A Survey on Machine Learning Algorithms for Cardiovascular Diseases Prediction

Mrs J Amutha[1], Dr K Ruba Soundar[2], Mrs M Piramu[3], Dr.K.Murugesan[4]

[1] P.G. Scholor, [2]Professor & Head / CSE, [3]Assistant Professor/ CSE, [4]Professor/ECE
[1, 2, 3]P.S.R. Engineering College, Sivakasi – 626140
[4]Easwari Engineering College, Chennai

Abstract: **Heart is the most important part in all living organisms. Cardiovascular diseases or heart related diseases are at its peak in today's world. Cardiovascular diseases prediction in a living being is a critical challenge analysis in the medical field. Machine learning algorithms are used in effective decision making, perfection and correctness because of little fatigue problem. In this work a survey has been done among various machine learning algorithms such as SVM, Decision Tree, K-Nearest Neighbor (KNN), Artificial Neural Networks (ANN) and Random Forest with linear model to predict out of this heart disease. In performance level 92% is achieved through Support Vector Machine prediction model for heart diseases. Support Vector Machine method aims at finding large amount of feature by applying machine learning algorithm to improve the accuracy in the prediction of cardiovascular diseases.**

Keywords: **Heart Disease classification, Support Vector Machine, Decision Tree, K-Nearest Neighbor, Artificial Neural Networks, Random Forest.**

1. INTRODUCTION

Heart disease is a dangerous health problem and several people have been suffered by this disease around the world. The Heart disease occurs with common symptoms of chest pain, chest tightness, breath shortness, physical body weakness and, feet are swollen. We are using this project as a mean to come across an efficient technique using for the recognition of heart disease, because the current diagnosis techniques of cardiovascular disease are not effective in early time identification. In Model world, everyone is running out of time. So it is important to diagnose and give proper treatments to save more people around the world. The traditional treatment for diagnosis of heart disease is done by the analysis of the medical history of the patient; consist only of the physical examination report and analysis of related symptoms by a physician. So, that the result of obtained diagnosis methods are not accurate in identifying the patient of heart disease. Furthermore, it is computation and expensively difficult to analyze.

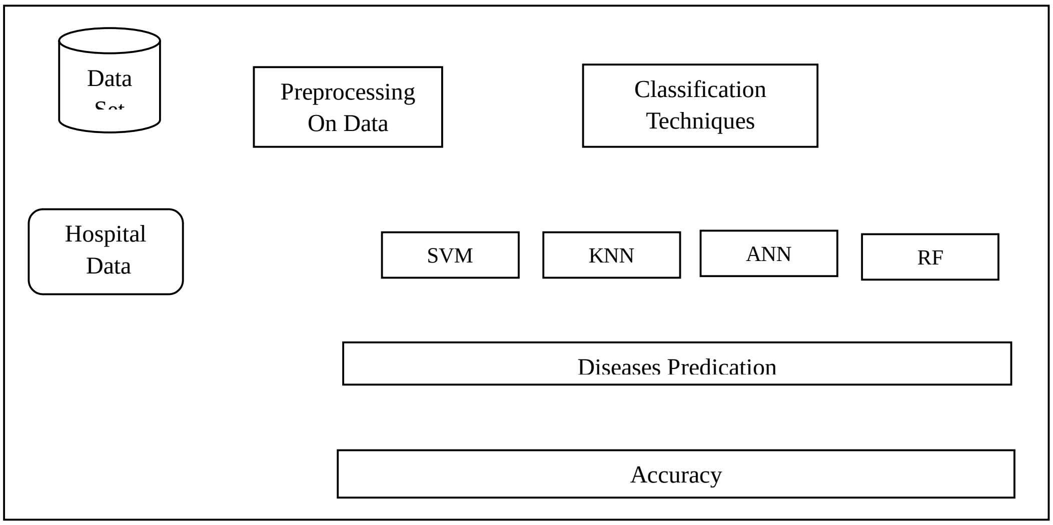
Specialist decision system based on machine learning method and features are used to effectively diagnosis the heart disease as a result, except for decreasing the ratio of death. The machines learning using predictive model need proper data for training and testing. However, the performance of machine learning method can be used if balanced dataset is used for training and testing of the model. Furthermore, the model predictive capabilities can improve by using proper and related features from the data. Therefore, data balancing and feature selection is significantly important for model performance improvement. The identification of disease in most cases depends on a complex combination and huge volume of medical data. The algorithms have been effective assisting in making decisions and predictions from the traditional machine learning algorithms that aims in improving the accuracy of heart disease and prediction has been applied. In diseases, accurate diagnosis is primary. Although, this method is used to predict and diagnosis.

2. LITERATURE REVIEW

In literature various machine learning based diagnosis techniques have been proposed by researchers to diagnos heart diseases. We are using research paper to study some current machine learning based diagnosis techniques in order to explain the important of the proposed work. [1] Naïve Bayes classifier is based on Bayes theorem. It is based on the Probabilities of the class attribute based on the product of prior probability of class attribute and the possible conditional probabilities involving the values of a single attribute. Logistic Regression (LR) is one of the classification method used to diagnose heart diseases. LR is similar to the Linear Regression but it is appropriate to use when the dependent variable is binary means only having 2 possible outcomes. The outcome is determined by set of independent variables call as predictor or explanatory variable. Like all other regression analyses, this method is also a predictive analysis. [2] so that it classifies heart diseases correctly without errors. Hence they intend to propose a model with highest accuracy, precision and with minimum root mean square error. The KNN algorithm has low accuracy in this scenario but it has good results for large number of samples. Considering these statistics, we have concluded that most efficient amongst all the algorithms studies are SVM, ANN and KNN for heart disease prediction. [3] The back propagation algorithm is a technique used in developing multilayer perceptron (MLP) neural networks in a supervised manner. The BP algorithm is also called as error back propagation algorithm which is based on the error correction learning rule. In forward pass an activity pattern is applied to the input nodes and it propagates through the network layer by layer. As a result, a set of outputs is produced as the actual response of the network. The weights at the functional points of the network are fixed in the forward pass. For the duration of the backward pass, the synaptic weights are all adjusted in accordance with an error-correction rule. [4] A method is based on neural network and genetic algorithm is proposed for the prediction of heart syndrome by exploiting main detrimental features. The proposed technique predicts the threat of heart syndrome by precision of 89%. Therefore, by using this technique a smart system is able to predict the syndrome.

3. PROPOSED SYSTEM

The subsequent diagram represents this heart disease prediction system Architecture. Processing of system starts with the data collection. The data-aspirant repository dataset is used in this work.



Data-aspirant repository dataset in training dataset consists of 14 attributes of data. All the attributes consist of numeric values. The initial 13 variables will be used for predicting 14th variables. The decision making of variable is at index 14.

3.1 Dataset Selection

Attribute of dataset is a property of dataset which is used by the system for heart disease prediction. The performance of machine learning model can be increased, if balanced dataset is used for training and testing. The initial step of the predication system is, the data collection and deciding about the training and testing dataset. We have used 70% of training dataset and 30% of dataset as testing dataset. Furthermore, the model predictive capabilities can be improved by using proper and related features from the data. Therefore, data balancing and feature selection are significantly important for model performance improvement.

3.2 Pre-Processing

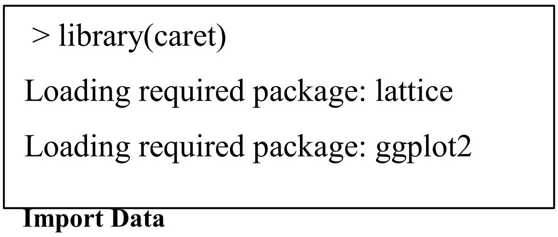
The data-spirant dataset is loaded and the data becomes ready for pre processing. The subset of 13 attributes take in age, sex, cp, treetops, chol, restecg, thalach, exang, olpeak, slope, ca, that is decision making from the pre-processed data set of heart disease. Support Vector Machine, K-Nearest Neighbors, Artificial Neural Network, Decision Tree, Random Forest, and Linear Model are used to develop the classification model. Confusion matrix method using the performed of model.

3.3 Classification

The different types of various features selection and modeling keep on repeating combinations of attributes. Training data should have used as in regression operation for predicting and comparing the output. Training data is loaded into classification operation and uses SVM, KNN, ANN, RF and Decision tree algorithms are classification that match the parameters with dataset and reduced the complexity.

3.4 Support Vector Machine

CART stands for Classification and Regression Trees methodology. R has a package to access the REST API called CARET. Open R-Studio generate a new R Script. Once you have done this, you will need to install and load the CARET.



```
> library(caret)
Loading required package: lattice
Loading required package: ggplot2
```

3.5 Import Data

Open up R-Studio, in the Files tab, click Upload, and choose your csv file. Click on the Workspace tab, on "*Import Dataset" -> "From heart_df*". A document program will open up, find the Csv record and click open. Click "Import".

3.6 Training Data

Data slicing splits method for data into train dataset and test dataset. Training data set can be used specifically for our model building by using prediction of heart disease. Even during consistency, we should not standardize our test set. The Following Figure 4.3 Represents for Summary. The caret package inbuilt a method of *createDataPartition()* for partitioning our data into train dataset and test dataset. We have been on passing three parameter values. The "y" parameter values take in variable giving into dataset using the separated. So that is in our case, decision making variable is at V14, so we are passing in dataset of sri_heart$V14. The last on the attributes of the value is $V14 for "p" parameters hold on value range upto 0 to 1. It is shown on dataset of the split. We are using the value of p=0.8. So that is dataset split range in 80:20 ratios.

4. Machine Learning Algorithms

4.1 Support Vector Machine

The initial steps of implementation, the training dataset are used to learn the ML model. It helps to divide the data into k equal subsets and to give a chance for each subset to be a part of training and testing phase. The working of cross validation operator considers as an efficient, as it repeats the learning phase k times, where every time the testing data selection is different from previous. Finally, it repeats the experiment k times and uses the average results. For training data SVM classifier, "*svmLinear*". method using than parameter passing through into the *train*() method. We are passing our decision making variable V14. The "V14~." denotes a formula for using all attributes in our classifier and V14 as the target variable.

4.2 K-Nearest Neighbor

It works on the basis of distance between the location of data and on the basis of this distinct data that are classified with each other. All the other groups of data are called neighbor of each other and the number of neighbor are decided by the user which, Play very crucial role in analysis of the dataset. Each cluster represented in two dimensional space whose coordinates are represented as (Xi,Yi) where Xi is the x-axis, Y represent y axis and i = 1,2,3,….n.

4.3 Artificial Neural Network

Artificial Neurons or Processing Elements (PE) are highly basic models of biological neurons. Each one neuron in ANN receives a number of inputs and an output which can be connected to other artificial neurons. Artificial neural networks are closely interconnected networks of Processing Elements; to accurate the strength of the connection between the units in response to externally supplied data. The network has 2 binary inputs I0 and I1.. One binary output Y. Connection Weights are W0 and W1..

4.4 Random Forest

Random forest model using in caret package. Now, our model is trained with K value as dataset. We are go to check the predict classes for our test dataset. Here, the two parameters using of *predict*() method. The first and second parameters are our training and testing model, held our testing data frame. Finallyss, the list of positive case return in method of predict(), We are saving it in a test_pred variable.

Training Set	Accuracy Result Using SVM	Accuracy Result Using KNN	Accuracy Result Using RF	Accuracy Result Using ANN
50%	85.93	82.96	83.70	81.48
55%	87.60	84.30	85.12	85.12
60%	87.96	85.19	86.11	85.19
65%	88.30	88.30	90.43	88.3
70%	90.12	86.42	90.12	80.25
75%	91.04	86.57	89.55	82.09
80%	92.59	83.33	90.74	81.48
85%	87.50	77.50	85.19	80.00
90%	85.19	76.92	85.00	77.78
95%	76.92	70.37	76.92	76.92

Table 1: Accuracy Result

5. Result

We have implements in SVM method of confusion matrix using in accuracy results. It shows that SVM method accuracy for testing dataset in 92.59%. A survey has been done among various range of training Data value from 50% To 95% are accuracy result for prediction of Machine Algorithm using heart diseases. Heart diseases prediction SVM, ANN, KNN and RF algorithm using in survey has been done among Training Data Range of Value in start 50% To 95%. The following Table 1 represents the Heart diseases prediction using Support Vector Machine.

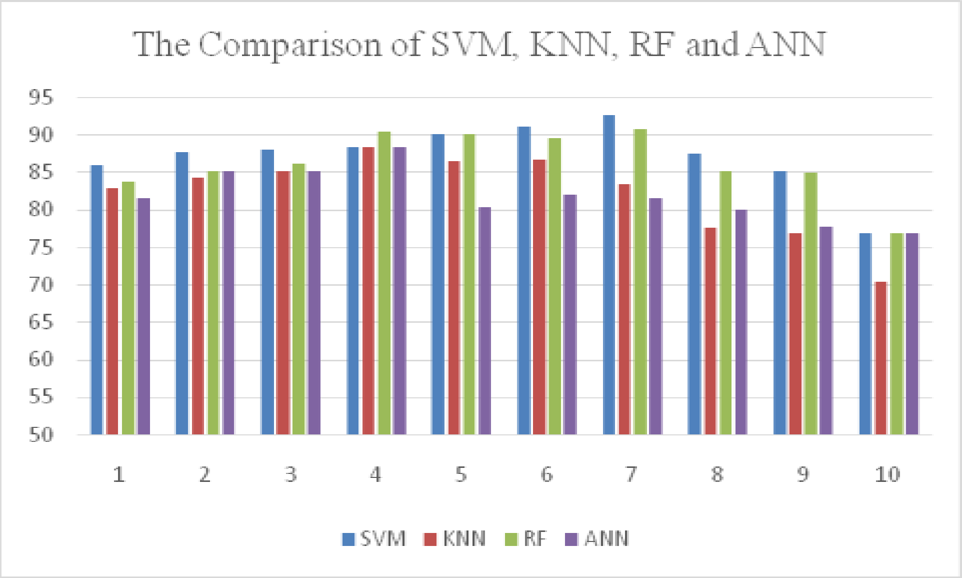


Figure 2: Comparison chart

Above Figure 2 represents the comparison chart of SVM, KNN, RF and ANN Algorithms Using Heart Diseases Prediction.

6. CONCLUSION

Identifying the processing of raw healthcare data of heart information will help in the long term saving of human lives and early detection of abnormalities in heart conditions. Machine learning algorithms techniques were used in this work to process raw data and provide new project to heart disease. We have cardiovascular diseases prediction is essential of medical field. However, the mortality rate can be high level has been controlled if the disease is detected at the early stages and preventive processes are adopted as soon as possible results. Further extension of this studying the method of using to through the surveys to real-world datasets instead of just theoretical approaches and simulations, our project proved to be quite accurate in the prediction of heart disease, the future course of this research analytics after applied. It can be performed with various method in mixtures of machine learning techniques for better prediction techniques. Furthermore, new feature selection methods can be developed to get a broader perception of the important features to increase the performance of heart disease prediction.

REFERENCES

[1] Ruba Soundar Kathavarayan, Murugesan Karuppasamy, "Identification of untrained facial image in combined global and local preserving feature space", 2010, International Journal of Biometrics and Bioinformatics (IJBB)

[2] Salma Banu N K, Suma Swamy, "Prediction of Heart Disease at early stage using Data Mining and Big Data Analytics: A Survey", 2016, International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)

[3] V Krishnaiah, G Narsimha, "Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review", International Journal of Computer Applications (0975-8887), Volume 136, No. 2, February 2016

[4] Tahira Mahboob, Rida Irfan, Bazelah Ghaffar, "Evaluating Ensemble Prediction of Coronary Heart Disease using Receiver Operating Characteristics", International Journal of Computer Applications, March 2016

[5] C Sowmiya, P Sumitra, "Analytical Study of Heart Disease Diagnosis Using Classification Techniques", 2017, IEEE International Conference On Intelligent Techniques In Control, Optimization And Signal Processing

[6] Ruba Soundar Kathavarayan, Lekshmi Kalinathan "Segmentation of hepatocellular carcinoma and dysplastic liver tumors in histopathology images using area based adaptive expectation maximization", 2018, Multimedia Tools and Applications