Different Machine Learning Models Based Heart Disease Prediction

Arunpradeep N., G. Niranjana

Abstract: ***Heart related disease is one of the crucial reasons for high amount of people's death in the whole countries and it's considered as life forbidding disorder, in addition to that this effect takes place in whole earth. Heart disease will affect the early stage of age peoples also. Thus, heart related disease creates the more challenges to people living and identify the causes and detection step is more important in nowadays. So, we need to develop of automatic system with more accurate and reliable for early detection of heart disease. For this reason, various machine learning models are developed to predict heart related disease; different medical data package is processed to automatic analysis with get more accuracy. In this paper, we discuss the available machine learning models such as KNN, SVM, DT and RF algorithms for prognosis of heart disease with high certitude, precision and recall.***

Keywords: Heart Disease prediction, KNN, SVM, Decision Tree, Random Forest, machine learning.

I. INTRODUCTION

Heart is a crucial part in the human body. Based on heart, blood is circulated to whole body. Heart working function is not give correct response means our organs such as brain and different organs are suddenly stop the working within few minutes then it will make human die. The different heart disease is developed by work stress, bad food habit and lifestyle change etc. Nowadays, most of the death cases occurring by the many people are affected by the heart disease [1]. In India, 1.7 million peoples are death by the heart disease; this information is released from the 2016 GBD (Global Burden of Disease) Report on Sep 15, 2017. Chronic disease will make the life care spend and decrease the individual work rate. World Health Organization made some estimation, based on the estimation WHO suggests India will lose $2370 billion from 2005-15 due to cardiac related disease [2].

Thus, early precision of cardiac related disease is very extensive in nowadays. Symptoms of heart disease is frequent back pain, breath tiny, pain in jaw, pain in neck, pain in chest, pain of arm and shoulders; this may be vary every person [3]. This is not effective way for finding the heart disease so visual appearance vise heart disease finding is implementing that means using X-ray test and angiography test. Above test result gives a good result but test expensive and medical equipment availability is less in some rural place. This problem is solved by using various machine leaning models to predict the heart related disease for reducing the human die. It is easy and effectively prediction model for heart disease without any tests.

Revised Manuscript Received on February 04, 2020.

Arunpradeep.N, Master Of Technology, Department of Internet of Things, SRM Institute of Science and Technology, KTR Campus, Chennai, Tamilnadu, India.

Dr. G. Niranjana, Associate Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, KTR Campus, Chennai, Tamilnadu, India

Based on the statistical factors of human i.e. age, blood pressure, cholesterol level and etc, this data mining approaches are implemented.

This paper tested with Kaggle data package and their results and the results were compared using the four machine learning models i.e. KNN, SVM, DT and RF for predicting the heart disease; finally presents the quantitative evaluation i.e. accuracy, precision and recall for analyzing the different models.

II. RELATED WORKS

In [4], Tan et al presented a hybrid approach that means combining of two machine-learning models such as SVM-support vector machine and genetic algorithm for classification. They used four different data packages like iris, diabetic, breast cancer, heart disease and hepatitis for making experiment; these data package was collected from the UCI repository. After executing of hybrid method to heart disease classification, it obtained the accuracy is 84.07 %, 78.26 % of accuracy is obtained for diabetes, 76.20 % of accuracy for breast cancer and for hepatitis disease the accuracy is 86.112 %.

In [5], Ottom et al presented a novel method for detection and monitoring of coronary artery disease. For this work, Cleveland heart data was collected from the UCI. 303 cases and 76 attribute used in data package; only 13 attribute is used for detection. They proposed three ML algorithms such as, SVM and Functional Tree for detection process. Finally, detection accuracy 88.3 % is obtained by SVM, achieved 83.8 % accuracy by and 81.5 % accuracy by FT.

In [6], Parthiban et al proposed an automatic learning model for classification of heart disease. They used two machine learning models such as KNN and SVM for this work. For this work, 500 patient's data packages are used. The 500 patient's data package was obtained from the Chennai Research Institute; it had 142 disease patients and 358 normal patients records. At last, they achieved 74 % accuracy by KNN and 94.60 % accuracy by SVM.

III. COMPARISON OF EXISTING ALGORITHMS

This paper presents the four different machine algorithms i.e. SVM, KNN, DT and RF for prediction or classification of heart disease based on medical data attributes. Fig.1 shows the flow of process for heart disease prediction follow as,

i. Cardiac infection data package is collected from the UC Irvine ML repository.
ii. The detailed packages has some NaN values; it is needed to replace by numerical values. This process is done in preprocessing step.
iii. After that, our data package is split into training and testing data for validation.

iv. Finally training data are trained by different algorithms and testing data are classified based on trained model.

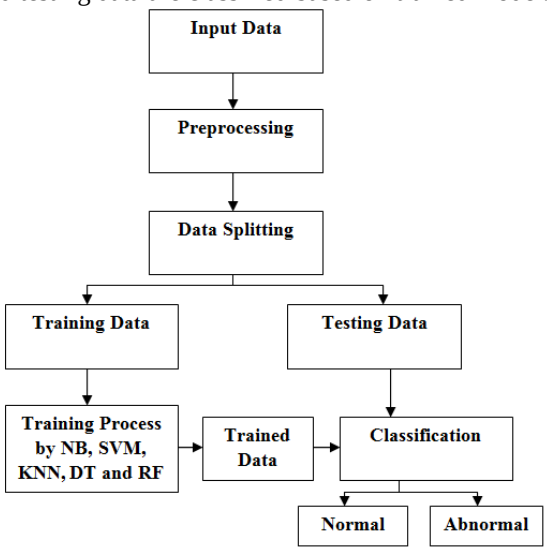

Fig.1 Overall Process for Heart Disease Prediction

1. Dataset Preparation

The cardiac disease data package is obtained from the UCI machine learning repository. In UCI machine learning repository, various domains data package is available. Based on the repository, many researches and academic paper are developed. This UCI is developed by David Aha in 1987and students are follows the UC Irvine. Cardiac infection data package data is collect from the four institutions [7] such as,

1. Cleveland Clinic Foundation.
2. Hungarian Institute of Cardiology, Budapest.
3. V.A. Medical Centre, Long Beach, CA.
4. University Hospital, Zurich, Switzerland.

Attribute	Representation	Attribute Information	Definition
Age	Age	NUMBER	Individual age (19 to 78)
Sex	Sex	NUMBER	Based on category {0=Female, 1=Male}
Chest twinge Type	Cp	NUMBER	Type of chest twinge (1: typical angina, 2:atypical angina, 3:non-anginal pain, 4:asymptomatic)
Rest Blood Pressure	Trestbps	NUMBER	Resting heart rate in mm Hg [94 200]
Serum Cholestoral	Chol	NUMBER	high-density lipoproteins in mg/dl [126,564]
Fasting Blood Sugar	Fbs	NUMBER	FBS > 120 mg/dl (0= False, 1=True)
Res-Electo cardio grapic	Restecg	NUMBER	Resting ECG results (0: normal, 1: ST-T wave abnormality, 2: LV hypertrophy)
Max Heart Rate	Thalach	NUMBER	extreme cardiac rate noted[71, 202]
Exercise Induced	Exang	NUMBER	Exercise induced angina (0: No, 1: Yes)
Old peak	Oldpeak	reaL	ST depression induced by exercise relative to rest[0.0, 62.0
Slope	Slope	NUMBER	Slope of the peak exercise ST segment (1: up-sloping, 2: flat, 3: down-sloping)
Major Vessels	Ca	NUMBER	Four greater vessels colored by fluoroscopy (values 0 -3)
Thal	Thal	NUMBER	deficiency types: value 3: normal, 6: fixed deficiency, 7: irreversible deficiency
Class	Class	NUMBER	Diagnosis of cardiac disease (1: Unhealthy, 2: Healthy)

2. Data Preprocessing

The data package contains original attributes and NaN values. In programming, we cannot the process the NaN values so these values are transform into another value i.e. numerical value. NaN values are replaced by the mean value of columns.

3. Data Splitting

The splitting step is used for creating the training and testing data to analyzing process. In that, our whole data package is divide into training and testing data; use 80% of data for training and 20% of data for testing.

4. Classification

In classification, splited training and testing data are valuated based on machine learning models. First, training data was trained by using four different machine learning models such as , KNN, SVM, DT and RF. After that testing data are validated based on trained data with high classification accuracy rate. Four different algorithms are explained details given as follows,

A. Support Vector Machine

One of the most popular supervised machine learning models is support vector machine which is used for classification and prediction. SVM finding the hyper-plane in the feature space that making variation between the labels or classes for classification.

An SVM model represents the training data points as points in the feature space, mapped in such a way that points belonging to separate classes are segregated by a margin as wide as possible.
The test data points are then mapped into that same space and are classified based on which side of the margin they fall.

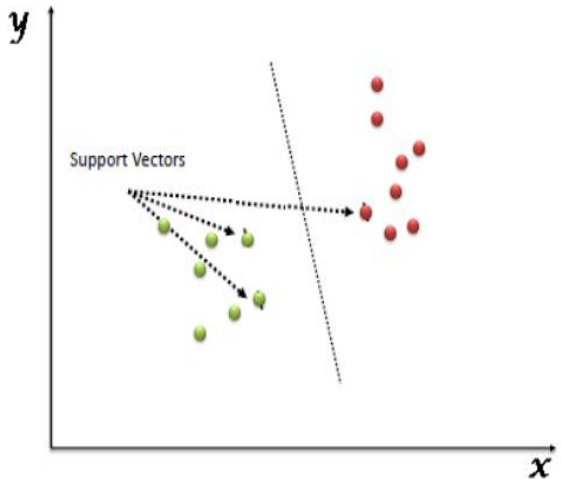

Fig.2 Support Vector Machine

B. K-Nearest Neighbour

Hodges et al. in 1951 presented a pattern classification based on nonparametric model which is called as K-Nearest Neighbour rule [8]. KNN is the one of the basic and simple but very intelligent classification algorithm. This algorithm did not create any assumptions for data and normally KNN used for classification process when no knowledge about the data distribution. The working of KNN is finding the k nearest data for the test data in the training set of data. K nearest data finding in training set is based on Euclidean Distance.

C. Decision Tree

Decision Tree is one of the supervised learning algorithms. Mostly classification problems are solved by using decision tree. It easily performs with continuous and categorical attributes. Based on significant predictors, the population is dividing into two or more similar set in DT. The first step of DT is calculating of entropy for each and every attribute. Next, based on the variables/ predictors the data package is splited with high information gain or less entropy. Above two steps are followed to remaining attributes.

$$Entropy\ (E) = \sum_{k=1}^{l} -q_k \log_2 q_k \qquad (2)$$

where l is refers to response variable modules count, q_k is the ratio of the count of the k[th] class procedures to a whole count of models.

$$Gain\ (E, G) = Entropy\ (E) - \sum_{v1\epsilon\ Values\ (G)} \frac{|E_{v1}|}{E}\ Entropy\ (E_{v1})$$

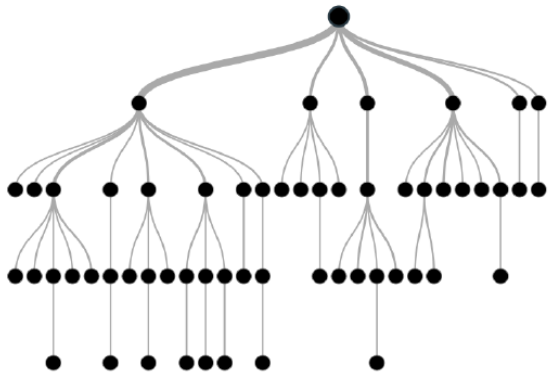

Fig.3. Decision Tree Structure

D. Random Forest

In supervised machine learning models, Random Forest is also one popular model. RF is worked for both classification and regression but it gives only better result to classification. In random forest, before getting the output or result many decision trees used. So, random forest is refers the combining of many decision trees. High number of trees would make the good result in RF. Voting system is used for classification and then decides the class whereas in regression it makes the mean prediction for all the outputs of each and every decision trees. Random forest easily and effectively worked with more number of data package with high dimensionality.

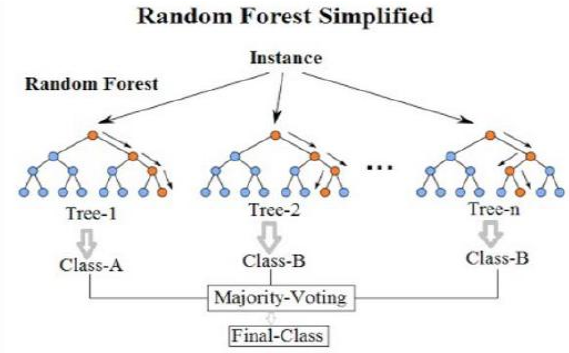

Fig.4. RF Working Structure

IV. EXPERIMENTAL RESULTS

In this part, we show the classified result from various prediction models. We used different parameters for make comparison with different models; the parameters i.e. Accuracy, Precision and Recall. Table 1 shows the comparison with four models,

Table 1: Quantitative Evaluation with different models

Method	SVM	KNN	DT	RF
Accuracy	0.865	0.753	0.745	0.847
Precision	0.843	0.785	0.776	0.825
Recall	0.866	0.753	0.745	0.847

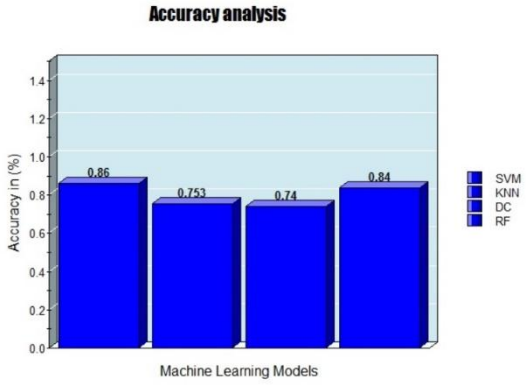

Fig.5 Accuracy Analysis

Fig.5 shows the accuracy analysis with different machine learning models.

PROPOSED SYSTEM:

To overcome all the basic and essential medical problems and some convoluted issues like heart (ECG checking) that occurred in an e-entry which prompts the town patient to be associated with the Specialist in the Emergency clinic. Do an Exploration on, medicinal sensors, which utilized to gather physiological information from patients and transmit it to a doctor. Few parameters tested using boards, smart watch, sensors. This undertaking plans to determine the job of body sensor organizes in medication to limit the requirement for parental figures and help the incessantly sick and older individuals. Particularly the town individuals to get exhorts from separate Specialists and to carry on with an autonomous life, other than giving individuals quality consideration and cloud database

SYSTEM ARCHITECTURE:

From the architecture it describes about the project related to IOMT with cloud to discuss about cloud storage, dataset, boards, analysis various parameters which implant to patient. Wearable device measures the test to patients, create a dataset, and upload to cloud. By using Arduino, Nordic semiconductor board to check some parameters with the help of some IDE software's and embedded studio software's to write and code and compile it. Create a dataset and upload to cloud. All cloud storage data received through no sql dbms tool using Cassandra for analysis in it. The analysis taken by prediction method for Naïve Bayes algorithm. After the analysis, report generated and submitted to doctor to predict the cardiac arrest to patient. This may produce some reliability and accuracy compare to other methods.

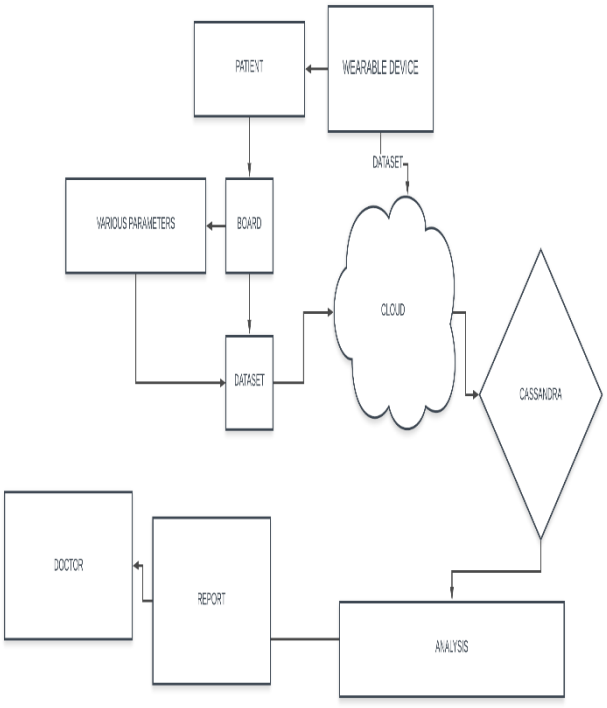

Fig .6 System Architecture

V. CONCLUSION

In this article we studied and analyzed the various machine learning algorithms i.e. KNN, SVM, DT and RF algorithms for predicting of cardiac related disease. Prediction of cardiac disease followed the step as information preparation in that details was taken from open source database; data preprocessing for replacing of NaN value; data splitting in that whole data package split into training and testing; finally different models based training data are trained and testing data are validated by the trained data in classification step. At last, above different machine learning models are compared in terms of finding of quantitative evaluation metrics such as accuracy, precision and recall.

REFERENCES

1. Ramadoss and Shah B et al. A. Responding to the threat of chronic diseases in India. Lancet. 2005
2. Global Atlas on Cardiovascular Disease Prevention and Control. Geneva, Switzerland: World Health Organization, 2011.
3. NagannaChetty, Kunwar Singh Vaisla, NagammaPatil, An Improved Method for Disease Prediction using Fuzzy Approach, ACCE 2015.
4. K. C. Tan, E. J. Teoh, Q. Yu, and K. C. Goh, A hybrid evolutionary algorithm for attribute selection in data mining, Expert Systems with Applications, 2009.
5. A. F. Otoom, E. E. Abdallah, Y. Kilani, A. Kefaye, and M. Ashour, Effective diagnosis and monitoring of heart disease, International Journal of Software Engineering and Its Applications, 2015.
6. G. Parthiban and S. K. Srivatsa, Applying machine learning methods in diagnosing heart disease for diabetic patients, International Journal of Applied Information Systems, 2012.
7. M. Lichman, UCI Machine Learning Repository. 2013.
8. J. Hodges et al. Discriminatory analysis, nonparametric discrimination: Consistency properties, 1981.

9. S.Rajathi and Dr.G.Radhamani et al. Prediction and Analysis of Rheumatic Heart Disease using kNN Classification with ACO , 2016.
10. Puneet Bansal and Ridhi Saini et al. Classification of heart diseases from ECG signals using wavelet transform and kNN classifier, International Conference on Computing, Communication and Automation (ICCCA2015).
11. Simge EKIZ and Pakize Erdogmus et al. Comparitive Study of heart Disease Classification, 2017
12. Renu Chauhan, Pinki Bajaj, Kavita Choudhary and Yogita Gigras et al. Framework to Predict Health Diseases Using Selection Mechanism, 2015 2nd International Conference on Computing for Sustainable Global Development (INDIA Com).
13. M.A.JABBAR , B.L Deekshatulu and Priti Chndra et al. Alternating decision trees for early diagnosis of heart disease, Proceedings of International Conference on Circuits, Communication, Control and Computing (I4C 2014).
14. Amir Hussain, Peipei Yang, Mufti Mahmud and Jan Karasek et al. A Novel Cardiovascular Decision Support Framework for effective clinical Risk Assessment. IEEE 2014.
15. Quazi Abidur Rahman, Larisa G. Tereshchenko, Matthew Kongkatong, Theodore Abraham, M. Roselle Abraham, and Hagit Shatkay et al. Utilizing ECG-based Heartbeat Classification for Hypertrophic Cardiomyopathy Identification, IEEE 2015.

AUTHORS PROFILE

Arunpradeep N, is currently pursuing Master's degree in Internet of Things at SRM Institute of Science and Technology Kattankulathur, Chennai, India. Pursued his Bachelor's degree in KSR College of Technology, Tiruchengode, Namakkal, India. His area of interest are Cloud Computing, Internet of Things, Network Security, He published Two journal in Composite Material, in IJAER

Dr G. Niranjana, is associate professor of Computer Science and Engineering from SRM Institute of Science and Technology Kattankulathur, Chennai, India. She earned her Ph.D. in Image Processing from SRM Institute of Science and Technology Kattankulathur, Chennai, India. She has over eleven year of experience in Teaching and Research Her area of interest are Image Processing, Pattern Recognition, Networks and Data Structures.