**Asian Journal of Convergence in Technology Volume IV Issue I**
**ISSN NO: 2350-1146 I.F-5.11**

# *MediBot :A predictor and analyzer for CVD*

Kadambari Deherkar, Gauri Rajgopal, Soukhyada Vaidya, Dikshita Iyer

Department of Computer Engineering,

Don Bosco Institute of technology, Mumbai, India
kadambari@dbit.in

***Abstract*— Cardio Vascular disease are now considered to be one of the leading causes for death. It is the essential now to predict such diseases beforehand so that people can take precautions and try to preclude CVD. Thus, it is important to devise a method that aims at predicting these diseases and in less time. This paper aims at developing a CVD predictor which also provides suggestions to the users for lifestyle improvement. For this purpose, this project is making use of the Random Forest Algorithm, which has shown to provide a better accuracy rate than the previously explored methods. The main aim of this project is to help the users catch the risk of a heart disease at an earlier age and curb this in order to lead a healthy and beautiful life. This project will include a Chatbot application called as the MediBot through which the user can enter their details and get an almost accurate prediction of the risk level.**

**Keywords— Data Analysis, Data Mining, Medical information systems, Prediction methods**

## I. INTRODUCTION

The diagnosis of heart diseases mainly happens when a patient is actually suffering from it. For some people, it is possible that they may be prone to heart diseases due to their lifestyle, diet, stress levels, family history and many other factors. In such cases, knowing the risk of suffering from CVD at an early stage itself would help preclude the possibility of a CVD. Heart Disease or CVD has proven to be one of the leading causes of death worldwide since the past few years. CVD is one of the most widely researched topics of today. There are still no accurate methods to predict the risk level of suffering from such a disease.

The main factors that contribute to CVD are age, stress level, Blood pressure level, Sugar levels and family history. A person whose family has had many CVD cases is more prone to suffer from a CVD as compared to any other person with no family history of a CVD. These risk factors can be identified at an earlier stage and the patient can be diagnosed to tend to develop a heart disease. This well help save his life.

In the past few years, extensive research has been done in this field in order to come up with modes for an accurate prediction system. Different methodologies have been compared for their accuracy [2]. Each of these methods have their own discrepancies [1]. An alternative approach for this same method with a better accuracy rate and the ability to handle voluminous data would be the usage of the Random Forest Algorithm [3].

The prediction is done by taking inputs from the user. The system is developed using the Random Forest Algorithm. The prediction and analysis of the data often requires all the attributes to be known. However sometimes, it may so happen that a fraction of the data may be missing. In such cases, prediction would fail since the predictor is not having the values for all the attributes that are required. Nevertheless, using the Random Forest Algorithm would overcome this disadvantage [3]. In this paper, we made use of UCI dataset.

## II. LITERATURE SURVEY

The main aim of the prediction systems is to recognize the high-risk people earlier and prevent the occurrence of CVD in them. This would help to save many lives all

over the world and provide everyone with a better lifestyle.Systems for prediction of heart diseases have been around for more than a decade now.Almost all the researches have utilized the Cleveland Database for Heart Disease in order to train their data. Usage of more recent data may provide varied results for the same system.

In [4] Devendra Ratnaparkhi, Tushar Mahajan and Vishal Jadhav have built a heart disease prediction system based on the Naïve Bayes algorithm. It focused on seven risk factors for heart diseases and it was a web based application. The Naïve Bayes method provided a more accurate and faster result compared to Decision Tree and Neural Networks.

In [5] a system for prediction was built based on Neural Networks which they found to be the most accurate system for non-liner data. The BP algorithm of neural networks was utilized for this purpose. However, the accuracy of this method is in question. There is a need to improve the accuracy of this method. There are found to be more accurate methods for the same.

In paper [3] the main aim was comparison of the algorithms: Genetic Algorithm(GA) , Particle swarm optimization (PSO), Simulated Annealing (SA), Random Forest (RF) for missing data. These algorithms were tested on three different situations: ability to predict forest areas affected by fire under various conditions, and classifying unseen credit and health records. Tested under these circumstances, the RF Algorithm proved to achieve the highest accuracy to classify the unseen records of the Heart Disease data. The main aim of this algorithm is to minimize the error between the expected outcome and the actual outcome.

In [6] Chatbot or Chatterbot is used as a human machine interactor. The chatbot built in this paper was for the Indonesian language and it had a knowledge database. The knowledge database (RDBMS) consisted of the anticipated user inputs and the responses to be given to those inputs. When the user enters some input, a sentence similarity score is generated for the input sentence by matching the input sentence inside the knowledge database. From among the various generated scores, the knowledge sentence with the highest score will be chosen as the input sentence and a corresponding output will be provided for that input sentence.

In paper [7] A HDPS was constructed to compare three algorithms: Naïve Bayes, J48 and Random Forest algorithm. The system was built using the WEKA tool and the main aim was to identify the most efficient algorithm. The Random Forest algorithm proved to be more efficient as compared to J48 and the Naïve Bayes algorithm predicted two more datasets accurately as compared to Random Forest. Keerthana T K compared the techniques of Naive Bayes, J48, Random forest for prediction of heart diseases using the WEKA Tool for prediction. She compared the accuracy of all three methods and it showed that Naïve Bayes had a precision of 83.88% whereas Random Forest gave a precision of 82.8%. The precision rate of Naïve Bayes proved to be the highest among all three. However, the dataset used was small. The Naïve Bayes algorithm works efficiently only when the dataset is small. This shows that Naive Bayes algorithm fails to work efficiently when the dataset is large. Whereas the Random Forest algorithm works efficiently even for large datasets.

TABLE 1. COMPARISON TABLE

| Sr no. | Name of the IEEE paper | Algorithm used in the paper | Accuracy of the algorithm |
|---|---|---|---|
| 1 | Prediction of Heart Disease at early stage using Data Mining and Big Data Analytics: A Survey [1] | GA Technique , J48 | 90% 98% |
| 2 | Predictions in Heart Disease Using Techniques of Data Mining [10] | Naïve Bayes Classifier | 86.12% |

| | | | |
|---|---|---|---|
| 3 | Prediction and Analysis of Heart Disease[9] | Decision Trees, Naive Bayes and K-mean. | 80.4%, 86.12%, 75.18% |
| 4 | Study of Heart Disease Prediction using Data Mining [2] | Naïve Bayes, Decision Tree, Neural Networks | 83.70%, 76.66%, 78.148% |
| 5 | Heart Disease Prediction System using Data Mining Method [7] | Random Forest Algorithm , J48 | 92.11% |

Though J48 has high accuracy it does not consider the missing data which constitutes as a part of accuracy. Hence random forest algorithm proves to be more accurate. The above analysis of the algorithms has been used as a base to implement the algorithm for MediBot.

## PROPOSED MODEL

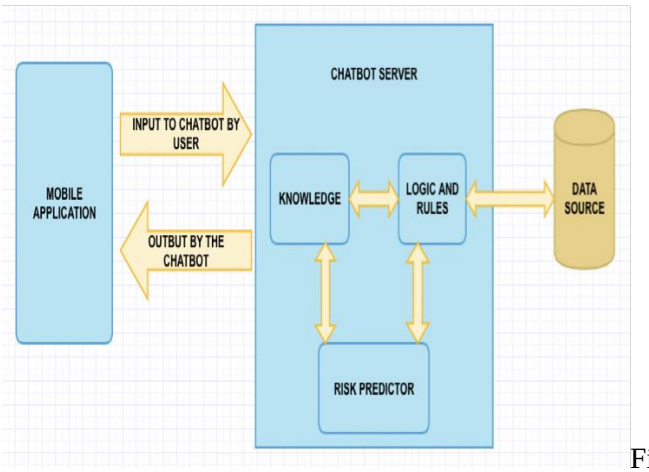


Figure 1. Design of Chatbot

**MOBILE APPLICATION:**
This is the user interface which allows the user to enter his details and check for the risk level of heart disease. The user enters details such as Age, Weight, BP, Sugar levels, etc. This data is given as input to the chatbot server, which processes the information and displays the risk level as the output.

**DATA SOURCE:**
The data source is the training data obtained from the doctors and various sources. This data is used in order to train the application. Patterns are recognized in this data and based on these patterns the logic and rules are derived to predict the accurate risk level of a person suffering from a cardio vascular disease.

**CHATBOT SEVRER:**
The chatbot server is the main module consisting of several sub-modules inside it as follows:

1. Knowledge

The knowledge is the database wherein the input of the user goes. The details entered by the user are stored in the knowledge module. Then these user details are given as input to the logic and rules module.

2. Logic and Rules

The logic and rules module consists of the logic based on which the input details are classified. This logic is derived from the data received by the doctors i.e. the training data. The data entered by the user is classified and it is split into various trees using the random forest algorithm.

3.Riskpredictor

The random forest algorithm then chooses the best of these trees which in turn gives the result of the risk level prediction. This risk level prediction is given as output to the user. The user can then ask for advice as to what changes can be made to his lifestyle in order to preclude a heart disease. When prompted, appropriate changes to diet and lifestyle are suggested to the user to curb the chance of a heart disease occurring.

The above proposed model consists of a training data and a test data set. Supervised learning is used to infer data from training data set.It includes a chatbot which will take data from user in the form of questions. The inputs from the dataset now will be classified on the basis of attributes taken as input and common most important attribute relevant will be filtered out. The prediction is made with the help of Random Forest Algorithm. As per the risk level the chatbot will act as a

consultant for the user. The risk level will be classified into three categories:
Medium risk, high risk, low risk.

The chatbot will need to be priory fed with the data set and also the domain knowledge of chatbot will be predefined. The chatbot will be able to identify queries from the user and fetch the correct output. The chatbot will also be able to identify the same query when put in different words or when conveying the same meaning. Chatbot should be able to identify up the keywords given by the user and then match it with the keyword present the database. The chatbot will have a list of keywords present in database.

### ALGORITHM FOR PROPOSED MODEL

The algorithm used for this project is the random forest algorithm. The reason behind using this algorithm is that it provides a higher accuracy rate for large data sets as compared to the other algorithms. It is used over Naive Bayes because it offers consistent and marked accuracy especially for multiple class classification task. Also, it works the best for missing data. Higher the number of trees in the forest, better is the rate of accuracy of the prediction made.

**Input:**
The input will consist of the following variables:
1. Age
2.Gender (1=Male, 0=Female)
3.Blood pressure
4.Fasting Sugar glucose levels (>120mg/dl)
[1=True,0=False]
5.Cholesterol Level
6.Smoking(1=Yes,0=No)
7.Frequencyintake of outside foods like burgers, pizza, sweets. (In a range of 1-7)
8. History of CVD (1=Yes, 0=No)

**Output:**
The output given will be a risk level prediction for the CVD. It will be in the range of 0-5 where 0 indicated no risk while 5 indicates very high risk.
0= No risk
1= Very low risk
2= Low risk
3=Medium risk
4= High risk
5= Very high risk

**Working:**
Random Forest algorithm works by producing many decision trees and it will give as output the class output by the individual trees. It is considered as one of the most accurate among all the learning algorithms available. For large data sets it works as a highly accurate classifier. There is no risk of over fitting of the data when using theRandom Forest Algorithm.

It is a rule based concept. When the tree is provided with the training dataset and the various attributes, the tree will come up with a set of rules for predicting the risk level of the heart disease. The random forest is developed by aggregating trees. Instead of creating just a single tree, multiple trees are created and the result from these trees are aggregated to obtain the final result. It deals with only 2 free parameters:
1. No. Of trees(Defaultvalue=500)
2. Variables randomly sampled as the candidate at each split

The random forest algorithm can be divided into two major stages:

**STAGE 1: CREATION OF FOREST**
Repeat following steps to create ‘n’ number of trees, thus resulting into a forest of trees:
For i=1 to n:

- Select ‘k’ attributes from among ‘m’ attributes of user randomly.

- From the ‘k’ attributes, one of the nodes, node ‘d’ will be selected as the node providing the best split.
- Split this node into child nodes

**STAGE 2: RISK PREDICTION**
Each tree has certain set of rules for predicting the risk level. At each tree, the attributes selected (k) will be utilized and the rules will be applied for a prediction of that particular tree.

- For each predicted tree, the number of votes are calculated. (Prediction of each tree is considered):

- Let $C_j$ be the class of the prediction of the $i^{th}$ random-forest tree.
- for i=1 to 'n':

Final Prediction = majority vote($C_j$)

As the attributes selected for the creation of the tree are selected at random, hence this algorithm is called as the random forest algorithm.
The random selection of attributes helps to provide an almost accurate prediction even in case of missing data values.

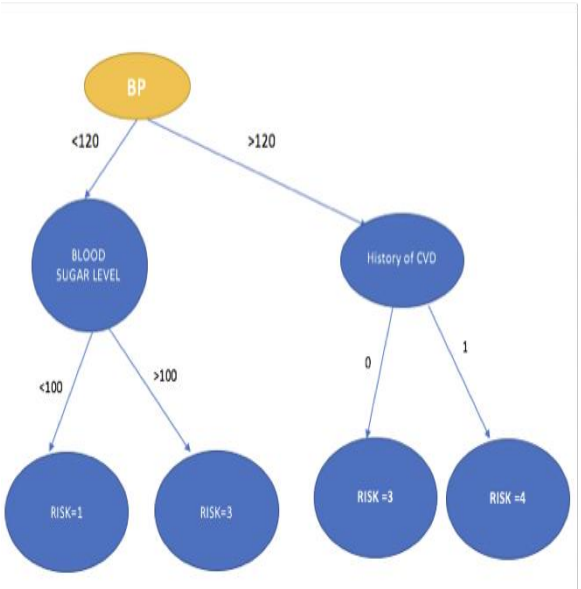


Figure 2. Example of decision tree for risk prediction

Once the risk level for that user is predicted, the user can then query the chatbot for suggestions on how to change one's lifestyle so that the risk of heart disease will be lowered.The chatbot will answer the user's query by providing some dietary and lifestyle recommendations.

## PROTOTYPE

We have for now, developed a prototype of the chatbot which work on terminal and was coded in the Python language.

```
=>hello
C:hi there
=>how are you?
C:I am feeling good
How may i help you?
=>What is my risk level?
C: Please enter the following details:
What is your age? 24
What is your gender?(F=0, M=1)0
What is your BP level?120
What is your fasting blood sugar level?170
What is your cholestrol level?150
How often in a week do you consume junk food? (0-7)??
Did anyone from your family suffer from CVD? (Yes=1, No=0)1
Do you smoke? (Yes=1, No=0)0
Do you drink? (Yes=1, No=0)0
Do you exercise daily? (Yes=1, No=0)0
=>
```

Figure 3. Screenshot of basic chatbot prototype

The inputs taken from the chatbot will then be processed by the random forest algorithm to provide the user with the Risk level prediction.

## SUMMARY AND FUTURE WORK

MediBot is an android based chatbot which aims at successfully predicting the risk level of a user for suffering from a Heart Disease. The main aim will be to provide users with an application that will keep them appraised of their susceptibility to a CVD and help them keep a check on the same.

The future work includes full implementation of the prediction module along with testing of results.The random forest algorithm that will be used for the prediction was shown to work better with large datasets in previous researches.

## ACKNOWLEDGEMENT

We take this opportunity to express our profound gratitude and deep regards to our mentor Kadambari Deherkar for her exemplary guidance, monitoring and constant encouragement throughout the course of this project.We would also like to take this opportunity to express a deep sense of gratitude to Dr Amiya Kumar Tripathy for his cordial support, valuable information and guidance, which helped us through various stages.

## REFERENCES

[1] 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)

[2] K.Sudhakar, Dr. M. Manimekalai , Study of Heart Disease Prediction using Data Mining , International Journal of Advanced Research in
Computer Science and Software Engineering

[3] Collins Leke, Bhekisipho Twala, and Tshilidzi Marwala, Modeling of Missing Data Prediction: Computational Intelligence and Optimization Algorithms, 2014 IEEE International Conference on Systems, Man, and Cybernetics October 5-8, 2014, San Diego, CA, USA

[4] Devendra Ratnaparkhi, Tushar Mahajan, Vishal Jadhav, Heart Disease Prediction System Using Data Mining Technique, International Research Journal of Engineering and Technology (IRJET) Volume: 02 Issue: 08 | Nov-2015

[5] Ankita Dewan, Meghna Sharma, Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification, 978-9-3805-4416-8/15, 2015 IEEE

[6] Bayu Setiaji, Ferry Wahyu Wibowo, Chatbot Using A Knowledge in Database, 2016 7th International Conference on Intelligence Systems, Modelling and Simulation

[7] Keerthana T K, Heart Disease Prediction System using Data Mining Method, International Journal of Engineering Trends and Technology (IJETT) – Volume 47 Number 6 May 2017

[8] Saimadhu Polamuri (2017, May 22), How the random forest algorithm works in machine learning, (1st edition), [Online], Available: http://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing

[9] Sonali S Jagtap, Prediction and Analysis of Heart Disease,International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 2, February 2017

[10] Monika Gandhi, Dr. Shailendra Singh, Predictions in Heart Disease using techniques of Data Mining, 2015 1st International Conference on Futuristic Trend in Computational Analyst and knowledge