ISSN: 2278 – 7798
International Journal of Science, Engineering and Technology Research (IJSETR)
Volume 6, Issue 1, January 2017

# Decision Support Framework to Pre-guide against Heart Disease

S.PADMAVATHI
Associate Professor
Thiagarajar College of Engineering,
Madurai,TamilNadu.

SRUTHI SIVAKUMAR
PG Student
Thiagarajar college of Engineering,
Madurai,TamilNadu.

*Abstract*—**The current evolutionary era of machine learning systems can be imparted for the development of real-world medical applications which would analyze the abnormal behavior of heart health. The heart health is susceptible to numerous problems due to internal and external factors remains unearthed until people suffer severe heart diminution problems. The Conjecture death of the person due to the Myocardial Infarction would occur unless their problematic symptoms are not analyzed in prior.**

**The web based decision support framework is designed for the provision of the pre-guidance report based on the decision generated using Bayesian network analysis. The report is generated in adherence with the patient's non-medical and medical factors which are obtained from the past medical records. The decision model is built using the Bayesian Classifier algorithm to predict the possibility of getting the disease. The model analyzes the casual intervention effects of the non-medical and medical factors of each individuals and would generate a pre guidance health report based on the probabilistic analysis of all the factors obtained from each of the individuals. The web-service would obtain the health metric data from end users in the form of question with health metric values as options. The pre guidance report would be a prognostic analysis of the possibility of individual contingent to the myocardial attack(heart disease).**

Keywords— **Bayesian Networks, Myocardial Infarction, pre- guidance report, casual intervention, questionnaire data, past-clinician medical data.**

## I. INTRODUCTION

The main focus on Medical field research are categorized into two different studies namely the observational studies and then experimental studies. The observational studies are done based on the past medical records where the patients are observed over a period of time and the cumulative collections of their reports are then analyzed statistically to visualize the final results .The current decade has a very large amount of unprocessed medical data that can be used for the artificial intelligence research in order to bring out positive results over the medical field problem.

The experimental studies which are done over the medical field actively intervene over the unknown observations by conducting several runs of designed experiments. The observational studies over the data are done at first and then based on the problem ,the experimental studies should be developed to derive out a solution . The various observational studies are grouped under four different types namely : The case series , case-control series, cross-sectional and the cohort.

In case series methodology , a very quick gathering of data from a group of participants or a group of patients are primarily collected and then describe the decision results based on the disease pattern analyzed over the collected dataset . In this methodology, static set of patient's data are obtained in which one sample subset are alone taken into considerations while performing the experiments.

Case-control series, is another methodology in which the two different groups are taken into consideration which would be used to perform the research experiments. One such set of data would be considered as control group which acts as primary data used as training dataset for the disease pattern analysis and another subgroup of dataset is used for validation purpose in which the analyzed pattern of the trained dataset is working properly to give desired decision as an outcome.

Cross-sectional studies is mainly based on the survey dataset which is randomly collected samples over the larger population group taken into consideration . The cohort studies is the method in which the group of people whose data are obtained by the regular follow up over a period of time.

Meta-analysis over the dataset makes use of the pre-existing research methodology and finally would combine their results in order to obtain overall final decision. The above mentioned studies are done to overcome the most of the common

problems that besets the clinical research while developing cloud based intelligent systems . The final decision are based on the results obtained from the research methodologies with proper expert opinion as a background work in comparison with the meta-analysis work. The various sections of this paper is organized as follows Section II reviews about the various research studies over the medical domain by suing various machine learning algorithms Section III provides design and methodology. Section IV presents the performance evaluation. Finally, the paper is concluded in Section V.

## II. RELATED WORK

The recent trends of incorporation of Artificial intelligence and decision support systems in many of the medical applications has offered a good platform for health monitoring capabilities in Ambient Assisted Living (AAL) lifestyle condition. Continuous medical assessment of health indicators for elderly people should be done at regular interval so that they can live on their own without depending on any support . Statistical process control which would apply for the identification of trends and exhibiting behavior are used to develop a Decision Support System, which supports several medical decision making situations, such as monitoring health assessment indicators on an on going regular basis, diagnosis support and risk assessment. [5]The Fuzzy Cognitive Maps knowledge representation and inference schema used for the disease classification. Personalization is an important aspect of remote disease management systems [7].It involves the forecasting of progression of a disease based on the analysis of patient data set by a knowledge model ,which involves the Markov decision process and also the probabilistic temporal Bayesian related states do not change over time which can be employed for forecasting any medical decision based on the input. The conditional probability table ,the Causal independence is implemented when multiple causes from parent nodes lead to a common effect to the child node through interaction of independent processes. This type of models allow decomposing a probability distribution based on the Boolean variable interactions among local parameters so that the decomposition makes it easier to deal with problems involving a large number of causes[7].

Many of the real-world applications has lot of raw medical data without any intelligence technique applied over it ,if it's being processed properly so that it can benefit the people who can be pre cautious in suspecting their own health disorder. One such technique is Bayesian Network which has the ability to simulate causal interventions in an attempt to perform risk management for medical decision making, perform diagnostic inference and also manages missing information and allows the incorporation of expert knowledge.The information is causally structured in coherence to the expert knowledge [4]. The Hybrid Intelligent Systems provides a solution for the real-world increasingly complex problems which rises the ambiguity issues , uncertainty problems and high-dimensionality explosion of data. They allow to use both a prior knowledge and compose raw medical data into innovative decision solutions. Multi-Classifier Systems (MCS) which is ensemble of different classifiers from heterogeneous or homogeneous modeling backgrounds to give the final decision[6]. The hybrid classifier system can easily be implemented in efficient computing environments such as parallel and multi-threaded computer

architectures. Another current research area focuses over the implementation of AI techniques in the distributed computing systems like Grid or Cloud computing especially when a database or data set is partitioned for privacy reasons, so the partial solutions must be computed on each partition and only the final decision is by the combination of the networked decision by all the algorithm implemented.

The health care services has been designed to provide a high quality guideline based care with decreased medication errors. Different ideas have been proposed for providing e-health care[8] for the web users . The solutions lack a comprehensive integration with its intended environment and needs adaptability to specific infrastructure. The success of Web 2.0 and increasing trend toward social platform oriented web services has resulted in number of health applications such as Patients Like Me, Sugar Stats, Cure Together TU Diabetes that allow patients to maintain their own health records and seek advice for their health status.

## III. PROPOSED DECISION SUPPORT ARCHITECTURE

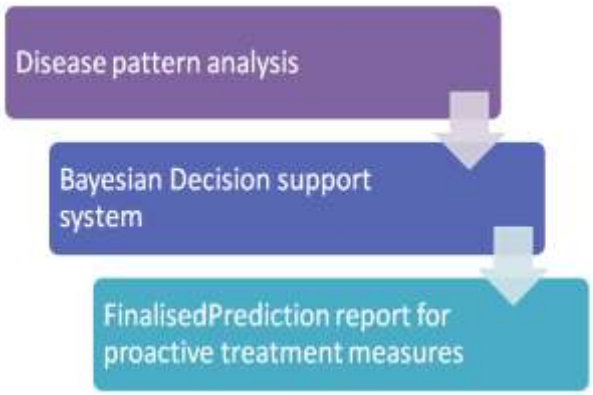


Figure 1: The interaction between various components of the Decision support framework.

The design of the proposed architecture is done based on the steps represented in the Figure.1,in which the previous clinician records are classified by the Bayesian classifier model. The decision rules are imposed based on the dependency of various attributes over the dataset. The results are based on the predicted disease pattern which states the possibility of getting the heart disease.

The interaction of the intelligent Model-View-Controller System represented in the Figure.2, which would obtain the patient's current health status as input and evaluates the input based on the Bayesian Classifier model and then send a report based on their input. The graphical report would be generated as a result of evaluation user input in adherence to Bayesian model, which would state their nature of getting heart disease. The proposed architecture can be deployed by using any of the cloud technology platforms like Amazon EC2,Open stack, Microsoft Azure etc.,. The decision support system is in need for the cloud technology because the end users provide their health oriented values over the web system has to be safely obtained to do the data analysis. The Private cloud can be used for the deployment of the system and in addition to that Scalability of the system can also be achieved since large number of users using this system to analyze their health condition and proactively they can take necessary treatments according to their health status.

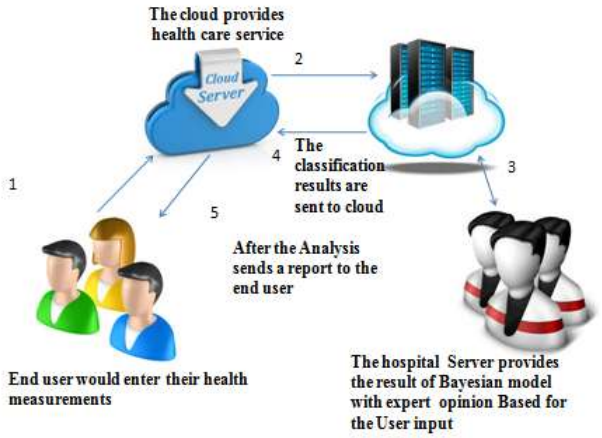


Figure 2: The interaction between various components of the Decision support framework.

The Cohort methodology is used to obtain the disease pattern from the patients in order to apply the supervised machine learning technique for developing a decision support framework . Decision support framework which is developed based on three phase as follows:

1. Pattern analysis
2. Acute Event Identification
3. Identification of correlated medical attributes.

The dataset are at first subjected to pattern analysis such that the user input data from the web portal are matched by the knowledge model to provide them a suggestion to be precautious in taking care of their health issues.

The pattern identification regarding the biochemical value and their externally observable symptoms are done at first by cleaning heart dataset. Each of the attribute is taken into consideration and their sub factors are subdivided for the analysis. The statistical analysis of the dataset as follows:

### *A. PATTERN ANALYSIS*

#### *1) Baseline extraction:*

The first step in the pattern analysis the estimation of baseline features by converting the attribute of multi-value vector into a single value variable by taking their central tendency (mean) and standard deviation are computed based on the equation (1) and (2)respectively . For example taking into consideration of cholesterol value of patients who were suspected and reported for Myocardial infarction (MI )attack are at first analyzed by comparing their own cholesterol level to the mean value of sub quartiles of patients those who had the same cholesterol level who were high probably prone to poor heart health. The standard deviation analysis of the cholesterol level is also done to estimate the status of the patients health to analyze the deviation of their measurements at various time series. The other numerical value obtained from their biochemical test are also subjected to the baseline extraction in order to identify the health condition susceptible to the Heart disorder.

$$central\ tendency = \sum(n_i)\ /n \qquad (1)$$

$$standard\ \ deviation \qquad (2)$$

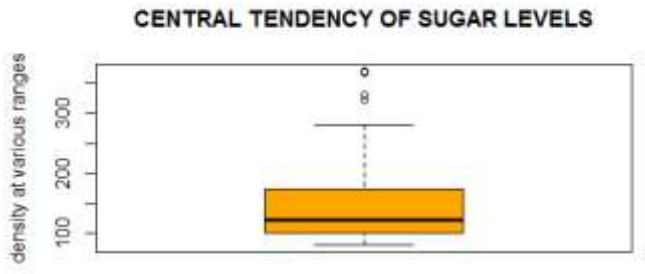


Figure 3: The Box plot represents the central tendency of the appropriate sugar level .

Standard deviation should be analyzed with their confidence interval measures can be estimated using the equation (3) since the Bayesian network work well for the normalized data values.

$$confidence\ interval = \bar{x} \pm z_{(} \qquad (3)$$

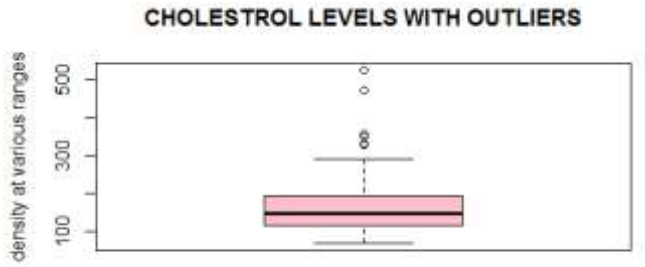


Figure 4: The box plot analyses the central tendency of the cholesterol level with the standard deviation measure with the outliers plotted crossing the standard deviation.

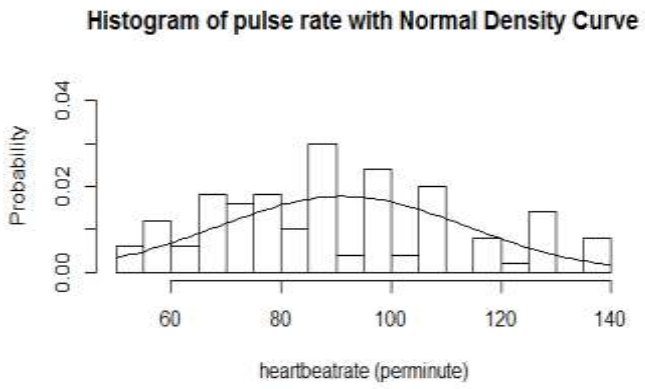


Figure 5: The normalization distribution analysis of the numerical health attribute with their standard deviation value to normalize these continuous medical values

### B. ACUTE EVENT IDENTIFICATION

The continuous monitoring of certain health features has to be done to have accurate probability measure .For example, blood sugar and blood pressure values may vary for each time instances ,which must be estimated regularly such that the deviation from the mean values can be taken into consideration while computing the probability measure.

### C. IDENTIFICATION OF CORRELATED MEDICAL FACTORS

The Bayesian rule is imparted as the decision core process works well over the independent factors. The correlated factors are constructed into a single synthetic node. The synthetic variables are introduced for the purposes :
a) reducing features dimensionality by combining different nodes together to reduce effects of combinatorial factor explosion (e.g. linking the unhealthy habits of the individual and the external health problem).

b) improving causal relationship between health model variables.

Specifically, for this part of the model the medical experts suggested that the observation of the inherent health defects in conjunction with external health disorder and unhealthy behavior condition (i.e. smoking , alcoholic addiction) that cannot be controlled by self management are believed to be contingent to causal risk factors for highly probable to heart failure. However, if we were to model all variables with direct links to heart attack (i.e. without introducing synthetic nodes) this would have resulted in a conditional probability table (CPT) of $(3^3 * 2^8)$=6192 possible state combinations. Clearly, this would have been problematic given that the dataset considered for parameter learning only consists of around 100 random data instances.

Reconstructing network model , with the expertly defined synthetic nodes, reduced the combinatorial explosion by more than 70% (i.e. from 6192 down to $(3^3 * 2^3)$=216, and therefore allowed the formulation of more accurate CPTs, but also improved the causal relationship between factors of health risk analysis.

### D. DECISION SUPPORT CORE FUNCTIONALITIES

Bayesian networks can be used for the establishment of the conditional probabilistic relation among uncertain independent factors of real-life medical scenarios. The decision making these scenarios are concluded based on the correlated or serially influential relation between the factors which are statistical analyzed by the Bayesian rule.

***Algorithm imparted:*** Naive Bayesian classification over clinician medical record.

***Input:*** The user inputs are obtained from the answered questions from the health query webpage.

***Output:*** pre-guidance report based on the likelihood probabilistic measure of given data in adherence to mined results over the clinician records.

1. Construct 'n' attribute vectored training dataset $D_i$ from the original dataset D.
2. Compute the prior probability based on the target class attribute having the favourable outcomes of having the disorder.
3. For each $D_i$ where i $\epsilon${1,2,....n} do

   3.1For j=1,2,...m where m is the number of instances of the dataset.

   3.1.1. Learn the individual attribute $P_j$ while constructing the Bayesian network model over the dataset.

   3.1.2. The Expectation –Maximization step is imparted while classification in order to handle the missing data.

   3.1.3. Estimate the maximum likelihood of the individual attribute in par with the favourable outcomes.

   3.1.3.1 calculate the posterior probability based on the prior probability and the maximum likelihood for each of the attribute of the dataset. The Estimation method as in (4)follows:

$$p\left(\frac{A}{B}\right) = \frac{p\left(\frac{B}{A}\right) * p(A)}{p(A)} \quad (4)$$

   3.2 Compute the joint probability of all attributes of the dataset is using the equation (5) :

$$p(p_1, \ldots \ldots, p_n) = \prod_{i=1}^{n} p(A_i / pa(A_i)) \quad (5)$$

Where $A_i$ denotes probability of individual attribute over the probability family of attributes Ai

   3.3 Apply the intervention combinatoric rule over the parameters to obtain various possibility of proneness to the heart disorder from the dataset.

4.Generate the predicted results by the Bayesian rule over the user input data.

### E. INTERVENTION METRICS FOR HEALTH RISK PREDICTION

An intervention is an action which can be performed to manipulate the effect of some desirable future expected outcome which we would like to predict . In medical informatics, an intervention is represented by some cause which has more than one effect as an outcome. The binary operations such as *AND,OR* relations and also using the *ranked average and weighted average* methods over the interrelated factors can resolve the state explosion problem caused by the dependent medical factors ,thereby providing solution without any over fitting issues over the attributes. The combinatoric rule imposed over the attributes which would compute and analyze the probability of getting the disease based on cause -effect interactions among the interdependent factors.

## IV. RESULTS AND DISCUSSION

The myocardial infarction case study is our base historical dataset of the patients who reported with heart diminution problems .It was used for the development of the Decision support Bayesian model in which the knowledge model is developed over the dataset attributes was designed in consent with the domain experts. The patient's dataset of 100 instances were split into 65 % as training dataset and 35% as testing dataset. The training dataset included 45 males and 20 females. The subject's age range was 25-45. The mean age was 36 years with the standard deviation of 3.8. The questionnaire is developed based on the attributes of the dataset such that the user can enter their own health data values to know about the health status. Out of the 65 % percent of the subjects 52 subjects were "highly prone" to get affected by Coronary heart disease (silent heart attack ) and 13 were " less likely "to get affected by the heart disorder.

### A.EVALUATION METRICS

In medical data diagnosis ,test sensitivity is used to check the ability of the algorithm which correctly identifies those with the high possibility to have heart disease(True positive rate),whereas test specificity is used to check the ability of the algorithm to correctly identify those who would not have chance of getting affected by the disease(True negative rate) ,these are measured and tabulated in Table.1

$$Sensitivity = TP/(TP + FN) \quad (6)$$

$$Specificity = TN/(FP + TN) \quad (7)$$

Here, the characteristics of these metrics:

- True positive (TP) denote the classification results were correct in stating that the person has the highest possibility of having the disease.
- False positive(FP) denotes that the misclassification error result stating the person who has medical complication issues as not susceptible to disease .
- True Negative (TN)states that the no health issues person is not vulnerable to the disease.
- False Negative(FN) is also a misclassification error which contradicts the condition similar to false positive .

Table 1:The classification accuracy in terms of sensitivity and specificity measure of Bayesian Classifier.

| Number of cross validation folds | Dataset split | | Sensitivity Measure | Specificity Measure |
|---|---|---|---|---|
| | Training data split | Testing data split | | |
| 5 | 70 | 30 | 0.633 | 0.328 |
| 5 | 80 | 20 | 0.567 | 0.44 |
| 5 | 85 | 15 | 0.600 | 0.433 |
| 10 | 70 | 30 | 0.730 | 0.274 |
| 10 | 80 | 20 | 0.450 | 0.426 |
| 10 | 85 | 15 | 0.667 | 0.333 |
| 15 | 70 | 30 | 0.750 | 0.250 |
| 15 | 80 | 20 | 0.7380 | 0.224 |
| 15 | 85 | 15 | 0.682 | 0.361 |

Bayesian Classification's Accuracy is measured by the diagnostic capabilities of a knowledge model ,so that the model can correctly diagnose the medical result's ground-truth. The Receiver Operating Characteristic curve (ROC) and the Area Under the Curve (AUC) of the Roc curve is means to evaluate the performance of a binary classifier system. It is created by plotting the fraction of true positives out of the positives (TP = true positive rate) in *x*-axis and the fraction of false positives out of the negatives (FP = false positive rate) in *y*-axis. The terms positive and negative describe the classifier's prediction results, and the terms true and false refer to whether the prediction results correspond to the decision of classifying whether they have the possibility of getting the disease. The ROC curves are plotted as follows:

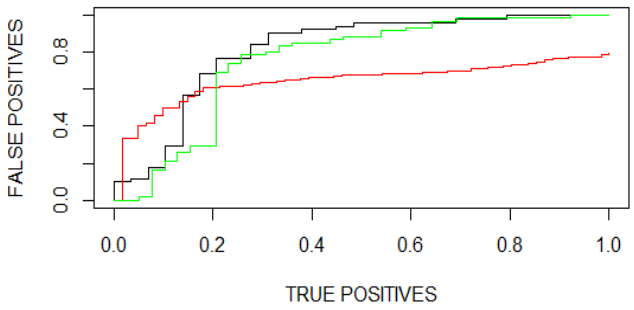


Figure 6:This is multiple roc curves generated for the true positives and false positives of the Bayesian classification for the cross validation of 5,10,15 folds .

## V.CONCLUSION

The classification accuracy from overall runs of the Bayesian classifier is about 78 percentage accurate in classifying the heart disease dataset. The limitations of this proposed approach is that the Bayesian knowledge model is developed from the past two years patients record which is small dataset which should be handled in future with large amount of Electronic medical heart dataset to improvise the classification accuracy . The model demands the independent medical features so that some of dependent factors are reduced into synthetic node based on intervention relationship analysis under the guidance of expert knowledge ,this drawback can be handled by the incorporation of ensemble classifier method which would increase the classification accuracy rate.

## REFERENCES

[1] Peng Wei Wang, Zhi Jun Ding, Chang Jun Jiang and Meng Chu Zhou," Design and Implementation of a Web-Service-Based Public-Oriented Personalized Health Care Platform",IEEE Transactions on Systems, Man and Cybernetics: Systems, Vol. 43, NO. 4, July2013,pp:941-957.

[2] Anthony Costa Constantinoua , Norman Fentona, William Marsha and Lukasz Radlinskib," From complex questionnaire and interviewing data to intelligent Bayesian Network models for medical decision support",Elsevier journal on Artificial Intelligence in Medicine, Vol.67, February 2016,pp:75-93.

[3] Jeferey G. Klann Peter Szolovits , Stephen M. Downs , Gunther Schadow,"Decision support from local data: Creating adaptive order menus from past clinician behavior",Elsevier journal on Journal of Biomedical Informatics, Vol.67, February 2016,pp:75-93.

[4] Anthony Costa Constantinou, Mark Freestone, William Marsh , Jeremy Coid,"Causal inference for violence risk management and decision support in forensic psychiatry ",Elsevier journal on Decision support systems, Vol.80,December 2015,pp:42-55.

[5] A.S. Billis, E. I. Papageorgiou, C. A. Frantzidis, M.S. Tsatali, A.C. Tsolaki and P. D.Bamidis.,"A Decision-Support Framework for promoting Independent Living and Ageing Well",IEEE Journal of Biomedical and Health Informatics , vol .19,No.1, January2015,pp199-209.

[6] Michal woznaik, Manuel Grana, Emilio corchado ,"A survey of multiple classifier systems as hybrid systems," Elsiever journal on Information Fusion,Vol.16,29 April 2013,pp:3-17.

[7] Marina Velikova , Josien Terwisscha van Scheltinga , Peter J.F. Lucas , Marc Spaanderman," Exploiting causal functional relationships in Bayesian network modeling for personalized healthcare" ,Elsevier journal on of Approximate Reasoning , Vol.55,April 2014,59-73.

[8] Pankaj Deep Kaur, Inderveer ,chana, " Cloud based intelligent system for delivering health care as a Service" ,Elsevier journal on Computer Methods and Programs in Bio Medicine, Vol. 113,September 2013,pp:346-359.

[9] Pablo Garcia-Sanchez, Jesus Gonzalez, Antonio M. Mora, Alberto Prieto," Deploying intelligent e-health services in a mobile gateway" , Elsevier journal on Expert systems with application, Vol.40, 2013,pp:1231-1239.

[10] Che-Wei Lina, Shabbir Syed Abdulb,c, Daniel L. Clinciub,d, Jeremiah Scholle,Xiangdong Jinf, Haifei Luf, Steve S. Chenf, Usman Iqbalb,Maxwell J. Heineckg, Yu-Chuan Li," Empowering village doctors and enhancing rural healthcare using computing in a rural area of mainland China" , Elsevier journal on Computer Methods and Programs in Bio Medicine, Vol. 113, October 2013,pp:585-592