IDEATION PHASE
LITERATURE SURVEY

Date	8 September 2022
Team ID	PNT2022TMID29391
Project Name	Visualizing and Predicting Heart Diseases with an Interactive Dashboard

Abstract- In order to forecast cardiac disease, this study discusses various data mining, big data, and machine learning techniques. Building an important model for the medical system to forecast heart disease or cardiovascular illness requires the use of data mining and machine learning. Medical professionals can assist patients by identifying cardiovascular illness before it manifests. Heart disease is one of the leading causes of death in the modern world. An important clinical challenge is the ability to forecast heart disease. But occasionally, a number of methods to forecast heart disease in data mining are found. Numerous methods for predicting heart disease were described in this survey publication.

Keywords – Data mining, cardiovascular disease, heart disease, machine learning

I. INTRODUCTION

Various disorders that impact the human heart are referred to as heart disease. The terms "heart disease" and "cardiovascular disease" are frequently used interchangeably. Heart disease is a general term that covers a wide range of heart-related medical conditions. The irregular health state that directly affects the heart and all of its components is characterized by these medical conditions. A heart attack, stroke, or chest pain can all be caused by illnesses that are generally made possible by restricted or obstructed blood arteries due to heart disease. Heart diseases are also characterized by other illnesses that affect the muscle, valves, or rhythm of the heart. Heart disease comes in many different forms. The most similar types are heart failure (HF) and Coronary Artery Disease (CAD). Data mining is a complicated process that uses intricate algorithms to extract implicit, previously undiscovered possibly useful information known as knowledge from medical data. Big data (BD) is a term used to describe large records of information. Data mining and big data are two distinct concepts. These two strategies accomplish the same job, which centres on gathering a sizable amount of data, managing them, and creating reports on the data by removing the knowledgeable information. Using Big Data, data mining fundamentally involves identifying meaningful patterns in data that contain specific information.

II. LITERATURE SURVEY

[1] The authors of this research used the decision tree and the hill climbing algorithms. The data is pre-processed before the classification algorithms are applied. Cleveland data set is the one that was used. Evolutionary Learning (KEEL), an open-source data mining method, is used in the Knowledge Extraction process to fill in the missing values in the data set. The best collection of rules is then discovered using a hill climbing algorithm. The parameters and their

corresponding values are: MinItemsets: The minimum number of item-sets per leaf is two; Confidence: The minimum confidence value is 0.25; The authors search for the optimal collection of rules for the hill-climbing algorithm with a threshold of 10. For each level, a decision tree is built using a top-down technique, and a test for the actual node picked using a hill-climbing algorithm is used to choose a node. The decision tree is capable of producing the fundamental rules. Original rules are the first generated rules, and pruned rules are generated from these original rules. These rules are obtained, and they are sorted and free of duplication. Finally, a limited set of regulations known as Class wise Rule distribution are produced. About 86.7% of the time, the system is accurate.

[2] The author of this paper uses hybrid machine learning to forecast cardiac disease. Cleveland data set is the one that was used. Data pre-processing is the first step. In this, the tuples with missing values are eliminated from the data set. Age and sex data set attributes are also not used because, in the authors' opinion, they are private and have no bearing on prediction. The 11 remaining qualities are significant because they include crucial clinical records. Their own Hybrid Random Forest Linear Approach (HRFLM), a hybrid of the Random Forest (RF) and Linear method, has been proposed (LM). The authors employed four algorithms in the HRFLM algorithm. The first algorithm divides the incoming dataset into sections. It is built on a decision tree, which is used to process each dataset sample. The dataset is divided into leaf nodes after the feature space has been determined. Partitioning of the data set is the first algorithm's output. Following that, they apply rules to the data set in the second algorithm, and the output is the classification of the data using those rules. Utilizing a Less Error Classifier, features are extracted in the third algorithm. The goal of this approach is to determine the classifier's minimum and maximum error rates. This algorithm produces features with classified attributes as its output. In forth algorithm they apply Classifier which is hybrid method based on the error rate on the Extracted Features. Finally, they have compared the results obtained after applying HRFLM with other classification algorithms such as decision tree and support vector. As a result of RF and LM producing superior outcomes to previous algorithms, HRFLM, a novel and original algorithm, was developed. The HRFLM's accuracy originally improved with number splits before stabilizing at a certain level. The acquired accuracy of 88.7% is more than that of the SVM and decision tree. The authors advise combining different machine learning algorithms in order to increase accuracy even further. They also advise focusing on the creation of cutting-edge feature selection approaches that would aid in the extraction of important information.

[3] In this study, the authors suggest a system with two linear Support Vector Machine-based models (SVM). The first is referred to as L1 regularized, and the second is referred to as L2 regularized. By setting the coefficient of any unneeded characteristics to zero, the first model is used to remove those features. For prediction, the second model is employed. In this section, disease prediction is carried out. They suggested a hybrid grid search algorithm to optimize both models. Based on parameters like accuracy, sensitivity, specificity, the Matthews correlation coefficient, ROC chart, and area under the curve, this technique optimizes two models. The Cleveland data set was utilized. 70% of the data were used for training, and 30% for testing, using holdout validation. Each experiment is run twice for different values of C1, C2, and k, where C1 is the hyperparameter of the L1 regularized model, C2 is the hyperparameter of the L2 regularized model, and k is the number of

features in the chosen subset. When C1 = 0.200, k = 11, and C2 = 0.500 are used, the L1-linear SVM model gives a maximum testing accuracy of 91.11% and a training accuracy of 84.05%. This is the first trial. L1 and L2 linear SVM models with RBF kernels are cascaded in the second experiment. This is giving maximum testing accuracy of 92.22% and training accuracy of 85.02% for value of C1 = 0.060, k = 8 and C2 = 400.00 and G (Hyperparameter of the L2-SVM model with RBF kernel) = 0.015. They have obtained an improvement in accuracy over conventional SVM models by 3.3%.

[4] In this paper authors deal with various supervised machine learning algorithms such as Random Forest, Support Vector Machine, Logistic Regression, Linear Regression, Decision Tree with 3fold, 5fold and 10fold cross-validation techniques. They have used Cleveland data set having 303 tuples, with some tuples having missing attributes. In the preprocessing of data, they just removed the missing value tuple from the data set which are six in number and then from the remaining 297 tuples, they divided the data as training 70% and testing 30%. First algorithm applied is Linear Regression. In this, they have defined the dependency of one attribute over others which can be linearly separated from each other. Basically, the classification takes place with the help of the group of attributes used for binary classification. They have obtained best results in 10fold which is 83.82%. Logistic regression classification is done using a sigmoid function. This algorithm applied for heart disease prediction shows maximum accuracy with 3 and 5fold cross-validation and it is 83.83%. Support Vector Machine is the classification algorithm in supervised machine learning. In this the classification is done by hyperplane. The maximum accuracy achieved by SVM in 3fold cross-validation is 83.17%. To determine the decision tree's maximum accuracy, the authors of this study experimented with various number splits and leaf node counts. The maximum accuracy is attained with 37 number splits and 6 leaf nodes, which is 79.12%. When employed with cross-validation, the decision tree's accuracy was 79.54% with 5fold. When applied to a nonlinear data set, the random forest technique performs better than the decision tree. The collection of decision trees known as a random forest was produced by several root nodes. Voting can be done from this group of decision trees, and classification can then be done from the decision that receives the most votes. Authors have used different number splits, different number of tree per observation and different number of folds for cross-validation. For random forest, 85.81% accuracy is achieved by 20 Number of splits, 75 Number of trees and 10 number of folds

[5] The authors of this work discuss machine learning strategies for heart disease prediction, including decision trees and the Naive Bayes algorithm. In the first algorithm, a decision tree is constructed utilizing a set of circumstances that result in True or False conclusions. Other algorithms, such as SVM and KNN, base their conclusions on dependent variables and either vertical or horizontal split conditions. However, a decision tree is a structure that resembles a tree with a root node, leaves, and branches, and it is based on the decisions made in each tree. Decision trees also aid in recognizing the significance of the dataset's properties. They have also used Cleveland data set. Dataset splits in 70% training and 30% testing. This algorithm gives a 91% accuracy. The second algorithm is Naive Bayes. It is used for classification. It can handle complicated, nonlinear, dependent data and hence is found suitable for heart disease dataset as this dataset is also complicated, dependent and nonlinear in nature. This algorithm gives an 87% accuracy.

[6] Bo Jin, Chao Che et al. (2018) proposed a "Predicting the Risk of Heart Failure With EHR Sequential Data Modeling" model designed by applying neural network. This paper used the electronic health record (EHR) data from real-world datasets related to congestive heart disease to perform the experiment and predict the heart disease before itself. We tend to used one-hot encryption and word vectors to model the diagnosing events and foretold coronary failure events victimization the essential principles of an extended memory network model. By analyzing the results, we tend to reveal the importance of respecting the sequential nature of clinical records.

[7] Aakash Chauhan et al. (2018) presented "Heart Disease Prediction using Evolutionary Rule Learning". This study eliminates the manual task that additionally helps in extracting the information (data) directly from the electronic records. To generate strong association rules, we have applied frequent pattern growth association mining on patient's dataset. This will facilitate (help) in decreasing the amount of services and shown that overwhelming majority of the rules helps within the best prediction of coronary sickness

[8] "Prediction and Diagnosis of Heart Disease by Data Mining Techniques" designed by Boshra Bahrami, Mirsaeid Hosseini Shirvani. This paper uses various classification methodology for diagnosing cardiovascular disease. Classifiers like KNN, SVO classifier and Decision Tree are used to divide the datasets. Once the classification and performance evaluation the Decision tree is examined as the best one for cardiovascular disease prediction from the dataset.

[9] M.Satish, et al. (2015) used different Data Mining techniques like Rule based, Decision Tree, Navie Bayes, and Artifical Neural Network. An efficient approach called pruningclassification association rule (PCAR) was used to generate association rules from cardiovascular disease warehouse for prediction of Heart Disease. Heart attack data warehouse was used for pre-processing for mining. All the above discussed data mining technique were described.

[10] Lokanath Sarangi, Mihir Narayan Mohanty, Srikanta Pattnaik (2015) "An Intelligent Decision Support System for Cardiac Disease Detection", designed a cost-efficient model by using genetic algorithm optimizer technique. The weights were optimized and fed as an input to the given network. The accuracy achieved was 90% by using the hybrid technique of GA and neural networks

V. CONCLUSION

This overview of the literature conveys the idea that numerous methods have been investigated for diagnosing cardiovascular disease. Big data, machine learning, and data mining can be used to great success to analyze the prediction model with the highest degree of accuracy. The primary goal of this study is to diagnose cardiovascular disease or heart disease utilizing a variety of techniques and procedures to obtain a prognosis.

REFERENCES

[1] Purushottam, Kanak Saxena and Richa Sharma, "Efficient heart disease prediction system." Procedia Computer Science 85 (2016): 962-969.

[2] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava, "Effective heart disease prediction using hybrid machine learning techniques." IEEE Access 7 (2019): 81542-81554.

[3] Ali, Liaqat, et al, "An optimized stacked support vector machines based expert system for the effective prediction of heart failure." IEEE Access 7 (2019): 54007-54014. www.ijcrt.org © 2020 IJCRT | Volume 8, Issue 8 August 2020 | ISSN: 2320-2882 IJCRT2008170 International Journal of Creative Research Thoughts (IJCRT) www.ijcrt.org 1606

[4] Singh, Yeshvendra K., Nikhil Sinha, and Sanjay K. Singh, "Heart Disease Prediction System Using Random Forest", International Conference on Advances in Computing and Data Sciences. Springer, Singapore, 2016.

[5] Santhana Krishnan. J, Geetha S., "Prediction of Heart Disease Using Machine Learning Algorithms", 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)

[6] Bo Jin ,Chao Che, Zhen Liu, Shulong Zhang, Xiaomeng Yin, And Xiaopeng Wei, "Predicting the Risk of Heart Failure With EHR Sequential Data Modeling" ,IEEE Access 2018.

[7] Aakash Chauhan , Aditya Jain , Purushottam Sharma , Vikas Deep, "Heart Disease Prediction using Evolutionary Rule Learning", "International Conference on "Computational Intelligence and Communication Technology" (CICT 2018).

[8] Boshra Bahrami, Mirsaeid Hosseini Shirvani, "Prediction and Diagnosis of Heart Disease by Data Mining Techniques", Journal of Multidisciplinary Engineering Science and Technology (JMEST) ISSN: 3159-0040 Vol. 2 Issue 2, February–2015.

[9] M.Satish, D Sridhar, "Prediction of Heart Disease in Data Mining Technique", International Journal of Computer Trends & Technology (IJCTT), 2015.

[10] Lokanath Sarangi, Mihir Narayan Mohanty, Srikanta Pattnaik, "An Intelligent Decision Support System for Cardiac Disease Detection", IJCTA, International Press 2015.