PREDICTIVE ANALYTICS FOR CARDIOVASCULAR DISEASE DETECTION USING MACHINE LEARNING

NIRANJAN REDDY[1] SUNIL.B[2] V.AKSHAY[3] G.TEJA[4] D.ARVIND[5]

[1]**Dept.of IT,SNSIT,HYD,** niranjanreddy@sreenidhi.edu.in
[2]**Dept.of IT,SNSIT,HYD,** sunilb@ sreenidhi.edu.in
[3]**Dept.of IT,SNSIT,HYD,** 20311a12l6@sreenidhi.edu.in
[1]**Dept.of IT,SNSIT,HYD,** 21315a1219@sreenidhi.edu.in

Abstract: Cardiovascular diseases (CVDs) have become a leading cause of mortality worldwide, posing a significant health concern. Timely diagnosis is crucial for effective treatment. Leveraging machine learning (ML) techniques offers a promising avenue to analyze large and intricate medical datasets. Researchers have increasingly employed ML algorithms to aid healthcare professionals in diagnosing heart-related ailments. Given the heart's pivotal role in the human body's functioning, accurately predicting occurrences of heart diseases is paramount. The abundance of patient data maintained by medical facilities presents an opportunity for predictive analytics to anticipate future disease occurrences. Various ML techniques, including Artificial Neural Network (ANN), Random Forest, and Support Vector Machine (SVM), have been deployed for heart disease prediction. However, despite advancements in ML, accurate and swift detection methods remain a challenge. Hence, there is a pressing need to develop efficient detection techniques to mitigate the high mortality rates associated with heart diseases. Researchers are intensifying their efforts to harness ML algorithms in software development, facilitating both prediction and diagnosis of heart ailments. This research aims to predict heart disease in patients utilizing ML algorithms.

Keywords: Neural Network, Machine Learning, Supervised learning, Support vector machine, Random forest.

INTRODUCTION

Heart is an important organ of the human body. It pumps blood to every part of our anatomy. If it fails to function correctly, then the brain and various other organs will stop working, and within few minutes, the person will die. Change in lifestyle, work related stress and bad food habits contribute to the increase in the rate of several heart-related diseases. Heart diseases have emerged as one of the most prominent causes of death all around the world. According to World Health Organization, heart related diseases are responsible for taking 17.7 million lives every year, 31% of all global deaths. In India too, heart-related diseases have become the leading cause of mortality. Heart diseases have killed 1.7 million Indians in 2016, according to the 2016 Global Burden of Disease Report, released on September 15,2017. Heart-related diseases increase the spending on health care and also reduce the productivity of an individual. Estimates made by the World Health Organisation (WHO), suggest that India has lost up to $237 billion, from 2005- 2015, due to heart-related or Cardiovascular diseases. Thus, feasible and accurate prediction of heart-related diseases is very important.

Medical organizations, all around the world, collect data on various health-related issues. These data can be exploited using various machine learning techniques to gain useful insights. But the data collected is very massive and, many times, this data can be very noisy. These datasets, which are too overwhelming for human minds to comprehend, can be easily explored using various machine learning techniques. Thus, these algorithms have become very useful, in recent times, to predict the presence or absence of heart-related diseases accurately

The usage of information technology in the health care industry is increasing day by day to aid doctors in decision- making activities. It helps doctors and physicians in disease management, medications, and discovery of patterns and relationships among diagnosis data. Current approaches to predict cardiovascular risk fail to identify many people who would benefit from preventive treatment, while others receive unnecessary intervention. Machine-learning offers an opportunity to improve accuracy by exploiting complex interactions between risk factors. We assessed whether machine-learning can improve cardiovascular risk prediction

I. LITERATURE SURVEY

ChalaBeyene et al[1], recommended Prediction and Analysis of the occurrence of Heart Disease Using Data Mining Techniques. The main objective is to predict the occurrence of heart disease for early automatic diagnosis of the disease within result in a short time. The proposed methodology is also critical in a healthcare organization with experts that have no more knowledge and skill. It uses different medical attributes such as blood sugar and heart rate, age, sex are some of the attributes are included to identify if the person has heart disease or not. Analyses of the dataset are computed using WEKA software. Senthilkumar Mohan et al[2],implemented hybrid machine learning for heart disease prediction. The data set used is Cleveland data set. The first step is data pre-processing step. In this the tuples are removed from the data set which has missed the values. Attributes age and sex from data set are also not used as the authors think that it's personal information and has no impact on predication. The remaining 11 attributes are considered important as they contain vital clinical records. They have proposed their own Hybrid Random Forest Linear Method (HRFLM) which is the combination of Random Forest (RF) and Linear method (LM). In the HRFLM algorithm, the authors have used four algorithms. First algorithm deals with partitioning the input dataset. It is based on a decision tree which is executed for each sample of the dataset. After identifying the feature space, the dataset is split into the leaf nodes. Output of first algorithm is Partition of data set. After that in second algorithm they apply rules to the data set and output here is the classification of data with those rules. In third algorithm features are extracted using Less Error Classifier. This algorithm

deals with finding the minimum and maximum error rate from the classifier. Output of this algorithm is the features with classified attributes. 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 a decision tree and support vector machine. In result as RF and LM are giving better results than other, both the algorithms are put together and new unique algorithm HRFLM is created. The authors suggest further improvement in accuracy by using combination of various machine learning algorithms.

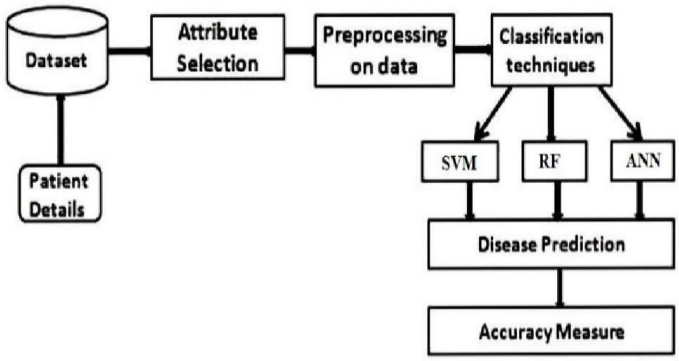
Ali, Liaqat, et al[3],propose a system containing two models based on linear Support Vector Machine (SVM). The first one is called L1 regularized and the second one is called L2 regularized. First model is used for removing unnecessary features by making coefficient of those features zero. The second model is used for prediction. Predication of disease is done in this part. To optimize both models they proposed a hybrid grid search algorithm. This algorithm optimizes two models based on metrics: accuracy, sensitivity, septicity, the Matthews correlation coefficient, ROC chart and area under the curve. They used Cleveland data set. Data splits into 70% training and 30% testing used holdout validation. There are two experiments carried out and each experiment is carried out for various values of C1, C2 and k where C1 is hyperparameter of L1 regularized model, C2 is hyperparameter of L2 regularized model and k is the size of selected subset of features. First experiment is L1-linear SVM model stacked with L2-linear SVM model which is giving maximum testing accuracy of 91.11% and training accuracy of 84.05%. The second experiment is L1-linear SVM model cascaded with L2-linear SVM model with RBF kernel. This is giving maximum testing accuracy of 92.22% and training accuracy of 85.02.They have obtained an improvement in accuracy over conventional SVM models by 3.3%. Singh, Yeshvendra K. et al[4],deal with various supervised machine learning algorithms such as Random Forest, Support Vector Machine, Logistic Regression, Linear Regression, Decision Tree with 3 fold, 5 fold and 10 fold 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 10 fold 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 5 fold 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 S V M in 3 fold cross-validation is 83.17%.For Decision Tree in this paper, the authors have used different number splits and different number of leaf nodes to find the maximum accuracy. With 37 number splits and 6 leaf nodes maximu accuracy is achieved which is 79.12%. When used with cross-validation, accuracy achieved by the decision tree 79.54% with 5 fold. Random forest algorithm used on nonlinear data set gives better results as compared to the decision tree. Random forest is the group of decision tree created by the different root nodes. From this group of decision tree, voting can be done first and then classification can be done from the one getting maximum votes. Authors have used different number splits, different number of tree different number of folds for cross-validation. For random forest, 85.81% accuracy is achieved by 75 Number of trees and 10 number of folds.

II. DATASET

We performed computer simulation on one dataset. Dataset is a Heart dataset. The dataset is available in UCI Machine Learning Repository [10]. Dataset contains 303 samples and 14 input features as well as 1 output feature. The features describe financial, personal, and social feature of loan applicants. The output feature is the decision class which has value 1 for Good credit and 2 for Bad credit. The dataset-1 contains 700 instances shown as Good credit while 300 instances as bad credit. The dat set contains features expressed on nominal, ordinal, or interval scales.



III. PROPOSED SYSTEM

4.1 Random Forest

Random Forest is a supervised machine learning algorithm. This technique can be used for both regression and classification tasks but generally performs better in classification tasks. As the name suggests, Random Forest technique considers multiple decision trees before giving an output. So, it is basically an ensemble of decision trees. This technique is based on

the belief that more number of trees would converge to the right decision. For classification, it uses a voting system and then decides the class whereas in regression it takes the mean of all the outputs of each of the decision trees. It works well with large datasets with high dimensionality

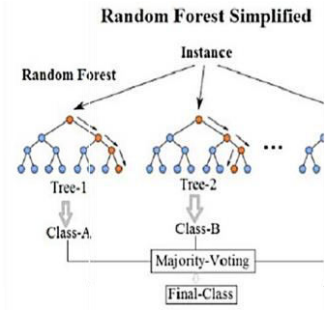


Fig2. Random Forest

4.2 Support Vector Machines (SVMs)

Support vector machines exist in different forms, linear and non-linear. A support vector machine is a supervised classifier. What is usual in this context, two different datasets are involved with SVM, training and a test set. In the ideal situation the classes are linearly separable. In such situation a line can be found, which splits the two classes perfectly. However not only one line splits the dataset perfectly, but a whole bunch of lines do. From these lines the best is selected as the "separating line".

A SVM can make some errors to avoid over-fitting. It tries to minimize the number of errors that will be made. Support vector machines classifiers are applied in many applications. They are very popular in recent research. This popularity is due to the good overall empirical performance. Comparing the naive Bayes and the SVM classifier, the SVM has been applied the most

4.3 Artificial Neural Network

These are used to model/simulate the distribution, functions or mappings among variables as modules of a dynamic system associated with a learning rule or a learning algorithm. The modules here simulate neurons in nervous system and hence ANN collectively refers to the neuron simulators and their synapsis simulating interconnections between these modules in different layers .

Neural Network is built by stacking together multiple neurons in layers to produce a final output. First layer is the input layer and the last is the output layer. All the layers in between is called hidden layers. Each neuron has an activation function. Some of the popular Activation functions are Sigmoid, ReLU, tanh etc. The parameters of the network are the weights and biases of each layer. The goal of the neural network is to learn the network parameters such that the predicted outcome is the same as the ground truth. Back-propagation along loss-function is used to learn the network parameters.

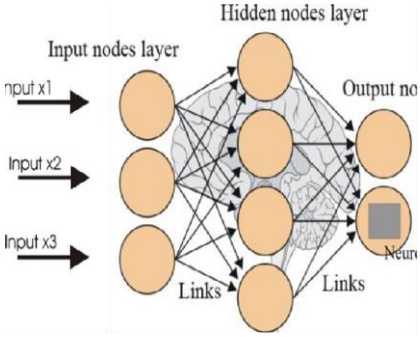


Fig3. Neural Netowrks

IV. SOFTWARE USED

5.1 Python

To collect data a web scraper programmed in Python was used. According to Wikipedia Python's syntax allows programmers to express concepts in fewer lines of codes. Guido van Rossum at CWI in the Netherlands started Python's implementation in December 1989. Python 2.0 was released on October 16th 2000 and Python 3.0 was released December 3rd 2008.

Why use Python for web scraping and not another thing? Python offers a module called 'urllib2', which has suitable functions to open websites and extract information easily. Python is used to program the web scraper that is in charge of collecting the weather data for the model.

V. RESULT AND DISCUSSION

This project aims to know whether the patient has heart disease or not [15].The records in the dataset are divided into the training set and test sets. After preprocessing the data. The data classification technique namely support vector machine, artificial neural network, random forest were applied. The project involved analysis of the heart disease patient dataset with proper data processing. Then, 3 models were trained and tested with maximum scores as follows:

1. Support Vector Classifier: 84.0 %
2. Neural Network: 83.5 %
3. Random Forest Classifier: 80.0 %

VI. CONCLUSION

This project provides the deep insight into machine learning techniques for classification of heart diseases. The role of classifier is crucial in healthcare industry so that the results can be used for predicting the treatment which can be provided to patients. The existing techniques are studied and compared for finding the efficient and accurate systems. Machine learning techniques significantly improves accuracy of cardiovascular risk prediction through which patients can be identified during an early stage of disease and can be benefitted by preventive treatment. It can be concluded that there is a huge scope for machine learning algorithms in predicting cardiovascular diseases or heart related diseases. Each of the above-mentioned algorithms have performed extremely well in some cases but poorly in some other cases.

REFERENCES

[1]. 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.

[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.

[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]. Prerana T H M1, Shivaprakash N C2 , Swetha N3 "Prediction of Heart Disease Using Machine Learning ,Algorithms- Naïve Bayes, Introduction to PAC Algorithm, Comparison of Algorithms and HDPS" International Journal of Science and Engineering Volume 3, Number 2 – 2015 PP: 90-99

[6]. B.L DeekshatuluaPriti Chandra "Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm" International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013.

[7]. Michael W.Berryet.al, Lecture notes in data mining, World Scientific(2006)

[8]. S. Shilaskar and A.Ghatol, "Feature selection for medical diagnosis :Evaluation for cardiovascular diseases," Expert Syst. Appl., vol. 40, no. 10, pp. 4146–4153, Aug. 2013.

[9]. C.-L. Chang and C.-H. Chen, "Applying decision tree and neural network to increase quality of dermatologic diagnosis," Expert Syst. Appl., vol. 36, no. 2, Part 2, pp. 4035–4041, Mar. 2009.

[10]. T. Azar and S. M. El-Metwally, "Decision tree classifiers for automated medical diagnosis," Neural Comput. Appl., vol. 23, no. 7–8, pp. 2387–2403, Dec. 2013. [10] Y. C. T. Bo Jin, "Support vector machines with genetic fuzzy feature transformation for biomedical data classification.," Inf Sci, vol. 177, no. 2, pp. 476–489, 2007.

[11]. N. Esfandiari, M. R. Babavalian, A.-M. E. Moghadam, and V. K. Tabar, "Knowledge discovery in medicine: Current issue and future trend," Expert Syst. Appl., vol. 41, no. 9, pp. 4434–4463, Jul. 2014.

[12]. E. Hassanien and T. Kim, "Breast cancer MRI diagnosis approach using support vector machine and pulse coupled neural networks," J. Appl. Log., vol. 10, no. 4, pp. 277–284, Dec. 2012.

[13]. Sanjay Kumar Sen 1, Dr. Sujata Dash 21Asst. Professor, Orissa Engineering College, Bhubaneswar, Odisha – India.

[14]. Domingos P and Pazzani M. "Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier", in Proceedings of the 13th Conference on Machine Learning, Bari, Italy, pp 105-112,1996.

[15]. Elkan C. "Naive Bayesian Learning, Technical Report CS97-557", Department of Computer Science and Engineering, University of California, San Diego, USA, 1997.

[16]. B.L Deekshatulua Priti Chandra "Reader, PG Dept. Of Computer Application North Orissa University, Baripada, Odisha – India. Empirical Evaluation of Classifiers Performance Using Data Mining Algorithm"