Prediction of Cardiovascular Disease Using Machine Learning Algorithm

Anand.P[1], Abinaya.B[2], Akshaya Priya.R.K[3], Chandraleka.P[4], Harni.R[5]
[1]*Assistant professor, Information Technology, Saranathan College of Engineering, Trichy*
[2,3,4,5] *Information Technology, Saranathan College of Engineering, Trichy*

***Abstract*- The proposed work suggests the design of a health care system that provides various services to monitor the patients using wireless technology. It is an intelligent Remote Patient monitoring system which is integrating patient monitoring with various sensitive parameters. The system is said to be an intelligent system because of its diagnosis capability, timely alert for medication etc. The current statistics shows that heart disease is the leading cause of death and which shows the importance of the technology to provide a solution for reducing the cardiac arrest rate. Apart from that the proposed work compares different algorithms and proposes the usage of Random Forest algorithm for heart disease prediction.**

***Index terms*- Heart Disease, Dataset, Random Forest Tree algorithm**

I.INTRODUCTION

The World Health Organization (WHO) classifies cardiovascular diseases as the number one cause of death globally. In total, 17.9 million people died from cardiovascular diseases in 2016, representing 31% of all global deaths. Cardiovascular diseases are disorders of the heart and blood vessels. Four out of five cardiovascular diseases deaths are due to heart attacks and strokes. Individuals at risk of cardiovascular diseases may demonstrate raised blood pressure, be overweight or obese among the adult population, cardiovascular diseases are the main health problem in general. It mainly affects the heart and the arteries of the brain, heart and legs. Therefore, the lack of blood supply not only damages the heart, but also the legs and brain, which can lead to health disorders prompting a risk of heart attacks, thrombosis or rupturing of blood vessels, among others

The main risk factors were defined in the Framingham Heart study published in 1952 and are listed as follows

- Age
- Gender
- Body Mass Index
- Smoking Condition
- Homocysteine
- Reactive C-Protein
- Fibrinogen
- Previous familiar cases
- Diet
- Cholesterol HDL Triglycerides Lipoprotein
- Sedentary Condition
- Glucose Tolerance and Metabolic System
- High blood pressure

The WHO defines unhealthy diet, physical inactivity, tobacco use and excessive use of alcohol as the most important behavioral risk factors for heart disease. These “intermediate risk factors” can be measured in primary care facilities and indicate an increased risk of developing a heart attack and other complications Some of this information can be provided immediately, while in the other cases tests need to be done. These can include blood tests or an electrocardiogram. An electrocardiogram is a diagnostic tool that is routinely used to measure and record different electrical potentials of the heart. Willem Einthoven developed the ECG method in the early 1900s, and while it is a relatively simple test to perform, the interpretation of ECG tracing requires a significant amount of training

The P wave of the ECG looks at the atria. The QRS complex looks at the ventricles and the T wave evaluates the recovery stage of the ventricles while they are refilling with blood. The ST slope and ST depression, induced by exercise, is part of the database which is used for the method in this paper. Generally, many health care organizations are facing a major challenge to offer high quality provisions, like diagnosing patients correctly and administrating treatment at reasonable costs. Machine learning techniques have been widely used to mine information from medical databases. In Machine Learning, classification (e.g.: is this specific patient sick or healthy) is a supervised form of learning that can be used to design models describing important data classes. Using those machine learning techniques can support researchers or physicians in making medical decisions and they can answer important and related questions concerning health care.

II. PROBLEM STATEMENT

Heart disease can be managed effectively with a combination of lifestyle changes, medicine and, in some cases, surgery. With the right treatment, the symptoms of heart disease can be reduced and the functioning of the heart improved. The predicted results can be used to prevent and thus reduce cost for 3 surgical treatments and other expensive.

The overall objective of my work will be to predict accurately with few tests and attributes the presence of heart disease. Attributes considered form the primary basis for tests and give accurate results more or less. Many more input attributes can be taken but our goal is to predict with few attributes and faster efficiency the risk of having heart disease. Decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the data set and databases. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients.

Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. According to (Wurz & Takala, 2006) the opportunities to improve care and reduce costs concurrently could apply to as much as 30% of overall healthcare spending. The successful application of data mining in highly visible fields like e-business, marketing and retail has led to its application in other industries and sectors. Among these sectors just discovering is healthcare. The healthcare environment is still information rich‟ but knowledge poor‟. There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in the data for African genres.

III. OUR WORK

Method to detect possible heart disease using the Random Forests algorithm. Cardiovascular diseases are the number 1 cause of death globally – an estimated 17.9 million people died from it in 2016. This machine learning work contributes to healthcare and can detect heart disease on the basis of clinical data and test data from different patients. The result and contribution of this paper is to identify whether a patient has heart disease or not, based on the information of clinical data and test results and so support doctors in making decisions about patient treatments.

IV. RANDOM FOREST TREE ALGORITHM

The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

A random forest is great with high dimensional data since we are working with subsets of data. It is faster to train than decision trees because we are working only on a subset of features in this model, so we can easily work with hundreds of features But as stated, a random forest is a collection of decision trees With that said, random forests are a strong modeling technique and much more robust than a single decision tree. They aggregate many decision trees to limit over fitting as well as error due to bias and therefore yield useful results.

The Decision Tree is a tree-based flowchart model, in which each internal node represents a "test" on an attribute. Each branch represents the outcome of the test and the leaves are a class distribution. The different paths from the root to a leaf represent a classification rule.

The machine learning technique used in this paper is the Random Forests. It is used to classify whether a person has a heart disease or not, based on clinical information and test results about a group of patients.

The Random Forests algorithm is a popular and very efficient algorithm for both classification and regression problems. The principle of Random Forests is to combine many binary decision trees by using several bootstrap samples coming from the learning sample and choosing randomly at each node a subset of variables.

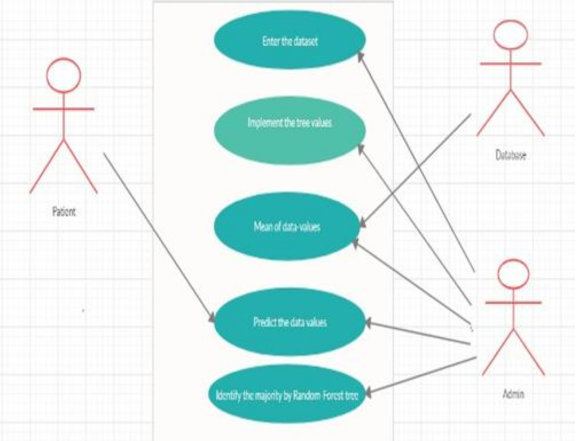


Fig 1.1 System environment

Predict Accuracy

In order to have more reliable and accurate prediction results, ensemble method is a well-proven approach practiced in research for attaining highly accurate classification of data by hybridizing different classifiers. The improved prediction performance is a well-known in-built feature of ensemble methodology. This study proposes a weighted vote-based classifier ensemble technique, overcoming the limitations of conventional DM techniques by employing the ensemble of two heterogeneous classifiers: random forest and classification via decision tree.

Architecture

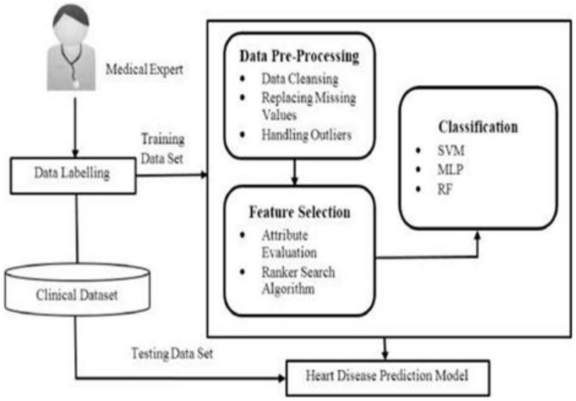


Fig 1.2 System facilitating the subsequent learning and generalization steps, and in some cases leading to better human

Modules description

Preprocessing in Data Mining:

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.

The overall objective of our work is to predict more accurately the presence of heart disease. In this paper, UCI repository dataset are used to get more accurate results. Two data mining classification techniques were applied namely Decision trees and Random forest algorithm.

Attributes with categorical values were converted to numerical values since most machine learning algorithms require integer values. Additionally, dummy variables were created for variables with more than two categories. Dummy variables help Neural Networks learn the data more accurately.

Feature Extraction

In machine learning, pattern recognition and in image processing, feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, interpretations. Feature extraction is related to dimensionality reduction.

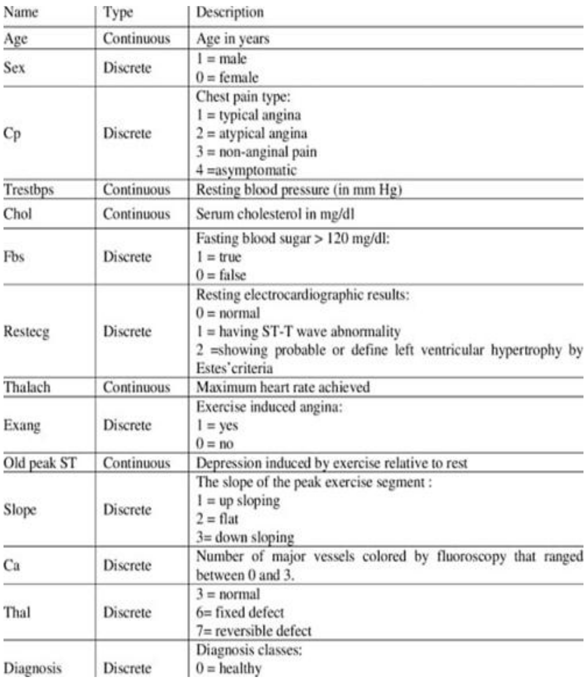
The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data.

Cross Validation

In order to achieve a reliable result from the Random Forests, cross-validation was used. Cross validation divides the data set into a specific number of subsets. Each subset is used by repeating both as a training record and as a test record. The error estimates of all rounds are then summarized and averaged [18]. As used in the method by Rieg et al., a 10 times 10-Cross-Validation was applied. As the result, the algorithm revealed which subjects were correctly classified

Code

```
Name: “stdout”, Text: [
Target    1.000000\n,
Exang     0.436757\n,
Cp        0.4333798\n,
Oldpeak   0.430696\n,
Thalach   0.421741\n,
Ca        0.391724\n,
Slope     0.345877\n,
Thal      0.344029\n,
Sex       0.280937\n,
Age       0.225439\n,
Trestbps  0.144931\n,
Restecg   0.137230\n,
Chol      0.85239\n,
Fbs       0.028046\n,
Name: target,      dtype:float64\n.
Print(dataset.corr()[\”target\”].abs().so
rt_values(ascending=False))”
```



Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	1 = male 0 = female
Cp	Discrete	Chest pain type: 1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 =asymptomatic
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar > 120 mg/dl: 1 = true 0 = false
Restecg	Discrete	Resting electrocardiographic results: 0 = normal 1 = having ST-T wave abnormality 2 =showing probable or define left ventricular hypertrophy by Estes’criteria
Thalach	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise induced angina: 1 = yes 0 = no
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment : 1 = up sloping 2 = flat 3= down sloping
Ca	Discrete	Number of major vessels colored by fluoroscopy that ranged between 0 and 3.
Thal	Discrete	3 = normal 6= fixed defect 7= reversible defect
Diagnosis	Discrete	Diagnosis classes: 0 = healthy

Fig 1.3 Attribute description Random Forest Tree Algorithm

When training, each tree in a random forest learns from a random sample of the data points. The samples are drawn with replacement, known as bootstrapping, which means that some samples will be used multiple times in a single tree. The idea is that by training each tree on different samples, although each tree might have high variance with respect to a particular set of the training data, overall, the entire forest will have lower variance but not at the cost of increasing the bias.

```
fromsklearn. ensemble import Random Forest Classifier
# Create the model with 100 trees
model=RandomForestClassifier(n_ estimators=100,
bootstrap=True, max_features='sqrt') # Fit on training data
model.fit(train, train_labels)
```

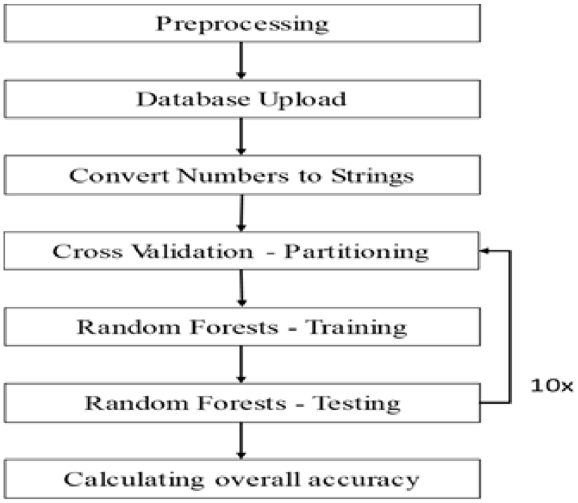

Preprocessing
Database Upload
Convert Numbers to Strings
Cross Validation - Partitioning
Random Forests - Training
Random Forests - Testing
10x
Calculating overall accuracy

Fig 1.4 Flow diagram

In the diagram is a way of representing a flow of process. The Preprocessing is the initial process and the prediction is decided in the section. The Uci data are collected and upload to in this section. The patients’ data are to be compared to the threshold value. Patients details are insert to random forest algorithm for getting more accuracy. Uci data was stored the database and the values are retrieve easily. Build a decision tree and the values are inserted. Then tested the random forest in python. The each and every value is to be tested. At last the accuracy will be given, its more 7 than others algorithm. Finally we calculate overall accuracy.

V. CONCLUSIONS

This paper presented a new approach to heart disease classification, using the Random Forest machine learning algorithm and attributes based on clinical data and patient test results. It reached an overall accuracy of 84.448%. The highest accuracy was reached while using an additional 10 times cross-validation in the process and it outperforms other machine learning techniques using the same database.

Using the Random Forests algorithm without the cross validation secured an overall accuracy of 95%. To further improve the accuracy of the algorithm, updating the database with more information and attributes could help to increase the level of accuracy already achieved. As mentioned in the introduction to this paper, it is already known that Age, Gender, Body Mass Index, Smoking Condition, Homocysteine, Reactive C-Protein, Fibrinogen, Previous familiar cases, Diet, Cholesterol HDL Triglycerides Lipoprotein, Sedentary Condition, Glucose Tolerance and Metabolic System, High blood pressure are risk factors for heart attacks. Although all the listed reasons for heart diseases are known, their epidemiological relevance is different from case to case. Therefore, the attributes probably need to be weighed correctly. In order to estimate the risk of suffering from a heart disease, a global evaluation should be added to the information in the database as well. One solution could be the Anderson Table.

VI. FUTURE WORK

We will triangulate simple ECG sensor data with other physiological sensor data (i.e., heart rate variability, electroencephalography, electro dermal activity, eye fixation, eye pupil diameter). Furthermore, we will experimentally evaluate whether our novel approach is also robust under various conditions of a user's cognitive workload, concentration, and mindfulness. In addition, we will report common method bias evaluations and the results of transferring our novel spectral method to ECG, where we already achieved outstanding results in predicting diseases such as schizophrenia, epilepsy, and sleep disorder based on electroencephalographic data. Finally, we will conduct an empirical implementation study to evaluate acceptance and trust by physicians and patients and if the automated approach improves the coordination between physicians more efficiently.

REFERENCES

[1] World Health Organization, Cardiovascular diseases (CVDs)," 17 May 2017.

[2] J. A. Sanz et al., "Medical diagnosis of cardiovascular disease using an interval-valued fuzzy rule-based classification system," Applied Soft Computing, vol. 20, pp. 103–111, 2014.

[3] T. R. Dawber, G.F. Meadors, and F.E. Moore Jr., Epidemiological approaches to heart disease: The Farmingham study," American Journal of Public Health 41 (3), pp. 279–281, 1951.

[4] L. Biel, O. Petterson, L. Philipson, and P. Wide, "ECG Analysis: A New Approach in Human Identification," IEEE Transactions on Instrumentation and Measurement, vol. 50, no. 3, pp. 808–812, 2001.

[5] B. Wedro, D. L. Kulick, and C. P. Davies, „ Electrocardiogram (ECG, EKG)," eMedicineHealth, online.

[6] J. M. Keller, M. R. Gray, and J. A. Givens Jr., "A Fuzzy K-Nearest Neighbor Algorithm,"IEEE Transactions on Systems, Man, And Cybernetics, vol. Smc-15, no. 4, 1985.

[7] S. Cost, and S. Salzberg, "A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features," Machine Learning, vol. 10, pp. 57–78, 1993.

[8] T. R. Patil, and S. S. Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification", International Journal of Computer Science and Applications, vol. 6, no.2, pp. 256–261, 2013.

[9] H. Zhang, "The Optimality of Naive Bayes," FLAIRS Proceedings, AAAI, 2004.

[10] S. Tong, and D. Koller "Support Vector Machine Active Learning with Applications to Text Classification," Journal of Machine Learning Research, vol. 2, pp. 45–66, 2001.

[11] J. A. K. Suykens, and J. Vandewalle, „Least Squares Support Vector Machine Classifiers," Neural Processing Letters, vol. 9, no. 3., pp. 293-300, 1999.

[12] J. R. Quinlan, "Induction of Decision Trees," Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.