



# SECURE DATA ANALYTICS FOR HEART DISEASE PREDICTION

Mr. Rahul Sannappa Belli  
HOD

Ms, Shukranti Sambhaji Patil  
Lecturer

Mr, Amol Shivaji Tambekar  
Student

Mr. Nitesh Subhash Rane  
Student

Department of Computer  
Engineering

Dr. A. D. Shinde Institute  
Of Technology  
Kolhapur, India

**Abstract:** Technological rapid growth in biomedical applications generate high volume of personal data. This biomedical data raises privacy concerns as it reveals sensitive data such as health status and peoples living style. Information generated by biomedical mobile applications need to keep private. The proposed system keeps private data locally on mobile and only data required for heart disease prediction is uploaded to server. Data analytics for Heart disease prediction is implemented using two algorithms Logistic Regression and Nave Bayes. This paper proposes a rule based model to compare the accuracies of applying rules to the individual results of Nave Bayes and logistic regression on the mHealth application database in order to present an accurate model of predicting heart disease.

**Index Terms -** Data Analytics, Nave Bayes, Logistic Regression, mHealth application

## I. INTRODUCTION

Data mining is a process of extracting useful information from large amount of data set. The resulted data is used for prediction purpose so that it can be beneficial for various purposes like improving business process, finding causes of diseases likewise. Different techniques involved in data mining are classification, clustering, association etc. Data mining has immense applicability in diverse area like Biomedical Analysis, Telecom Industry, Intrusion Detection System, Fi- nancial Data Analysis etc. In biomedical analysis, different algorithms are used to design model for healthcare prediction. The algorithms used are Nave Bayes, SVM, logistic regression etc.

Around 17.3 billion people die in the world for year of 2008[9]. In spite of the fact that cardiovascular diseases are controllable by taking cautions which were analyzed using different data mining techniques. Last two decades, a lot of research is going on in health care industry to find out the causes of various diseases as precaution is always better than prevention. Cardiovascular diseases includes heart failure, cardiomyopathy, coronary heart disease etc. and the common causes for these diseases are diabetes, smoking,

high cholesterol, hypertension. Data set used for the data analytics is downloaded from UCI Machine Learning Repository [10].

## II. REVIEW OF LITERATURE

Now a day's multiple people can use forward biomedical sensors and mobile application. That technology generate large amount of biomedical data which include some personal information of patient about lifestyle. So in this paper personal information can be kept secretly and logistic regression Technique can be used to predicate disease prediction and treatment [1].

In today's environment large companies or hospital can be store the patient's sensitive information on host data center. An efficient algorithm / techniques are used for designing predictive model for disease diagnosis and treatment. In this paper the information can be keeping locally and use Holomorphic encryption. Holomorphic encryption can be handle enumeration of such encryption in which cannot be decryption and not need of decryption key [3].

The Support Vector Machine (SVM) is a state-of-the-art classification method introduced in 1992. The SVM classifier is widely used in bioinformatics due to its high accuracy, ability to deal with high-dimensional data such as gene expression, and exhibility in modeling diverse sources of data. In this paper, PPSVC techniques can be used to tackle the privacy violation problem of the classification model of the SVM[5].

MediNet is a system that can be used to develop to personalize the self-healthcare process for patients with diabetes and cardiovascular disease using a mobile phone network. It can be use current and past information from monitoring devices to recommendations. It can provide for the uniqueness of each patient by personalizing its recommendations based on personal level characteristics of the patient, as well as groups of patients share those characteristics [2].

Cloud-assisted mobile health (CAM) can be monitoring mobile communications and cloud computing technologies to provide feedback decision support, has been considered as a revolutionary approach to improving the quality of healthcare service while lowering the healthcare cost. CAM, which can effectively provide security for privacy of clients and the intellectual property Of mHealth service providers.[4]

## III. SYSTEM ARCHITECTURE

For heart disease prediction, this paper proposes a combination of models which is shown in Fig 1. This secure data analytics approach is divided into five modules involving mHealth application, Preprocessing, Training Model, Applying Logistic Regression, applying Naive Bayes, and Result Comparison and Heart Disease prediction. Patient database is collected from mHealth application and also taken from UCI repository for training purpose.

## Proposed System:

The above mentioned five models are described as below.

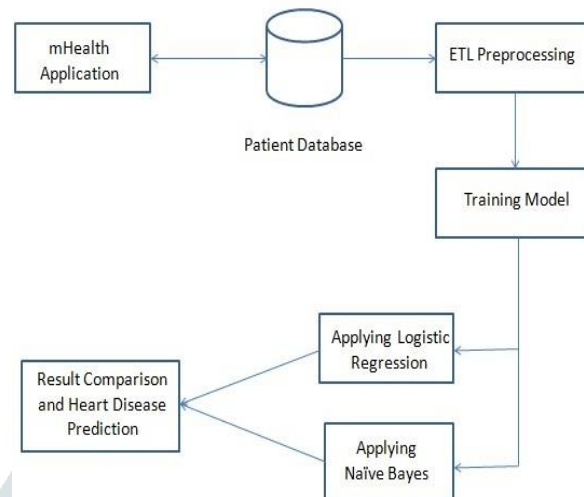


Fig. 1. Block Diagram of Proposed System

### A. mHealth Application

Mobile health technologies are rapidly growing which includes wearable devices and embedded sensors. As technology growing rapidly, mHealth application development using mobile health technology is also growing in a same way. mHealth applications record all day to day activity of individual and inform if any cautions need to take. But it comes with privacy concern.

### B. ETL Pre-processing

ETL stands for Extract, Transform, and Load. It is a tool which is combination of three functions which is mentioned above. It is used to get data from one database and transfer to other database.

Data preprocessing is a data mining technique used to transform sample raw data into an understandable format. Real-world collected data may be inconsistent, incomplete or contains an error. This paper proposes ETL and preprocessing combined together to process on existing patient data and load it into mHealth server database.

Following are data processing techniques.

- **Data Cleaning:** Resolve inconsistency and eliminate noise in data.
- **Data Integration:** Incorporate data from different sources into one rational source such as data warehouse.
- **Data Transformation:** In data transformation, data transform from one source to another source. It involves following terms.

1. Normalization 2. Aggregation 3. Generalization

### C. Training the Model

Each of the two models has been trained using different methods. For logistic regression, the first step to training is to find the significant attributes by calculating their individual P-values. As a rule of thumb, if it

is below 0.05, only then is the attribute significant. The Hosmer-Lemeshow test is also required to check for goodness fit of the model. The corresponding P-value must abide by a 5k

Naive bayes are a classification method works on bayes theorem which assumes independence among attributes. The basic assumption is that presence of a particular feature in a class unrelated to presence of any other. Naive bayes requires less amount of data as compared to other method for estimation of parameters for classification.

#### D. Applying Logistic Regression and Naive Bayes

For prediction of categorical dependent variable out- come, from set of independent variables [8]. Logistic regression is mainly used to for prediction and also calculating the probability of success. Logistic Regression involves fitting an equation of the form to the data[8].

$$y = e^{\Lambda(b_0 + b_1 * x)} / (1 + e^{\Lambda(b_0 + b_1 * x)})$$

(1) x=input values      y=output values

b<sub>0</sub> = bias or intercept term b<sub>1</sub>=coefficient for single input value (x)

The Naive bayes classifier is based on Bayes theorem with independent assumptions between predictors Continuous values associated with each class are distributed according to a Normal distribution [6]. Naive Bayes classification algorithm is based on Bayes theorem.

$$P(A_k/B) = \frac{P(A_k \cap B)}{P(A_1 \cap B) + P(A_2 \cap B) + \dots + P(A_n \cap B)} \quad (2)$$

A<sub>k</sub> = set of mutually exclusive events

B = any event from the sample space such that P(B) > 0 Here, In proposed system, Bayes theorem can be written in below given way.

$$P(D/S) = P(D/S) * P(D)/P(S) \quad (3)$$

D = Disease S= Symptom

## IV. RESULT COMPARISON AND HEART DISEASE PREDICTION

This module includes the comparison of results from Logical Regression and Naive Bayes algorithm. Also this system finds out the accuracy percentage from both the algorithms.

## V. CONCLUSION

This system proposes confidential scheme for predicting heart disease using two different models, Naive Bayes and Logistic Regression. As identified through survey, it is a need to have combinational approach to increase the accuracy of prediction for heart disease.

## ACKNOWLEDGMENT

I express my sense of gratitude towards my project guide Mr. R.S. Belli for his valuable guidance at every step of study of this project, also his contribution for the solution of every problem at each stage.

I am thankful to Mr. R. S. Belli head of the department Computer Engineering, all the staff members and who extended the Preparatory step of this project. I am very much thankful to respected Principal Mr. A. S. Shelake for his support and providing all facilities to complete the project report. Finally I want to thank to all my friends for their support suggestions. Last but not least I want to express thanks to my family for giving me support and confidence at each and every stage of this project.

## REFERENCES

1. Yanmin Gong , Yuguang Fang , Yuanxiong Guo ,Private Data Analytics on Biomedical Sensing Data Via Distributed Computation 1545-5963 (c) 2015 IEEE.
2. P. Mohan, D. Marin, S. Sultan, and A. Deen, Medinet: personalizing the self-care process for patients with diabetes and cardiovascular disease using mobile telephony, in Engineering in Medicine and Biology Society,2008. EMBS 2008. 30th Annual International Conference of the IEEE. IEEE, 2008, pp. 755758.
3. J. W. Bos, K. Lauter, and M. Naehrig, Private predictive analysis on encrypted medical data, Journal of biomedical informatics, vol. 50, pp.234243, 2014.
4. H. Lin, J. Shao, C. Zhang, and Y. Fang, Cam: cloud-assisted privacy preserving mobile health monitoring, Information Forensics and Security, IEEE Transactions on, vol. 8, no. 6, pp.985997, 2013.
5. K.-P. Lin and M.-S. Chen, On the design and analysis of the privacy preserving svm classifier, Knowledge and Data Engineering, IEEE Transactions on, vol. 23, no. 11, pp. 17041717, 2011.
6. R. Agrawal and R. Srikant, Privacy-preserving data mining, in ACM Sigmod Record, vol. 29, no. 2. ACM, 2000, pp. 439450.
7. Shalabi, L.A., Z. Shaaban and B. Kasasbeh, Data Mining: A Preprocessing Engine, J. Comput. Sci., 2: 735-739, 2006.
8. Mythili T., Dev Mukherji, Nikita Padalia, and Abhiram Naidu, A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL), International Journal of Computer Applications (0975 8887), Volume 68 No.16, April 2013.
9. <http://www.world-heart-federation.org/cardiovascular-health/global-facts-map/>
10. Robert Detrano 1989 Cleveland Heart Disease Database V.A. Medical Center, Long Beach and Cleveland Clinic Foundation