and Validation Technique

Rapid Miner: Rapid Miner is employed in academia, training, and research. This application we used for data processing, visualization results, model validity, and optimization. One of the most predictive analytical techniques utilized by Gartner, it quickly recognized the knife as the leader of the most sophisticated magic quadrilateral systems theory. Choosing the proper validation technique for specific Datasets is essential. The hold-out validation method is the most effective strategy. we are training 80% of the dataset and testing 30% of it, we applied a holdout validation technique to get good results. Furthermore, we measured the accuracy, sensitivity, specificity, and F1- Score by the implied confusion Metrix. A thorough examination is provided in the visualization and display of the performance indicators bar graphs.

Attribute names	Attribute information
1. Age	the patient's age years)
2. Sex	male or female (binary)
3. Diabetes	diabetes if the patient (Boolean)
4. Smoking	Whether or not the individual smoked (Boolean)
5. Anemia	reduction in hemoglobin or red blood cells (Boolean)
6. High Blood Pressure	If the patient has high blood pressure (Boolean)
7. Creatinine phosphokinase (CPK)	CPK enzymatic blood concentration (mcg/L)
8. Time	Observation period (Days)
9. Serum creatinine	Blood serum creatinine concentration (mg/dL)
10. Serum sodium	serum sodium concentration in the blood (mEq/L)
11. Ejection fraction	percentage of blood that leaves the heart after each contraction (%)
12. Platelets	A blood platelet's presence (kilo platelets/mL)
13. Death event [Target]	During the time of the follow-up, if the patient passed away (Boolean)

Table II: Features List (Dataset attribute names)

Machine Learning Algorithms

After completing the data processing, training, and categorization, various machine learning algorithms were employed, such as regression analysis, logistic regression, support vector machines (SVMs), random forest, k-nearest neighbor, the Decision Tree Algorithm, and XGBoost. The best-performing algorithm was chosen and the results of the different algorithms on the dataset are presented as follows.

Logistic Regression

A supervised learning approach called logistic regression is employed to forecast a dependent categorized predicted value. To categorize a massive dataset includes regression models is very much useful. Where this algorithm predicts the probability of certain classes based on some dependent variables [14].

Mathematically, logistic regression is represented by the equation:

$$y = e\ (b0 + b1x) / 1 + e(b0 + b1x)$$

The input value is represented by 'x', the expected outcome is represented by 'y', the bias or intercept term is represented by 'b0', and the input coefficient is represented by 'b1' in the given equation.

This algorithm uses the Sigmoid function for apply the model to discontinuous possibilities. This function transforms quantitative outputs into a likelihood statement between 0 and 1. To improve the accuracy in logistic regression the following steps need to be followed:

i. Need to import the required libraries.
ii. Input and visualize the dataset.
iii. Handling Null/Missing Values and cleaning unnecessary data from the data set.
iv. Deal or Analysis with the outliers.
v. Define dependent and independent variables and then divided the data into a training and testing set
vi. Use Ensemble and Boosting Algorithms.
vii. Hyperparameter Tuning.

Additionally, we determined the average scores for each predictive modeling classification, including accuracy, precision, recall, the F1-Score, and the area under the curve (AUC). We presented a summary of the study in Table III. Table III compares the average performance results for various types of selected machine learning classifiers, such as logistic regression, support vector machines, XGBoost, LightGBM, decision trees, and bagging. The model's performances were evaluated using accuracy, precision, recall, F1Score, and AUC as metrics.

IV. RESULT AND DISCUSSION

The study found that maintaining a normal platelet count improves the survival rate, however, the correlation between the two is low. Also, maintaining a normal sodium level lowers the risk of death after a heart failure, while high blood pressure increases the risk of death after cardiac failure. The authors noted that having a higher ejection fraction appears to decrease the risk of death after cardiac failure, but due to the small sample size, it's not possible to infer any conclusions from the extreme values. The correlation between the variables is low, except for the relationship between sex and smoking.

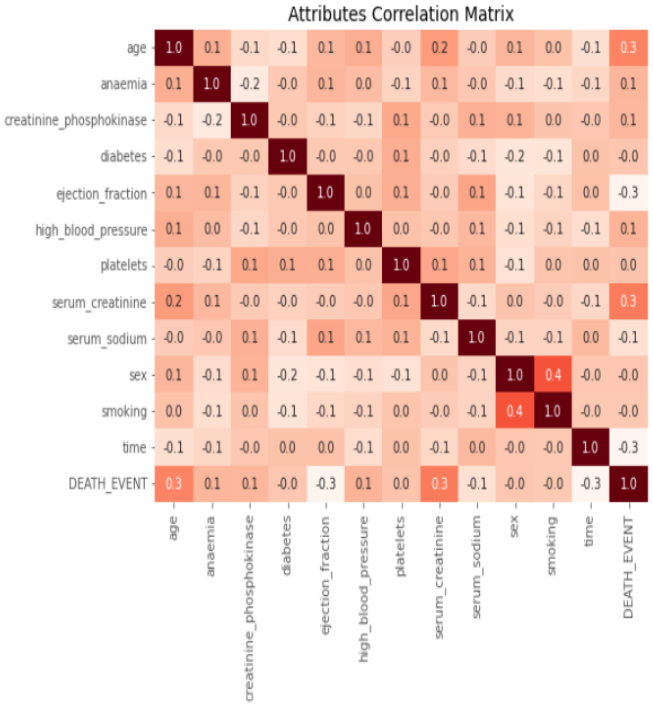

Fig 2: Correlation matrix between dataset attributes

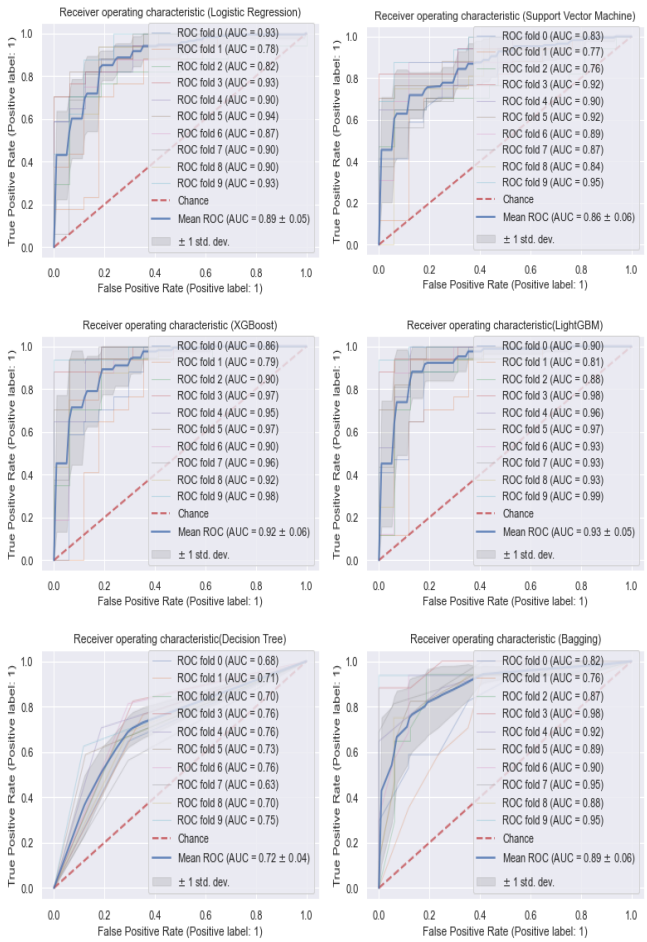

Fig 3**:** Area under curve output of a different kind of machine learning Algorithm

An Area under the curve of 0.5 generally means that while there is no unequal treatment (i.e., the ability to determine a patient's likelihood of developing cardiovascular disease based on the test), while an Area under curve of 0.7 to 0.8 means that achievement is appropriate, an AUC of 0.8 to 0.9 means that results are excellent, and an AUC of much more than 0.9 means that the system has performed exceptionally well [13]. The performance of the specified machine learning models using 10-fold cross-validation is illustrated in Figure 3, which includes AUC plots and the average results.

Machine Learning Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**	**AUC (%)**
1. Logistic Regression	79.06	81.00	79.60	79.90	89.05
2. Support Vector Machine	72.90	71.04	80.07	74.80	86.06
3.XGBoost	83.85	86.00	83.00	82.90	92.06
4.LightGBM	84.60	87.30	81.90	83.00	92.95
5. Decision Tree	72.80	72.90	73.87	72.85	72.04
6. Bagging	82.01	87.96	74.45	71.00	89.06

Table III Outcomes of different machine learning classifiers

Table III presents the results of six different machine learning classifiers: LightGBM, XGBoost, Logistic Regression, Bagging, Support Vector Machine, and Decision Tree. These classifiers achieved AUC scores of 92.95%, 92.06%, 89.05%, 89.06%, 86.06%, and 72.04%, respectively. It indicates that LightGBM and XGBoost performed the best among all classifiers, with LightGBM displaying exceptional performance in terms of accuracy and AUC. Chart 1 also illustrates the comparison of results obtained from different machine learning models.

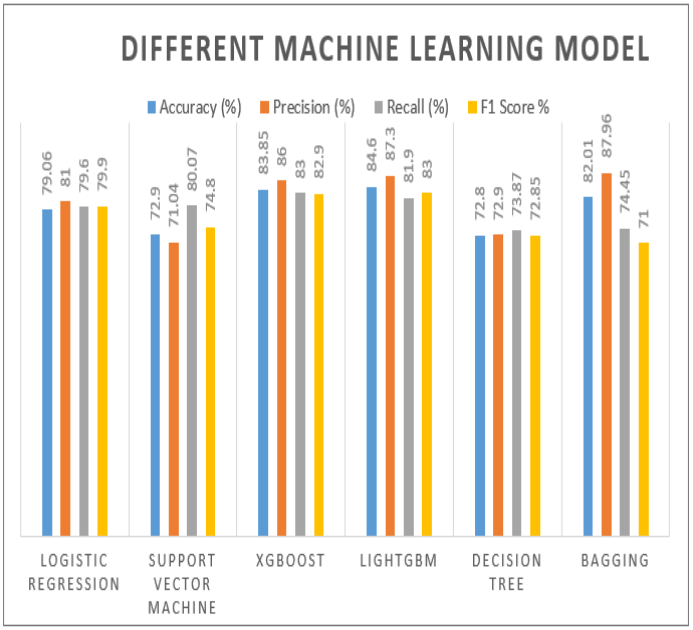

Chart 1. Analysis of the different machine learning models

V. CONCLUSION AND FUTURE WORK

Day by day the patient of myocardial disease is increasing it is good news to detect this disease through machine learning which is very compact in the diagnosis of this disease. Our proposed model is effective and sometimes efficient for healthcare to detect this disease at an early age. At least the healthcare system is getting some of the technological advantages to save human life. Another aim is to build a machine learning model to give a cost-effective diagnosis system for rural areas people it is not necessary to test people medically because the system can detect whether someone has a myocardial disease or not because we train the system by analysis of thousands of same previous cases. We observe many of the machine learning projects outperform in predicting myocardial disease at an early age, but our proposed model gives a profound result such as LightGBM getting 84.60% accuracy, 87.30% precision, and 83.00% in the recall. For this to used widely, more research must be conducted. Forecasting Cardiovascular health and pattern analysis may become a reality and greatly aid the development of the medical Sector if big data problems are resolved and advanced innovation, such as blockchain, is used. We might potentially focus on Myocardial disease detection in the future using different types of vocal features-based datasets by deep learning models. To counteract this dangerous situation and lower the likelihood of cardiovascular disease mortality, a technology that can predict cardiovascular disease recovery is required with the rising number of patients experiencing cardiovascular failures.

REFERENCE

[1] Dwivedi, A.K. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Compute & Application 29, 685–693 (2018). https://doi.org/10.1007/s00521-016-2604-1

[2] Harshit Jindal[1], Sarthak Agrawal[1], Rishabh Khera[1], Rachna Jain[2], and Preethi Nagrath[2] Published under license by IOPPublishingLtd IOP Conference Series: Materials Science and Engineering, Volume 1022, 1st International Conference on Computational Research and Data Analytics (ICCRDA 2020) 24th October 2020, Rajpura, India Harshit Jindal, *et al* 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* 1022 012072

[3] Anbarasi, M., Anupriya, E., & Iyengar, N. C. S. N. (2010). Enhanced prediction of heart disease with feature subset selection using genetic algorithm. International Journal of Engineering Science and Technology, 2(10), 5370-5376.

[4] Masethe, H. D., & Masethe, M. A. (2014, October). Prediction of heart disease using classification algorithms. In Proceedings of the World Congress on Engineering and Computer Science (Vol. 2, No. 1, pp. 25-29).

[5] Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked, 16, 100203.

[6] Yazdani, A., Varathan, K.D., Chiam, Y.K. et al. A novel approach for heart disease prediction using strength scores with significant predictors. BMC Med Inform Decis Mak 21, 194 (2021).

[7] Hassan CAU, Iqbal J, Irfan R, Hussain S, Algarni AD, Bukhari SSH, Alturki N, Ullah SS. Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers. Sensors (Basel). 2022 Sep 23;22(19):7227. DOI: 10.3390/s22197227. PMID: 36236325; PMCID: PMC9573101.

[8] Groepenhoff F, Eikendal ALM, Onland-Moret NC, Bots SH, Menken R, Tulevski II, Somsen AG, Hofstra L, den Ruijter HM. Coronary artery disease prediction in women and men using chest pain characteristics and risk factors: an observational study in outpatient clinics. BMJ Open. 2020 Apr 26;10(4): e035928. Doi: 10.1136/bmjopen-2019-035928. PMID: 32341045; PMCID: PMC7204862.

[9] M.Rouse, (2018) “The essential guide to managing HR technology trends”. [Accessed 15 July 2019] Link:www.searchenterpriseai.techtarget.com.

[10] D. Sayad, “Hierarchical Clustering”, clustering hierarchical. Link: www.saedsayad.com. [Accessed: 17-july-2019]

[11] P. Shakeel, Baskar et. al, (2018). “Cloud-based framework for the diagnosis of diabetes mellitus using K-means clustering”, vol.6, no.1.

[12] Mandrekar, J. (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. Journal Of Thoracic Oncology,5(9), 1315- 1316.doi: 10.1097/jto.0b013e3181ec173d.

[13] Maalouf, M. (2011). Logistic regression in data analysis: an overview. *International Journal of Data Analysis Techniques and Strategies*, *3*(3), 281-299.

[14] Islam, Md. , nipun , Shah Ashisul Abed., Islam, Majharul., Rakib Raht, Md Abdur., Miah,Jonayet., Kayyum,Salsavil., Shadaab, Anwar ., Faisal, Fiaz al et al. "An Empirical Study to Predict Myocardial Infarction Using K-Means and Hierarchical Clustering." *International Conference on Machine Learning, Image Processing, Network Security and Data Sciences*. Springer, Singapore, 2020.

[15] S. Kayyum et al., "Data Analysis on Myocardial Infarction with the help of Machine Learning Algorithms considering Distinctive or Non-Distinctive Features," 2020 International Conference on Computer Communication and Informatics (ICCCI),2020, pp.1-7, doi:10.1109/ICCCI48352.2020.9104104.

[16] Singh, P., Singh, S., & Pandi-Jain, G. S. (2018). Effective heart disease prediction system using data mining techniques. *International journal of nanomedicine*, *13*(T-NANO2014Abstracts),121–124. https://doi.org/10.2147/IJN.S124998.

[17] J. Miah, M. Mamun, M.M. Rahman, M. I. Mahmud, A. M. Islam, S. Ahmad, “MHfit: Mobile Health Data for Predicting Athletics Fitness using Machine Learning Models” 2022 2nd International seminar on machine learning, Optimization, and Data Science (ISMODE), 2022, (Preprint)