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

Heart Disease Prediction Using Machine Learning and Risk Analysis

Ms Sarika Sanjay Wabale*, Prof. Dr. N.R. Wankhede**

**M.E Student, Department of computer science and Engineering, Late G.N.Sapkal College Of Engineering, Nasik*
***Assistant Professor, Department of computer science and Engineering, Late G.N.Sapkal College Of Engineering, Nasik*

Abstract: Cardiovascular diseases (CVDs), which are diseases of the heart, are the main cause of the large number of fatalities that have occurred over the course of the most recent few years and have become the most dangerous disease in India and throughout the entire world. In this approach, there might be a need for a precise, workable, and reliable tool to study these illnesses in time for effective therapy. Numerous clinical datasets were used in conjunction with machine learning methods and techniques to conduct extensive and complex information research. In recent years, several analysts have used a variety of methodologies to provide the health care industry and internal specialists with the prediction of heart-related disorders. This study presents a survey of many models that are entirely based on such algorithms and techniques and examines their effectiveness. Models using supervised learning techniques, such as Support Vector Machines (SVM), Logistics Regression, Artificial Neural Network and random forest ensemble models are incredibly distinctive among the many researchers.

Keywords: Cardiovascular, datasets, Supervised learning algorithms, Support Vector Machines, Logistics Regression, Artificial Neural Network

1. Introduction:

Data Mining is a non-minor extraction of certain, beforehand obscure and potentially valuable information about data [1]. In short, it is a procedure of analyzing information from the substitute perspective of view and assembling the knowledge of it [2]. The discovered information can be used for various applications, for example healthcare industry. The Healthcare industry is"data rich", however lamentably not every one of the information is dug which is required for finding hidden patterns and effective decision making. Data Mining Techniques such as Propelled data mining techniques are Utilized to find learning in the database and for medicinal research, especially in the Heart disease prediction. A major challenge facing the healthcare industry is the nature of the administration. Poor analysis can prompt appalling outcomes which are unacceptable. The datasets are overwhelming for human personalities to fathom, can be effectively investigated utilizing different machine learning techniques. Accordingly, these algorithms have become very useful, in recent times, to

predict the presence or absence of heart related diseases accurately. The doctors are embracing many scientific technologies. Our project's objective is to foresee the odds of heart disease based on the patient's dataset and the doctor's perspective in check-up using machine learning. By utilizing the patient's medical records, a new system is proposed to foresee the chances of heart attack Attributes such as Blood pressure (bp), age, thickness of the artery, etc. are sustained into the dataset and algorithm [3]. In the proposed system demonstrates a survey of numerous fashions primarily based totally on such algorithms and strategies and analyzes their performance. Models rely on supervised mastering algorithms along with Support vector machine , Logistic Regression , Artificial Neural Network and ensemble fashions are located extraordinarily distinguished the various researchers. In the proposed system our plan is to taking the accuracy of algorithm and then finally classifies the heart disease into high risk, low risk, and medium risk with the help of certain attribute factors.

2. Literature Review:

C. Boukhatem, H. Y. Youssef and A. B. Nassif, proposes the system of cardiovascular disease detection model has been developed using three ML classification modeling techniques. This project predicts people with cardiovascular disease by extracting the patient medical history that leads to a fatal heart disease from a dataset that includes patients' medical history such as chest pain, sugar level, blood pressure, etc. This Heart Disease detection system assists a patient based on his/her clinical information of them been diagnosed with a previous heart disease. The algorithms used in building the given model are Logistic regression. With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for detection of heart diseases. [1]

This study compares the accuracy score of Decision Tree, Logistic Regression, Random Forest and Naive Bayes algorithms for predicting heart disease using UCI machine learning repository dataset. The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with accuracy score of 90.16% for prediction of heart disease. In future the work can be enhanced by developing a web application based on the Random Forest algorithm as well as using a larger dataset as compared to the one used in this analysis which will help to provide better results and help health professionals in predicting the heart disease effectively and efficiently.[2]

Senthil Kumar Mohan et al,[3] proposed Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques in which strategy that objective is to finding critical includes by applying Machine Learning bringing about improving the exactness in the expectation of cardiovascular malady. The expectation model is created with various blends of highlights and a few known arrangement strategies. This system produces an improved exhibition level with a precision level of 88.7% through the prediction model for heart disease with hybrid random forest with a linear model (HRFLM) they likewise educated about Diverse data mining approaches and expectation techniques, Such as, KNN, LR, SVM, NN, and Vote have been fairly famous of late to distinguish and predict heart disease.

Sonam Nikhar et al [4] has built up the paper titled as Prediction of Heart Disease Using Machine Learning Algorithms by This exploration plans to give a point by point portrayal of Naïve Bayes and decision tree classifier that are applied in our examination especially in the prediction of Heart Disease.

Some analysis has been led to think about the execution of prescient data mining strategy on the equivalent dataset, and the result uncovers that Decision Tree beats over Bayesian classification system Prof. Kailas Devadkar (PhD), [5] Prediction of Heart Disease Using Machine Learning", In this paper proposed system they used the neural network algorithm multi-layer perception (MLP) to train and test the dataset. In this algorithm there will be multiple layers like one for input, second for output and one or more layers are hidden layers between these two input and output layers. Each node in input layer is connected to output nodes through these hidden layers. This connection is assigned with some weights. There is another identity input called bias which is with weight b, which added to node to balance the perception. The connection between the nodes can be feed forwarded or feedback based on the requirement.

Abhay Kishore et al,[6] developed Heart Attack Prediction Using Deep Learning in which This paper proposes a heart attack prediction system using Deep learning procedures, explicitly Recurrent Neural System to predict the probable prospects of heart related infections of the patient. Recurrent Neural Network is a very ground-breaking characterization calculation that utilizes Deep Learning approach in Artificial Neural Network. The paper talks about in detail the significant modules of the framework alongside the related hypothesis. The proposed model deep learning and data mining to give the precise outcomes least blunders. This paper gives a bearing and point of reference for the advancement of another type of heart attack prediction platform for the prediction stage.

Prediction and Analysis of the Occurrence of Heart Disease Using Data Mining Techniques was advised by ChalaBeyene et al[7]. The major goal is to foresee the development of heart disease in order to quickly and automatically diagnose the condition. The suggested methodology is essential in a healthcare company with professionals that lack updated knowledge and abilities. To determine if a person has heart disease or not, various medical characteristics are used, such as blood sugar and heart rate, age, and sex. WEKA software is used to compute analyses of the dataset.

To predict cardiac illness, Senthilkumar Mohan et al.[8] used hybrid machine learning. Cleveland data set is the one that was used. Data pre-processing is the first step. In this, the tuples from the data set that were missing values are eliminated. 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 Technique (HRFLM), which combines 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. Its foundation is a decision tree which is executed for each sample of the dataset. 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 the fourth strategy, they use a hybrid classifier that is based on the extracted feature error rate. The results of applying HRFLM were compared with those of other classification methods, including decision trees and support vector machines, in the end. As a result of RF and LM producing superior outcomes to previous algorithms, HRFLM, a novel and original algorithm, was developed. The authors propose combining different machine learning techniques to increase accuracy even further.

Ali, Liaqat, and colleagues[9] offer a system with two linear Support Vector Machine-based models (SVM). The first is referred to as L1 regularised, and the second is referred to as L2 regularised. 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 optimise both models. Based on parameters like accuracy, sensitivity, specificity, the Matthews correlation coefficient, ROC chart, and area under the curve, this technique optimises two models. The Cleveland data set was utilised. 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 regularised model, C2 is the hyperparameter of the L2 regularised model, and k is the number of features in the chosen subset. The L1-linear SVM model is layered with the L2 linear SVM model in the first experiment, which provides the highest testing accuracy of 91.11% and training accuracy of 84.05%. L1- and L2-linear SVM models with RBF kernels are cascaded in the second experiment. This results in training accuracy of 85.02 and maximum testing accuracy of 92.22%. They have achieved a 3.3% increase in accuracy over traditional SVM models.

The author used different algorithms, it can be said that machine learning is proven to be quite useful in forecasting heart disease, which is one of the most important social issues in the modern world. There may soon be new approaches to make machine learning more beneficial in the field of healthcare as more and more research is being done in this area. Using the provided attributes, the algorithms utilized in this experiment worked pretty well. Finally, it may be concluded that by predicting cardiac disease, machine learning might lessen the harm done to a person's physical and emotional health. [10]

Naive Bayes Classifier and Decision Tree Classification are the two classification models.To compare the accuracy of these two methods, the same dataset is used for both of them. A heart disease patient was predicted by the Decision Tree model with an accuracy level of 91% and by the Naive Bayes classifier with an accuracy level of 87%. Thus, I finish my research by stating that the Decision Tree Classification technique is the greatest and best for handling medical data sets. [11]

Sonam NIkhar and et al [12] propose the heart disease prediction system using several classifier algorithms for the prediction of heart disease is discussed in this work. Naive Bayes classifier and decision tree classifier are the two methodologies; we have determined that the decision tree has a higher accuracy rate than the naive Bayes classifier. We will be working on a Selective Naive Bayes classifier in the future to improve the performance of the classifier. It is generally known that the Naive Bayesian classifier (NB) performs poorly

in some domains and extremely well in others. By deleting pointless and irrelevant features from the dataset and only selecting those that are most informative for the classification job, we hope to improve the performance of the Nave Bayesian classifier.

This work proposes a predictive model for multi-level risk prediction of developing heart failure using C4.5 decision tree classifier employing an open source data set of cardiac illnesses that is available online at UCI machine learning repository [13]. In order to increase the prediction accuracy, the dataset characteristics have been enhanced by the addition of three new features (i.e., risk factors). A performance evaluation has been subjected to 10-fold cross validation. To assess the predictive model, statistical metrics (such as sensitivity, specificity, and accuracy) were used. With 86.5% sensitivity, 95.5% specificity, and 86.53% accuracy, the prediction model outperforms numerous other models. The likelihood of acquiring heart disease would increase as more people participated in various additional risk factors in the future.[14]

3. Drawback of Existing System:

Currently different systems have been designed for predicting heart disease but accuracy is less which needs to improve. The current system which reads the data only and finds the accuracy of algorithm but this is not sufficient. Only finding of algorithm accuracy is not sufficient. So there is a need of a system which can predict heart disease and analyze a risk with the help of certain factors. The problem statement of the proposed system is to design and develop a system which will predict the heart disease based on the machine learning technique using SVM , Linear Regression, ANN algorithm etc. and classify the patient into different categories of risk to provide the risk analysis.

4. Proposed System:

Through performance analysis and exploration of classification algorithms, the suggested system in Fig. 1 forecasts cardiac disease. This study's goal is to accurately determine whether a patient has heart disease. The healthcare provider inputs the data from the patient's health report. The information is incorporated into a model that foretells the likelihood of developing heart disease.

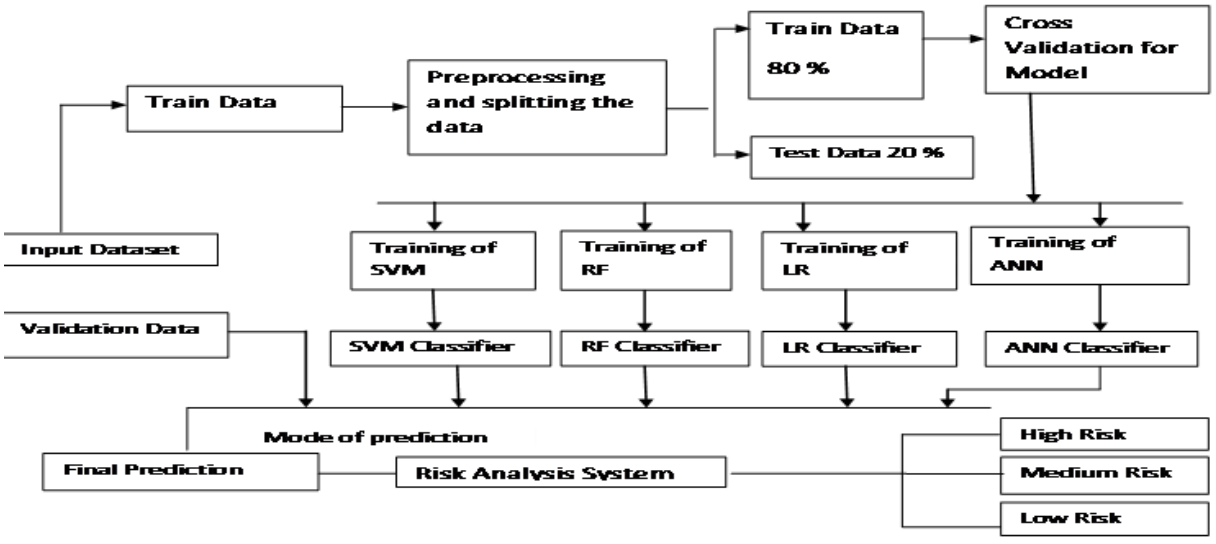

Figure1 Proposed System Architecture

This module preprocesses the dataset after reading the data from the csv file. If there are any missing values in the dataset, the system will fill them using preprocessing. Heart illness Dataset (https://www.kaggle.com/datasets/priyanka841/heart-disease-prediction-uci) was the dataset used. Four separate databases were combined, however only the UCI Cleveland dataset was used. Although there are 76 properties in total in this database, all published experiments only mention employing a subset of 14 features. Therefore, for our research, this system uses the UCI Cleveland dataset that has already been processed and is available on the Kaggle website. The fillna() function loops through your dataset and inserts a specific value into each row that is null. The processing of the dataset is the major focus. The dataset description is listed below. The suggested system extracts the features from this dataset.

1. id (Unique id for each patient)
2. age (Age of the patient in years)
3. origin (place of study)
4. sex (Male/Female)
5. cp chest pain type ([typical angina, atypical angina, non-anginal, asymptomatic])
6. trestbps resting blood pressure (resting blood pressure (in mm Hg on admission to the hospital))
7. chol (serum cholesterol in mg/dl)
8. fbs (if fasting blood sugar > 120 mg/dl)
9. restecg(resting elecardiographic results)
   -- Values: [normal, stt abnormality, lv hypertrophy]
10. thalach: maximum heart rate achieved
11. exang: exercise-induced angina (True/ False)
12. oldpeak: ST depression induced by exercise relative to rest
13. slope: the slope of the peak exercise ST segment
14. ca: number of major vessels (0-3) colored by fluoroscopy
15. thal: [normal; fixed defect; reversible defect]
16. num: the predicted attribute

This papers main objective is to predict the heart disease. In this system module, both the generated dataset and the characteristics from user input are used. The system compares user input to the test dataset and then generates a response based on algorithms. Based on the characteristic criteria, this module categorizes the patient as high risk, medium risk, or low risk. After disease prediction, the risk classification system's approach is to group patients into high-, medium-, and low-risk categories depending on things like age, ancestry, sex, cholesterol, etc. Based on the factors, the system divides the patients into high, medium, and low risk groups. The proposed system will be assessed using many criteria, including as

Precision: It can be defined as the number of correct outputs provided by the model or out of all positive classes that have predicted correctly by the model, how many of them were actually true. It can be calculated using the below formula.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}}$$

Recall: - It is defined as the out of total positive classes, how our model predicted correctly. The recall must be as high as possible.

$$\text{Recall} = \frac{\text{TP}}{TP + FN}$$

Accuracy:

One metric for evaluating arrangement models is accuracy. The little amount of expectations that our model accurately predicted is called exactness. Officially, exactness has a definition that goes along with it

$$Accuracy = \frac{\text{Number of correct Predictions}}{\text{Total No of predictions}}$$

The accuracy is will be calculated with following formula

$$Accuracy = \frac{\text{TP} + \text{TN}}{TP + TN + FP + FN}$$

F1 Score: - The harmonic mean of the model's precision and recall is known as the F-score, which is a method of combining the model's precision and recall. The F-score is frequently used to assess machine learning models, particularly those employed in natural language processing, as well as information retrieval systems like search engines.

$$F1 = 2 * \frac{\text{Precision} * \text{recall}}{Precision + recall}$$

Conclusion:

In this system we have reviewed the large number papers. The literature survey shows that the previous work is done on the dataset only. The author have used the algorithm on the dataset and predicted the algorithm is best. But in real case this is not sufficient so our proposed approach is to predict the heart disease and classify the heart disease into medium risk, high risk and low risk category.

5. References:

1. C. Boukhatem, H. Y. Youssef and A. B. Nassif, "Heart Disease Prediction Using Machine Learning," 2022 Advances in Science and Engineering Technology International Conferences (ASET), 2022, pp. 1-6, doi: 10.1109/ASET53988.2022.9734880.
2. A Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958.
3. S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
4. A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1275-1278, doi: 10.1109/ICECA.2018.8474922.
5. C. Boukhatem, H. Y. Youssef and A. B. Nassif, "Heart Disease Prediction Using Machine Learning," 2022 Advances in Science and Engineering Technology International Conferences (ASET), 2022, pp. 1-6, doi: 10.1109/ASET53988.2022.9734880.
6. M. S. Raja, M. Anurag, C. P. Reddy and N. R. Sirisala, "Machine Learning Based Heart Disease Prediction System," 2021 International Conference on Computer Communication and Informatics (ICCCI), 2021, pp. 1-5, doi: 10.1109/ICCCI50826.2021.9402653.
7. Mr. ChalaBeyene, Prof. Pooja Kamat, “Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Technique”, International Journal of Pure and Applied Mathematics, 2018.
8. Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava, “Effective heart disease prediction using hybrid machine learning techniques” IEEE Access 7 (2019): 81542-81554.
9. 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.
10. V. Sharma, S. Yadav and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020, pp. 177-181, doi: 10.1109/ICACCCN51052.2020.9362842.
11. Sonam Nikhar, A.M. Karandikar."Prediction of Heart Disease Using Machine Learning Algorithms", International Journal of Advanced Engineering, Management and Science(ISSN: 2454-1311),vol.2,no. 6, pp.617-621,2016.

12. Mr.Santhana Krishnan.J, Dr.Geetha.S," Prediction of Heart Disease Using Machine Learning Algorithms",2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), doi:10.1109/ICIICT1.2019.8741465.

13. A. Srivastava and A. k. Singh, "Heart Disease Prediction using Machine Learning," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2022, pp. 2633-2635, doi: 10.1109/ICACITE53722.2022.9823584.

14. A. J. Aljaaf, D. Al-Jumeily, A. J. Hussain, T. Dawson, P. Fergus and M. Al-Jumaily, "Predicting the likelihood of heart failure with a multi level risk assessment using decision tree," 2015 Third International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE), 2015, pp. 101-106, doi: 10.1109/TAEECE.2015.7113608

15. Abhay Kishore1, Ajay Kumar2, Karan Singh3, Maninder Punia4, Yogita Hambir5," Heart Attack Prediction Using Deep Learning", International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 04 | Apr-2018.

16. A.Lakshmanarao, Y.Swathi, P.Sri Sai Sundareswar," Machine Learning Techniques For Heart Disease Prediction", International Journal Of Scientific & Technology Research Volume 8, Issue 11,November 2019.

17. Avinash Golande, Pavan Kumar T," Heart Disease Prediction Using Effective Machine Learning Techniques", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1S4, June 2019.

18. Shah, D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. SN COMPUT. SCI. **1**, 345 (2020). doi./10.1007/s42979-020-00365-y

19. Ramalingam, V V Dandapath, Ayantan - Raja, M PY - 2018/03/19 SP - 684 T1 - Heart disease prediction using machine learning techniques: A survey VL - 7 DO - 10.14419/ijet.v7i2.8.10557 JO - International Journal of Engineering & Technology