A Comprehensive Review of Heart Disease Prediction Using Machine Learning Approach

[1]Shubha Soni, [2]Dr. Neeraj Sharma, [3]Dr. Harsh Pratap Singh
[1]Research Scholar, Computer Science and Engineering Sri Satya Sai University of Technology and Medical Sciences, Sehore
[2]Associate Professor, Computer Science and Engineering, Sri Satya Sai University of Technology and Medical Sciences, Sehore
[3]Assistant Professor, Computer Science and Engineering, Sri Satya Sai University of Technology and Medical Sciences, Sehore

Abstract

Prediction of heart disease becomes very important for the doctors or practitioners which helps in accurate decision making for the development of heart patients. Heart is one of the most significant part of the body. It helps to clean and circulate blood to whole parts of the body. Machine learning technique is widely used for the understanding the symptoms of heart patients. Machine learning is an analytical tool used when a task is large and difficult to program, such as transforming medical record into knowledge, pandemic predictions, and genomic data analysis. In this paper, we present the review of literature work done for prediction of heart disease and also discusses some machine learning techniques of heart prediction such as decision tree, naïve bayes algorithm, logistic regression, random forest etc.

Keywords: Machine Learning, Heart Disease, Prediction, Decision tree, Heart Patients, Regression

Introduction

According to the Centers for Disease Control and Prevention (CDC), heart disease is the leading cause of death in the United States. Around 1 in 4 deaths Trusted Source in the U.S. occur due to heart disease, and the condition affects all genders as well as all racial and ethnic groups. Heart disease does not discriminate. It is the leading cause of death for several populations, including white people, Hispanics, and Black people. Almost half of Americans are at risk for heart disease, and the numbers are rising. While heart disease can be deadly, it's also preventable in most people. By adopting healthy lifestyle habits early, you can potentially live longer with a healthier heart [1]. The most important behavioural risk factors of heart disease and stroke are unhealthy diet, physical inactivity, tobacco use and harmful use of alcohol. The effects of behavioural risk factors may show up in individuals as raised blood pressure, raised blood glucose, raised blood lipids, and overweight and obesity. These "intermediate risks factors" can be measured in primary care facilities and indicate an increased risk of heart attack, stroke, heart failure and other complications. Cessation of tobacco use, reduction of salt in the diet, eating more fruit and vegetables, regular physical activity and avoiding harmful use of alcohol have been shown to reduce the risk of cardiovascular disease. Health policies that create conducive environments for making healthy choices affordable and available are essential for motivating people to adopt and sustain healthy behaviours. There are also a number of underlying determinants of CVDs [2]. These are a reflection of the major forces driving social, economic and cultural change – globalization, urbanization and population ageing. Other determinants of CVDs include poverty, stress and hereditary factors. In addition, drug treatment of hypertension, diabetes and high blood lipids are necessary to reduce cardiovascular risk and prevent heart attacks and strokes among people with these conditions. Recent studies have used machine learning techniques to diagnose different cardiac problems and make a prediction.

Types of Heart Disease

There are many types of heart disease, and each one has its own symptoms and treatment. For some, lifestyle changes and medicine can make a huge difference in improving your health. For others, you may need surgery to make your ticker work well again [3].

Heart Arrhythmias

When you have an arrhythmia, your heart has an irregular beating pattern. Serious arrhythmias often develop from other heart problems but may also happen on their own.

Heart Failure
With heart failure, your heart doesn't pump blood as well as it should to meet your body's needs. It is usually caused by coronary artery disease, but it can also happen because you have thyroid disease, high blood pressure, heart muscle disease (cardiomyopathy), or certain other conditions.

Heart Valve Disease
Your heart has four valves that open and close to direct blood flow between your heart's four chambers, the lungs, and blood vessels. An abnormality could make it hard for a valve to open and close the right way. When that happens, your blood flow could be blocked or blood can leak. Your valve may not open and close right.
Diseases of the heart valves include:

- **Endocarditis.** This is an infection that's usually caused by bacteria, which may enter the blood and take root in your heart during illness, after surgery, or after using intravenous drugs. It often happens if you already have valve problems. Antibiotics can usually cure it, but the disease is life threatening without treatment. If your heart valves are seriously damaged as a result of endocarditis, you may need valve replacement surgery.
- **Rheumatic heart disease.** This condition develops when your heart muscle and valves are damaged by rheumatic fever, which is linked to strep throat and scarlet fever. Rheumatic heart disease was more common earlier in the 20th century. But doctors are now able to prevent it by using antibiotics to treat the diseases that lead to it. If you do get it, the symptoms usually show up many years after the infection [3].

Pericardial Disease
Any disease of the pericardium, the sac that surrounds your heart, is called a pericardial disease. One of the more common diseases is pericarditis or inflammation of the pericardium.

Cardiomyopathy (Heart Muscle Disease)
Cardiomyopathy is a disease of your heart muscle, or myocardium. It gets stretched, thickened, or stiff. Your heart may get too weak to pump well.
There are many possible causes of the disease, including genetic heart conditions, reactions to certain drugs or toxins (such as alcohol), and infections from a virus. Sometimes, chemotherapy causes cardiomyopathy. Many times, doctors can't find the exact cause.

Congenital Heart Disease
Congenital heart disease happens when something goes wrong while the heart is forming in a baby that's still in the womb. The heart abnormality sometimes leads to problems right after birth, but other times there are not any symptoms until you become an adult. Septal abnormalities are among the most common congenital heart problems. These are holes in the wall that separates the left and right sides of your heart. You can get a procedure to patch the hole. Another type of abnormality is called pulmonary stenosis. A narrow valve causes a decrease in the flow of blood to your lungs. A procedure or surgery can open or replace the valve. In some babies, a small blood vessel known as the ductus arteriosus doesn't close up at birth as it should. When this happens, some blood leaks back into the pulmonary artery, which puts strain on your heart. Doctors can treat this with surgery or a procedure or sometimes with medication.

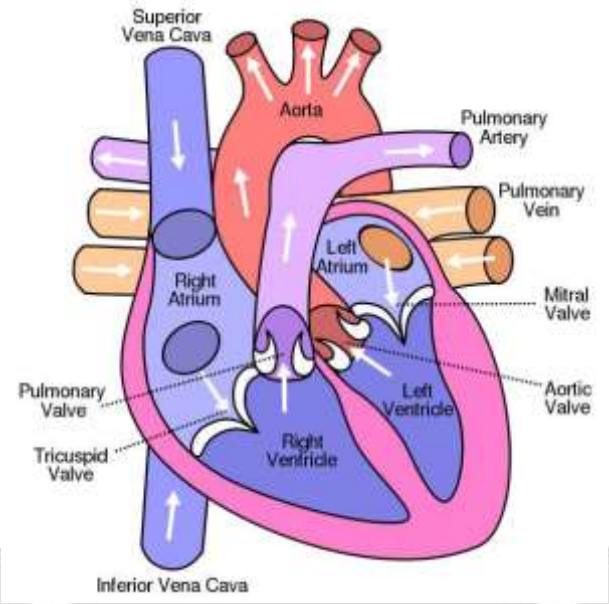


Fig. 1: Structure of Human Heart [16]

Symptoms of Heart Disease

Often, there are no symptoms of the underlying disease of the blood vessels. A heart attack or stroke may be the first sign of underlying disease [4]. Symptoms of a heart attack include:

- Pain or discomfort in the centre of the chest; and/or
- Pain or discomfort in the arms, the left shoulder, elbows, jaw, or back.

In addition the person may experience difficulty in breathing or shortness of breath; nausea or vomiting; light-headedness or faintness; a cold sweat; and turning pale. Women are more likely than men to have shortness of breath, nausea, vomiting, and back or jaw pain.

The most common symptom of a stroke is sudden weakness of the face, arm, or leg, most often on one side of the body. Other symptoms include sudden onset of:

- Numbness of the face, arm, or leg, especially on one side of the body;
- Confusion, difficulty speaking or understanding speech;
- Difficulty seeing with one or both eyes;
- Difficulty walking, dizziness and/or loss of balance or coordination;
- Severe headache with no known cause; and/or
- Fainting or unconsciousness.

Related Work

Anna Karen Garate-Escamila, Amir Hajjam El Hassani, Emmanuel Andres (2020) proposal of a dimensionality reduction method and finding features of heart disease by applying a feature selection technique. The information used for this analysis was obtained from the UCI Machine Learning Repository called Heart Disease. The dataset contains 74 features and a label that we validated by six ML classifiers. Chi-square and principal component analysis (CHI-PCA) with random forests (RF) had the highest accuracy, with 98.7% for Cleveland, 99.0% for Hungarian, and 99.4% for Cleveland-Hungarian (CH) datasets. From the analysis, ChiSqSelector derived features of anatomical and physiological relevance, such as cholesterol, highest heart rate, chest pain, features related to ST depression, and heart vessels. The experimental results proved that the combination of chi-square with PCA obtains greater performance in most classifiers. The usage of PCA directly from the raw data computed lower results and would require greater dimensionality to improve the results [5]. ***G. Jignesh Chowdary, Suganya. G, Premalatha. M (2020)*** introduced a novel ensemble methodology which uses the voting of Logistic Regression (LR), Random Forest (RF), Artificial Neural Network activated with ReLU function(NNR), K-Nearest Neighbors (KNN) and Gaussian Naive Bayes(GNB) to predict the possibility

of heart disease. The model is developed using Python-based Jupyter Notebook and Flask and is trained using the standard dataset from Kaggle. The model is tested and evaluated based on accuracy, precision, specificity, sensitivity, error. Testing witnessed an accuracy of 89% and a precision of 91.6%, along with a sensitivity of 86% and specificity of 91%. The results upon comparison with the individual models witness the better accuracy of using ensemble modeling and hence a better prediction leading to life saving [6]. ***Amin Ul Haq , Jian Ping Li , Muhammad Hammad Memon , Shah Nazir , and Ruinan Sun (2018)*** developed a machine-learning-based diagnosis system for heart disease prediction by using heart disease dataset. We used seven popular machine learning algorithms, three feature selection algorithms, the cross-validation method, and seven classifiers performance evaluation metrics such as classification accuracy, specificity, sensitivity, Matthews' correlation coefficient, and execution time. 0e proposed system can easily identify and classify people with heart disease from healthy people. Additionally, receiver optimistic curves and area under the curves for each classifier was computed. We have discussed all of the classifiers, feature selection algorithms, pre-processing methods, validation method, and classifiers performance evaluation metrics used in this paper. 0e performance of the proposed system has been validated on full features and on a reduced set of features. 0e features reduction has an impact on classifiers performance in terms of accuracy and execution time of classifiers. 0e proposed machine-learning-based decision support system will assist the doctors to diagnosis heart patients efficiently [7]. ***Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava (2019)*** proposed a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with different combinations of features and several known classification techniques. We produce an enhanced performance level with an accuracy level of 88:7% through the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM) [8]. ***Akram Ahmed Mohammed, Rajkumar Basa, Anirudh Kumar Kuchuru, Shiva Prasad Nandigama, Maneeshwar Gangolla (2020)*** proposed an integrated method of Random Forest Machine Learning Algorithm is with the Flask Web framework for predicting of Heart Disease is proposed. The ensemble learning methods are used for predicting heart disease. The proposed methodology involved integration of the Flask Web framework with the Random Forest machine learning technique to estimate the heart disease stages. Artery Blockage indicates the presence of heart disease. The higher the blockage, higher is the stage of heart disease. Stage 1 and Stage 2 indicate the presence of heart disease whereas Stage 3 and Stage 4 are called chronic heart disease and the risk of a heart attack at any day in such patients is very high. The Data required for the prediction contains parameters such as Age, Sex, Blood Pressure, Sugar levels which are collected from the Kaggle website. Experimental results say that predictions by using the proposed approach are consistently better than those obtained using the other methods [9]. ***Malavika G, Rajathi N, Vanitha V and Parameswari P (2020)*** extract the hidden information, which is useful to predict disease at the earlier. In medical field, predicting heart disease is treated as one of the intricate tasks. Therefore, there is a necessity to develop a decision support system to forecast the cardio vascular disease in a patient. Machine learning plays a vital part in disease prediction. In this paper, various machine learning methods were used to predict the heart disease and their performances were compared. The results obtained show the superiority of the Random forest algorithm [10]. ***Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain and Preeti Nagrath (2020)*** focused on which patient is more likely to have a heart disease based on various medical attributes. We prepared a heart disease prediction system to predict whether the patient is likely to be diagnosed with a heart disease or not using the medical history of the patient. We used different algorithms of machine learning such as logistic regression and KNN to predict and classify the patient with heart disease. A quite Helpful approach was used to regulate how the model can be used to improve the accuracy of prediction of Heart Attack in any individual. The strength of the proposed model was quiet satisfying and was able to predict evidence of having a heart disease in a particular individual by using KNN and Logistic Regression which showed a good accuracy in comparison to the previously used classifier such as Naive Bayes etc. So a quiet significant amount of pressure has been lift off by using the given model in finding the probability of the classifier to correctly and accurately identify the heart disease. The Given heart disease prediction system enhances medical care and reduces the cost. This project gives us significant knowledge that can help us predict the patients with heart disease It is implemented on the .pynb format [11]. ***Riddhi Kasabe and Prof. Dr. Geetika Narang (2020)*** evaluate different classification techniques in heart diagnosis. First, the heart numeric dataset is extracted and preprocess them. After that using extract the features that is condition to be find to be classified by machine learning. Compared to existing; machine learning provides better performance. After classification, performance criteria including accuracy, precision, F-measure is to be calculated. Machine learning provides better performance. The comparison measure expose that Random Forest is the best classifier for the diagnosis of heart disease on the existing dataset [12]. ***Dinesh Kumar G, Santhosh Kumar D, Arumugaraj K, Mareeswari V(2018)*** proposed research, data pre-processing uses techniques like the removal of noisy data, removal of missing data, filling default values if applicable and classification of attributes for prediction and decision making at different levels. The performance of the diagnosis model is obtained by using methods like classification, accuracy, sensitivity and specificity analysis. This project proposes a prediction model to predict

whether a people have a heart disease or not and to provide an awareness or diagnosis on that. This is done by comparing the accuracies of applying rules to the individual results of Support Vector Machine, Gradient Boosting, Random forest, Naive Bayes classifier and logistic regression on the dataset taken in a region to present an accurate model of predicting cardiovascular disease [13]. ***Aditi Gavhane, Gouthami Kokkula, Isha Pandya, Kailas Devadkar(2018)*** put a system in place to be able to detect the symptoms of a heart stroke at an early stage and thus prevent it. It is impractical for a common man to frequently undergo costly tests like the ECG and thus there needs to be a system in place which is handy and at the same time reliable, in predicting the chances of a heart disease. Thus we propose to develop an application which can predict the vulnerability of a heart disease given basic symptoms like age, sex, pulse rate etc. The machine learning algorithm neural networks has proven to be the most accurate [14]. ***Indu Yekkala, Sunanda Dixit (2018)*** insights not only diagnose the diseases but also predict and can prevent disease. One such use of these techniques is cardiovascular diseases. Heart disease or coronary artery disease (CAD) is one of the major causes of death all over the world. Comprehensive research using single data mining techniques have not resulted in an acceptable accuracy. Further research is being carried out on the effectiveness of hybridizing more than one technique for increasing accuracy in the diagnosis of heart disease. In this article, the authors worked on heart stalog dataset collected from the UCI repository, used the Random Forest algorithm and Feature Selection using rough sets to accurately predict the occurrence of heart disease [15].

Heart Disease Prediction Techniques

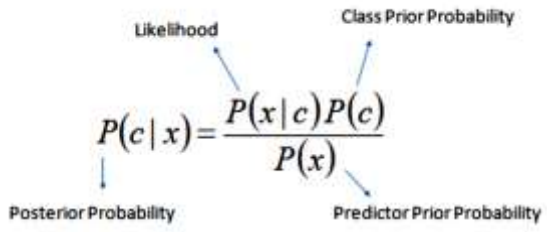
For prediction of heart disease various machine learning techniques has been developed and among from them some prediction techniques are explained below:

Decision Tree

The methodology used in the Decision tree is a commonly used data mining method for establishing classification and prediction systems based on multiple explanatory parameters for developing prediction models for a target instance. This path classifies a population into branch-like segments in a tree that construct an inverted tree with a root node, internal nodes, and leaf nodes. A decision tree is a non-parametric algorithm which can efficiently deal with huge, complicated data sets without involving multiple parametric structures. If the sample size is large enough, study data can be divided into training and validation data sets. Using the training data set to build a decision tree model and a validation data set decide on the appropriate tree size to achieve the optimal final model [17].

Naïve Bayes Classifier

Naive Bayes classifiers is a probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions between the features. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful in the field of medical science for diagnosing heart patients. Despite its simplicity, the Naïve Bayesian classifier[18] often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods. Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x|c)$. Naïve Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.



$$P(c \mid \mathbf{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- $P(c|x)$ is the posterior probability of *class* (*target*) given *predictor* (*attribute*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*

Where C and X are two events (e.g. the probability that the train will arrive on time given that the weather is rainy). Such Naïve Bayes classifiers use the probability theory to find the most likely classification of an unseen (unclassified) instance. The algorithm performs positively with categorical data but poorly if we have numerical data in the training set.

Logistic Regression

Logistic regression is a predictive model used to evaluate the relationship between the dependent variable (target) which is categorical data with nominal or ordinal scale and the independent variable (predictor) which is categorical data with interval or ratio scale. This algorithm can also be used for time series modelling to find the relationship between the variables involved. Logistic regression is an algorithm used to predict the probability of categorical dependent variables. In logistic regression, the dependent variable is shown as a binary variable that is valued at 1 (yes) Or 0 (no). The logistic regression model predicts as a function of X. The assumptions used in Logistic regression are as follows: binary logistic regression requires binary dependent variables, for binary regression, the factor 1 level of the dependent variable must represent the desired result, independent variables must be independent of each other. In this case, the model must have little or no multicollinearity and be linearly related to log opportunities [19]. Logistic regression used appropriate regression analysis to be performed when the dependent variable is dichotomous (binary). Logistic regression acts as a predictive analytical model. Logistic regression is applied to describe data and explain the relationship between one dependent binary variable with one or more independent variables at the nominal, ordinal, interval or ratio level. Logistic regression has several advantages and disadvantages [20]. The benefits of logistic regression include the following. First, logistic regression can show a significant relationship between the dependent variable and the independent variable. Second, logistic regression analysis can also be used to compare the effect of variables measured at different scales includings the effect of price changes and the number of promotional activities. This benefit helps market researchers or data analysts to eliminate and evaluate the best set of variables that will be used to build predictive models. Third, the logistic regression model is not only a classification model, but also provides information related to probability. To achieve a better result using Logistic Regression, first all independent variable must contain their valid value. Secondly, logistic regression works well for predicting categorical results and multinomial results. Third, there is no multicollinearity between variables in the dataset [21].

Random Forest

RF (see Figure 1 for an illustration) is a collection or ensemble of Classification and Regression Trees (CART) [22] trained on datasets of the same size as training set, called bootstraps, created from a random resampling on the training set itself. Once a tree is constructed, a set of bootstraps, which do not include any particular record from the original dataset [out-of-bag (OOB) samples], is used as test set. The error rate of the classification of all the test sets is the OOB estimate of the generalization error. [23] showed by empirical evidence that, for the bagged classifiers, the OOB error is accurate as using a test set of the same size as the training set. Thus, using the OOB estimate removes the need for a separate test set. To classify new input data, each individual CART tree (colored branches in Figure 1) votes for one class and the forest predicts the class that obtains the plurality of votes.

Fig. 2: Illustration of a random forest construct superimposed on a coronal slice of the MNI 152 (Montreal Neurological Institute) standard template. Each binary node (white circles) is partitioned based on a single feature, and each branch ends in a terminal node, where the prediction of the class is provided. The different colors of the branches represent each of the trees in the forest. The final prediction for a test set is obtained by combining with a majority vote the predictions of all single trees.

RF follows specific rules for tree growing, tree combination, self-testing and post-processing, it is robust to overfitting and it is considered more stable in the presence of outliers and in very high dimensional parameter spaces than other machine learning algorithms ([24, 25]. The concept of variable importance is an implicit feature selection performed by RF with a random subspace methodology, and it is assessed by the Gini impurity criterion index [26]. The Gini index is a measure of prediction power of variables in regression or classification, based on the principle of impurity reduction [27]; it is non-parametric and therefore does not rely on data belonging to a particular type of distribution. For a binary split (white circles in Figure 1), the Gini index of a node n is calculated as follows:

$$Gini(n) = 1 - \sum_{j=1}^{2} (P_j)^2$$

where pj is the relative frequency of class j in the node n. For splitting a binary node in the best way, the improvement in the Gini index should be maximized. In other words, a low Gini (i.e., a greater decrease in Gini) means that a particular predictor feature plays a greater role in partitioning the data into the two classes. Thus, the Gini index can be used to rank the importance of features for a classification problem.

Conclusion

Machine learning concept is widely used in medical filed for the early prediction of disease and diagnosis of heart patient on the basis of available dataset. The researchers have developed various machine learning techniques for the diagnosis of heart disease such as decision tress, naïve bayes, logistic regression etc. In this paper, we present the comprehensive review on the prediction and diagnosis of heart disease. The use of machine learning techniques provides the accurate results in prediction and which helps the practitioner to give better treatment to the heart patients. In this also discuss various machine learning techniques and after reviewing it is found that not a single technique gives better results. In future work, we will use the hybrid approach of machine learning techniques and their analysis will also be perform on some other measuring parameters such as precision, recall and F1 measure.

Reference

[1] Heart disease webpage on Mayo Clinic [Online]. Available: https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118, 2019.

[2] Chaitrali S.Dangare, and Sulabha S.Apte, “Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques”, International Journal of Computer Applications (0975 – 888), Vol. 47, No.10, June.2021.

[3] Guizhou Hu, Martin M. Root, “Building Prediction Models for Coronary Heart Disease by Synthesizing Multiple Longitudinal Research Findings”, European Science of Cardiology, 10 May 2005.

[4] T.Mythili, Dev Mukherji, Nikita Padaila and Abhiram Naidu, “A Heart Disease Prediction Model using SVM- Decision Trees- Logistic Regression (SDL)”, International Journal of Computer Applications, vol. 68, 16 April 2013.

[5] Anna Karen Garate-Escamila, Amir Hajjam El Hassani, Emmanuel Andres, “Classification models for heart disease prediction using feature selection and PCA”, Informatics in Medicine Unlocked 19 (2020) 100330.

[6] G. Jignesh Chowdary, Suganya. G, Premalatha. M, “Effective Prediction of Cardiovascular Disease Using Cluster of Machine Learning Algorithms”, Journal of Critical Reviews, ISSN- 2394-5125 VOL 7, ISSUE 18, 2020.

[7] Amin Ul Haq , Jian Ping Li , Muhammad Hammad Memon , Shah Nazir , and Ruinan Sun, “A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms”, Hindawi
Mobile Information Systems, Volume 2018, Article ID 3860146, 21 pages, https://doi.org/10.1155/2018/3860146.

[8] Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques”, In IEEE Access, 2019.

[9] Akram Ahmed Mohammed, Rajkumar Basa, Anirudh Kumar Kuchuru, Shiva Prasad Nandigama, Maneeshwar Gangolla, “Random Forest Machine Learning technique to predict Heart disease”, European Journal of Molecular & Clinical Medicine ISSN 2515-8260 Volume 7, Issue 4, 2020.

[10] Malavika G, Rajathi N, Vanitha V and Parameswari P “Heart Disease Prediction Using Machine Learning Algorithms “, Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-24-27

[11] Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain and Preeti Nagrath, “Heart disease prediction using machine learning algorithms” IOP Conf. Series: Materials Science and Engineering 1022 (2020) 012072.

[12] Riddhi Kasabe, Geetika Narang, "Heart Disease Prediction using Machine Learning", International Journal of Engineering Research & Technology (IJERT) ,Vol. 9 Issue 08, August-2020
[13] Dinesh Kumar G, Santhosh Kumar D, Arumugaraj K, Mareeswari V, "Prediction of Cardiovascular Disease Using Machine Learning Algorithms", Proceeding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India.
[14] Aditi Gavhane, Gouthami Kokkula, Isha Pandya, Kailas Devadkar, "Prediction of Heart Disease Using Machine Learning" Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018) IEEE Conference Record # 42487; IEEE Xplore ISBN:978-1-5386-0965-1
[15] Indu Yekkala, Sunanda Dixit, "Prediction of Heart Disease Using Random Forest and Rough Set Based Feature Selection", International Journal of Big Data and Analytics in Healthcare, Volume 3, Issue 1, January-June 2018.
[16] Nimai Chand Das Adhikari, Arpana Alka, and rajat Garg, "HPPS: Heart Problem Prediction System using Machine Learning". ACADEMY & INDUSTRY RESEARCH COLLABORATION CENTER (AIRCC), 2017.
[17] S Radhika, S Ramiya Shree, V Rukhmani Divyadharsini and A Ranjitha, "Symptoms Based Disease Prediction Using Decision Tree and Electronic Health Record Analysis", European Journal of Molecular & Clinical Medicine ISSN 2515-8260 Volume 7, Issue 4, 2020.
[18] K. Vembandasamy, R. Sasipriya P and E. Deepa, "Heart Diseases Detection Using Naive Bayes Algorithm", IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 9, September 2015
[19] E. Y. Boateng, and D. A. Abaye, "A Review of the Logistic Regression Model with Emphasis on Medical Research". Journal of Data Analysis and Information Processing, Vol.7. No.4, pp.190-207, 2019.
[20] J. Harrell and E. Frank, "Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis". Springer, 2015.
[21] Vaseem Naiyer, Jitendra Sheetlani, Harsh Pratap Singh, "Software Quality Prediction Using Machine Learning Application", Smart Intelligent Computing and Applications, 2020, Pages 319-327, Springer.
[22] Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). Classification and Regression Trees. Boca Raton, FL: CRC press.
[23] Breiman, L. (1996). Bagging predictors. Mach. Learn. 24, 123–140. doi: 10.1007/ BF00058655.
[24] Caruana, R., and Niculescu-Mizil, A. (2006). "An empirical comparison of supervised learning algorithms," in 23rd International Conference on Machine Learning, (Pittsburgh, PA: ACM Press), 161–168.
[25] Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., et al. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics 10:213. doi: 10.1186/1471-2105- 10-213.
[26] Ceriani, L., and Verme, P. (2012). The origins of the gini index: extracts from variabilità e mutabilità (1912) by corrado gini. J. Econ. Inequal. 10, 421–443. doi: 10.1007/s10888-011-9188-x.
[27] Strobl, C., Boulesteix, A. L., and Augustin, T. (2007). Unbiased split selection for classification trees based on the gini index. Comput. Stat. Data Anal. 52, 483–501. doi: 10.1016/j.csda.2006.12.030