JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Analysis and Prediction of Cardiovascular Disease using Machine Learning Technique

[1]Dr. S. Haseena , [2]E Priyadharshini , [3]V Shamili

[1]Assistant Professor (Selection Grade),

[1,2,3] Department of Information Technology,

[1,2,3]Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu.

Abstract : Globally, one of the main causes of death is heart disease. A significant challenge in clinical data analysis is predicting cardiovascular disease. The ability of AI to decide and anticipate a vast amount of information generated by the health sector has been established. By utilizing machine learning techniques, we offer a special technique for identifying crucial qualities. In order to acquire the correct results, doctors must attempt to save human lives. As a result, a system was created to forecast the likelihood of developing heart disease as well as common symptoms such age, gender, glucose, smoke, and cholesterol. Doctors utilize these to verify the health of their patients. In order to increase accuracy, autocorrelation, auto regression have been utilized along with machine learning techniques like multiple linear regression and correlation coefficient. This work increases the accuracy of diagnosing cardiac problems. The classification method used to develop the suggested system include support vector machine (SVM). By using this classification model, a single best-fit predictive model is produced. Numerous investigations have been made in order to identify heart infection, however the accuracy of the results

Index Terms **- cardiovascular disease, regression, correlation, prediction**

I. INTRODUCTION

A set of ailments that affect people's hearts and veins are referred to as heart disease. Different people experience different cardiac illness symptoms. Heart illness refers to a number of problems like hypertension, stroke, respiratory failure, and arrhythmia are distinguished by their adverse effects. The challenge for providers of medical services is to provide top-notch care at an affordable price. Poor therapy and an incorrect clinical diagnosis may lead to insufficient results. Cardiovascular disease detection and diagnosis is a never-ending effort that can be carried out by a qualified professional with extensive knowledge and expertise. Healthcare businesses may employ decision support systems (DSSs) to save expenses. Aspects of healthcare including patient records, different disease diagnoses, resource management, and others are typical. Cardiovascular disease has the biggest level of passing on the planet. In 2012, around 17.5 million people were killed from cardiovascular disease, implying that it comprises 31% for every single passing. Since 2017 the main source of death in our country is cardiovascular infections. Because of numerous contributing risk factors, including diabetes, high blood pressure, excessive cholesterol, irregular pulse rate, and many other factors, it is difficult to diagnose heart disease. Since cardiac illness has a complex character, it requires cautious management. Establishing remote monitoring devices and procedures for patient diagnosis has taken a lot of work. On the other hand, device weariness has been considered a barrier to conformity. Commercially available technologies have demonstrated the ability to overcome this barrier by reducing the need for human contact, and the accuracy of activity trackers has been demonstrated to be sufficient for health workers. The development of information technology systems has led to an ongoing production, processing, and evaluation of clinical records. Clinical reports have information that can be used to create innovative healthcare services that address social and economic challenges globally. For instance, clinical reports may include a range of numerical data, descriptions of medical conditions, pictures, etc. which are all can be utilized to develop content-based services that benefit patients and medical professionals. It has been extensively suggested that reducing mortality rates and improving selection for future diagnoses and interventions can be accomplished by quickly diagnosing cardiovascular disease in high-risk patients and speeding up detection using a prediction system. Clinicians may use a decision support system (DSS) framework to help them assess the likelihood of cardiovascular problems and prescribe the best medications to further avoid their occurrence. Additionally, a number of studies have shown that using a DSS can improve decision-making, treatment planning, and preventative services. Heart disease is effectively detected using an expert DSS based on a machine learning (ML) model. For preparation and testing, the ML foresight models require accurate data. ML is now being used in hospitals to help with the planning and management of infectious diseases, the structuring of administrative processes, and the personalization of medical care. Additionally, the device is being used in a variety of heart related fields, as well as the development of new medical procedures, the management of patient data, and the treatment of chronic illnesses. In recent years, the Carotid Artery Stenting (CAS) procedure has also grown in popularity as a therapy option in the medical field. Major adverse cardiovascular events (MACE) in senior heart disease patients are triggered by the CAS. Their assessment becomes crucial. The following shows the breakdown of this intended work. Section 2 goes into detail on the work related to the planned task. Section 3 goes into brief of analysis of heart disease Section 4 goes into detail about the proposed framework. The fifth section digs into the implementation of supervised learning technique. Section 6 goes into result analysis. Section 7 discusses the conclusion as well as future efforts.

II. RECENT STUDY

SonamNikhar[4] work on extrapolating a variety of medical conditions, including as diabetes, breast cancer, and cardiovascular diseases, utilizing multiple machine learning algorithms that produce a range of accuracy. To determine whether or not a patient had cardiac disease, classification is carried out using the Cleveland dataset [7]. Three primary benefits of the ML algorithms when paired with feature extraction goals: I to identify the distinctive characteristics, (ii) to assess PCA's effectiveness, and (iii) to look into the model that yields superior outcomes. Heart disease prediction using big and small data sets was briefly discussed by Himanshu et al. [9]. They mentioned that utilizing the SVM algorithms to accomplish prediction, small data sets require the least amount of training and testing time. Discussed heart disease prediction and demonstrated that some machine learning methods do not perform better for accuracy prediction, although hybridization can produce good accuracy. Devansh Shahet.al.[10] research on a significant body of heart disease data. They analyzed numerous machine learning algorithms and discovered the performing accuracy comparisons on various machine learning would yield the best accuracy. Researchers have proposed several ML-based diagnostic methodologies for HD. ML algorithms [11,12] have been extensively utilized in a variety of research such as disease recognition and identification. Guidi et al. [14] helped to create a DSS for heart failure analysis (HF). They investigated the effectiveness of neural networks (NN), SVM, CART-based fuzzy rules, and random forests, among other ML classifiers (RF). The CART model with random forests produced the best results, with an accuracy of 87.6 percent. Parthiban and Srivatsa [15] explored an SVM approach for detecting HD in diabetic patients with an efficiency of 95%. [16] employed a binary classifier that correctly identified heart disease with a 100% improvement in accuracy. S.Haseena et al[17] research on a Moth-Flame Optimization for Early Prediction of Heart Diseases. They analysed many machine learning algorithms and discovered the performing accuracy comparisons on various machine learning would yield the best accuracy. Detrano et al. [18] used ML Classification techniques to construct the HD classification system, which had 77 percent accuracy. Ahmed [19] proposed an HD identification algorithm employing IoT architecture with SVM. To forecast heart disease, the patient data was analyzed using an SVM. The researchers claim that their method predicted cardiac disease with 97.53 percent accuracy.

III. ANALYSIS OF CARDIOVASCULAR DATA

Analysis of Cardiovascular Data Based on spread cases reported in India, the Cardiovascular dataset is examined for cardiovascular disease. We gathered information on some of the worst-affected patients based on age. The data was obtained from www.kaggle.com. The dataset consists of 68,784 records of patient's data in 12 features, such as age, gender, height, weight, ap-high, ap-low, smoke, cholesterol, glucose, alcohol, physical activity, and cardio disease. The target class "cardio" equals to 1, when patient has cardiovascular disease, and if the value is 0, then the patient is healthy. The task is to predict the presence or absence of cardiovascular disease (CVD) using the patient examination results. The objective of this analysis is to find the correlation between for all spread cases from the dataset. Through this analysis, it is observed that there is a strong correlation between the complete dataset.

IV. PROPOSED SYSTEM DESIGN

The first step in the proposed system design is to obtain the data for this download from Kaggle. There are numerous steps in this process, as depicted in the block diagram in Fig. 1. The proposed system design to predict cardiovascular disease is done by four algorithms. They are correlation coefficient, Multiple Linear regression, autocorrelation and auto-regression. This dataset's attributes primarily took confirmed cases into account. This disease also causes to other healthy people too. The gathered data were examined as the model was being developed using Python software routines. A statistical description of the entire dataset was run in order to adequately comprehend it.

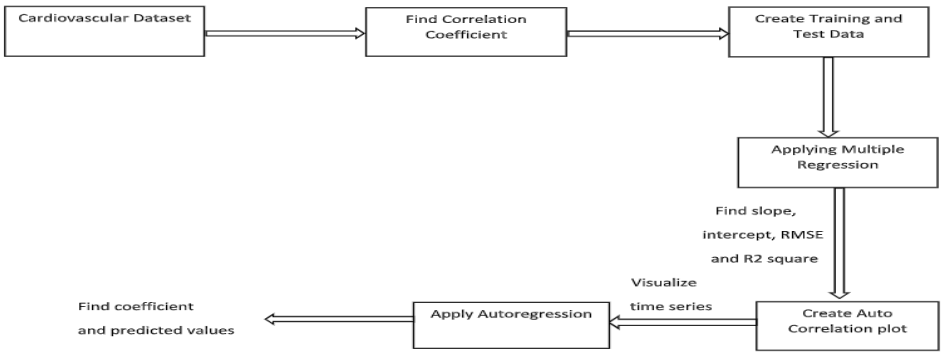

Fig. 1 Proposed model for predictions of confirmed, recovered, and death cases of cardiovascular disease.

In Fig. 1, the proposed model is summarized. Finding and calculating the dataset's degree of variables is crucial since it allows for better dataset preparation to satisfy the demands of the algorithms. Using Python software, correlation analysis and a recovery plan are applied to the data. It provides a statistical summary of instances that have been confirmed, recovered, or have resulted in death, and it also identifies a high correlation between recent data.

In order to explain the association between two independent variables (age and gender) and one dependent variable (spread cases), multiple linear regression techniques are used. In this instance, the dataset is split into training and test datasets with 70% training and 30% testing ratio.

After being trained on the dataset, this model has a very high degree of prediction ability, with an R2 score and a Root Mean Square Error (RMSE) The model can utilize regression against itself, and it can also use the autocorrelation plot to check for randomness in the data, according to the study of the entire dataset. After that, input your observations from the previous steps into an auto regression model. The values for the following time step are predicted using the time-series model. Results show that the predicted time series range is accurate. Find the lag and coefficients by fitting the model using the current dataset. Fig.2 Boxplot for dataset of cardiovascular disease. For confirmed, recovered, and dying cases, separate analyses are carried out based on the lag value.

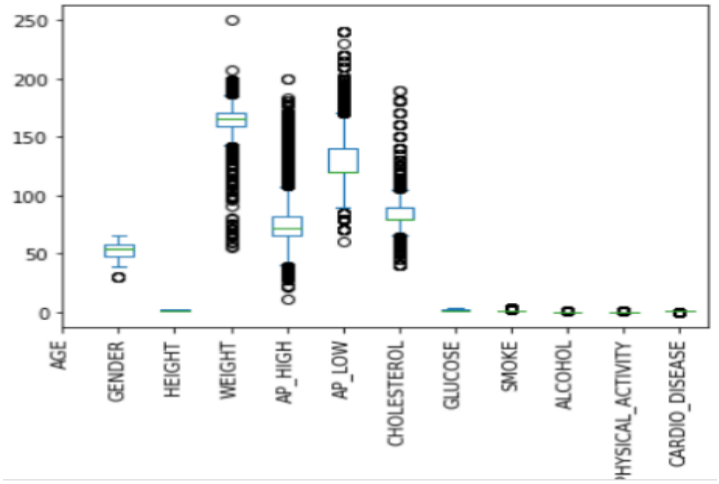

Fig.2 Boxplot for dataset of cardiovascular disease.

A) Data collection and pre-processing

The "cardiovascular" dataset from Kaggle was used for the analysis. The UCI Machine Learning Repository contains a heart disease dataset with 14 features. The median of the values in that column is used to replace all the unavailable values in the data. Numerical values are ascribed to categorical data. Data preprocessing is the process of transforming raw data into an understandable format. Raw datasets are usually characterized by incompleteness, inconsistencies, lacking behavior, and trends while containing errors. The preprocessing is essential to handle the missing values and address inconsistencies.

B) Data Cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct.

C) Correlation coefficient

By using the cardiovascular test data, the correlation coefficient method is used to determine the dependency between the variables that varies between -1 and 1. It is used to determine whether the columns are highly correlated or not. Pearson correlation method is the most common method to use for numerical variables; it assigns a value between − 1 and 1, where 0 is no correlation, 1 is a total positive correlation, and − 1 is a total negative correlation.

D) Multiple Linear Regression

It is a predictive analysis. It is used to predict the target variable using more than one independent variable. The multiple linear regression techniques are used to explain the relationship between two independent variables of the confirmed and recovered cases. and one dependent variable of the predicted cases. Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.

E) Auto Correlation

The auto-correlation analysis measures the relationship of the observations between the different points in time and seeks a pattern over the time series. The analysis of autocorrelation is a mathematical tool for finding repeating patterns such as identifying the plot for variation in death cases. The analysis of auto-correlation is a mathematical tool for finding repeating patterns such as identifying the plot for variation in death cases.

F) Auto Regression

Auto regression is a time series model that uses observations from previous time steps as input to a regression equation to predict the value at the next time step. It is used to predict the number of cases in the future. It is suitable where one wants to measure the influence of variables by correlating with the previous or past state.

$$X_t = b_0 + b_1 X_{t-1} + \varepsilon_t \quad (1)$$

Where, X_t is the value of time series at time t, b0 is the intercept at the vertical axis(y-axis), b1 is the slope coefficient, X_{t-1} is the value of time series at time t-1, ε_t is the error term or residual term or disturbance term. t is the time ranging from 1,2,3,…T

G) Data Visualization

Data visualization is a process of reading the dataset or the outcome of the data analysis and processing then to find out events that are likely to occur in the future.

V. IMPLEMENTATION

A) Support Vector Machine

Support vector machines are a classification method that divides input values by producing hyper planes. Based on the distribution of the data, hyper planes can take on a variety of shapes, but only those points are useful for distinguishing between the classes that are taken into account for classification.

The Support Vector Machine's implementation is explained as follows:

a) Load the data sets and clean the values; if a row lacks a value for a certain attribute, replace it with the dataset's median value for that row.

b) Divide the data set 80:20 into the train and test subsets.

c) Using test data to first construct a hyper plane before using SVM.

d) Determine the accuracy utilizing the training data is used.

e) Use the trained model with the test set of data.

The model makes use of a hyper plane to determine the closest proximity.

Therefore, the best technique to categorize such data would be using a hyper plane in the shape of a line. Fig.3 displays the hyper plane map for SVM for predicting heart disease. Patients with heart disease are represented in this by the yellow plot, while those without heart disease are represented by the blue dots. Fig.4 displays the count of positive and negative cases of cardiovascular disease. Fig.5 displays the form of most separated clusters for age, cardiovascular disease and AP_low. Fig.6 represents the degree of correlation of same variables between two successive time intervals.

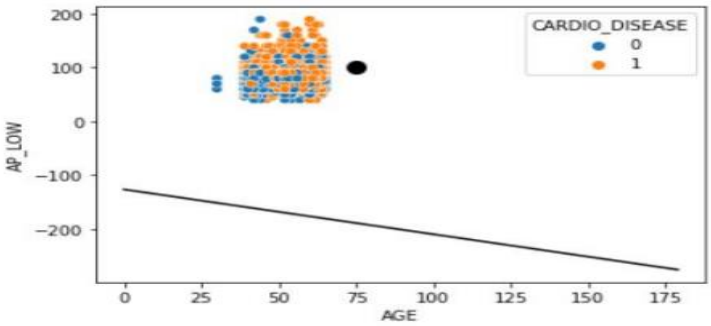

Fig.3 Hyper plane and distribution of data points on hyper plane for Heart Disease Prediction

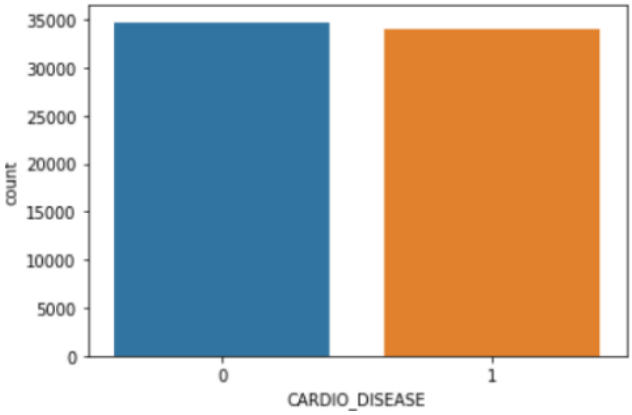

Fig.4 positive and negative cases of cardiovascular disease

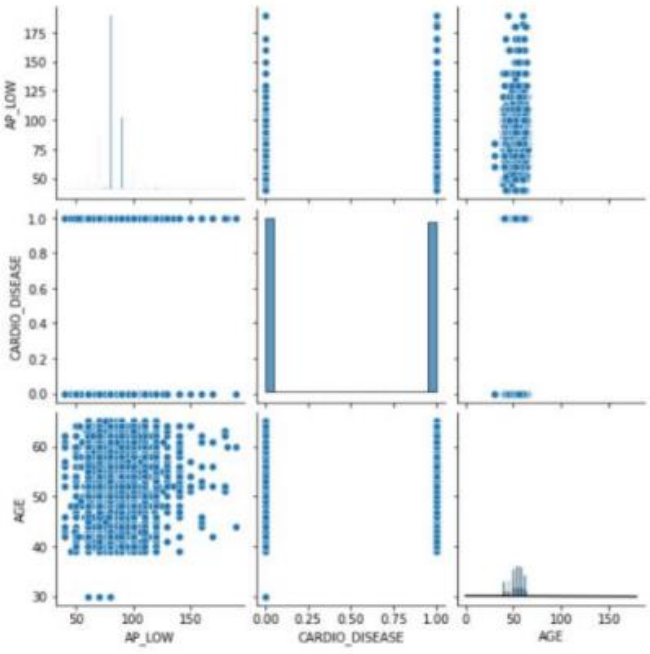

Fig.5 separated clusters.

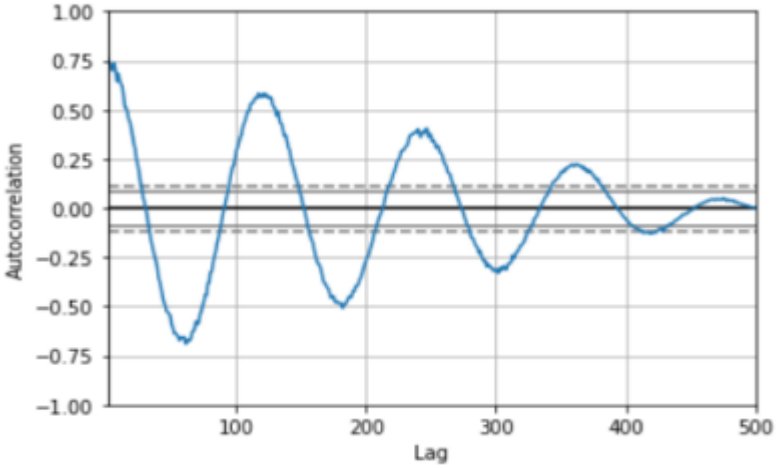

Fig.6 Autocorrelation Plot

VI. CONCLUSION

Heart disease prediction which uses a machine learning algorithm provides uses a prediction result if the user has heart disease or not. From the proposed system design the correlation coefficient and multiple linear regression are used to find the efficiency and accuracy. Also, this algorithm is used to find the heart disease prediction by knowing the correlation details between heart disease and another disease. The proposed model may be extended to predict the end of this pandemic in a particular region. Total causality and total economic losses may be predicted with the help of this model.

REFERENCES

[1] Sana Bharti, Shailendra Narayan Singh" Analytical study of heart disease compared with different algorithms": Computing, Communication & Automation (ICCCA), 2015InternationalConference.

[2] Geert Meyfroidt, FabianGuiza, Jan Ramon, Maurice Brynooghe" Machine learning techniques to examine large patient databases"-Best practice & Research Clinical Anaesthesiology, Elsevier Volume 23 (1) Mar 1, 2009.

[3] Sanjay Kumar Sen" Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms", International Journal of Engineering and Computer Science ISSN:2319-7242Volume6Issue 6 June 2017.

[4] SonamNikhar, A.M. Karandikar" Prediction of Heart Disease Using Different Machine Learning Algorithms"- Vol-2 Issue-6, June 2016.

[5] MatjazKuka, Igor Kononenko, Cyril Groselj, Katrina Kalif, JureFettich" Analysing and improving the diagnosis of ischaemic heart disease with machine learning" Elsevier -Artificial intelligence in Medicine, Volume23, May 1999. ISSN:2319-7242Volume6Issue 6 June 2017.

[6] Machine learning based decision support systems (DSS) for heart disease Diagnosis: a review. Online: 25 March 2017 DOI: 10.1007/s10462-01

[7]Dataset URL https://archive.ics.uci.edu/ml/machielearnind atabses/heartdisease

[8] Archana Singh, Rakesh k. (2020)." Heart disease Prediction Using machine Learning Algorithms" International Conferences on Electrical and electronics Engineering (ICE3).

[9] Himanshu Sharma and M A Rizvi. (2017). "Prediction of Heart Disease using Machine Learning Algorithms: A Survey" International Journal on Recent and Innovation Trends in Computing and Communication Volume: 5 Issue: 8 , IJRITCC August 2017.

[10] Devansh Shahet.al.(2020)"Heart Disease Prediction using Machine Learning Techniques" © Springer Nature Singapore Pte Ltd.

[11] M. B. B. Pepsi and S. N. Kumar, "Supervised learning techniques for classification of students' tweets," Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, no. 12, pp. 1714–1722, 2021.

[12] M. S. Bhuvaneswari and K. Muneeswaran, "A parallel approach for web session identification to make recommendations efficient," International Journal of Business Intelligence and Data Mining, vol. 19, no. 2, pp. 189–213, 2021.

[13] R. C. Deo, "Machine learning in medicine," Circulation, vol. 132, no. 20, pp. 1920–1930, 2015.

[14]G. Guidi, M. C. Pettenati, P. Melillo, and E. Iadanza, "A machine learning system to improve heart failure patient assistance," IEEE journal of biomedical and health informatics, vol. 18, no. 6, pp. 1750–1756, 2014.

[15]G. Parthiban and S. K. Srivatsa, "Applying machine learning methods in diagnosing heart disease for diabetic patients," International Journal of Applied Information Systems, vol. 3, no. 7, pp. 25–30, 2012.

[16]R. S. Singh, B. S. Saini, and R. K. Sunkaria, "Detection of coronary artery disease by reduced features and extreme learning machine," Clujul Medical, vol. 91, no. 2, pp. 166– 175, 2018.

[17] S. Haseena,S. Kavi Priya,S. Saroja , R. Madavan,M. Muhibbullah , and Umashankar Subramaniam "Moth-Flame Optimization for Early Prediction of Heart Diseases" Computational and Mathematical Methods in Medicine Volume 2022, Article ID 9178302,.

[18] H. Miao Kathleen, H. Miao Julia, and G. J. Miao, "Diagnosing coronary heart disease using ensemble machine learning," International Journal of Advanced Computer Science and Applications, vol. 7, no. 10, 2016.

[19]R. Detrano, A. Janosi, W. Steinbrunn et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," The American Journal of Cardiology, vol. 64, no. 5, pp. 304– 310, 1989.

[20]F. Ahmed, "An Internet of Things (IoT) application for predicting the quantity of future heart attack patients," International Journal of Computers and Applications, vol. 164, no. 6, pp. 36–40, 2017. [20]G. Guidi, M. C. Pettenati, P. Melillo, and E. Iadanza, "A machine learning system to improve heart failure patient assistance," IEEE journal of biomedical and health informatics, vol. 18, no. 6, pp. 1750–1756, 2014.