A Heart disease prediction using machine learning Algorithms

T.Kalaiselvi, G.Abinav, B.Meikandan

T.Kalaiselvi (AP/CSE), Dept of CSE, Erode Sengundhar Engineering College,Tamil Nadu ,India

G.Abinav UG student, Dept of CSE, Erode Sengundhar Engineering College,Tamil Nadu ,India

B.Meikandan UG student, Dept of CSE, Erode Sengundhar Engineering College,Tamil Nadu ,India

ABSTRACT

In day to day life, there are various factors that affect the mortal heart. Numerous problems are being at a rapid-fire pace and novel heart conditions are fleetly identified. In this stressful of world, Heart, being an essential organ in the body pumps blood through the body for blood rotation essential and its health is to be conserved for a healthy living. This design is driven by the fundamental motivation to create a heart complaint prediction model for determining the likelihood of a heart complaint in a case. Moreover, the objective of this exploratory study is to link the algorithms to this probability. Medical interpreters often face difficulty in recognizing potential heart problems in patients due to the need for exhaustive training and the use of rigorous medical tests. In this work, two data mining methods – KNN and SVM classification – are used to analyse and anticipate the chance of cardiovascular problems. The main purpose of this key exploratory research is to identify algorithms that can accurately classify individuals as either normal or abnormal. Hence, it is now possible to stop the loss of life at an earlier stage. The aforementioned methods definitely outperform other algorithms for predicting cardiac complaints. Python version 3.7 was used to create the design.

Keywords :- *Data mining, Prediction model Classification algorithms, Feature selection, Heart disease prediction*

1. INTRODUCTION

There may also be several inheritable factors through which a heart complaint type is passed down from generations. The World Health Organization reports that globally, more than twelve million deaths each year are attributed to various types of cardiovascular diseases. A "heart complaint" is an umbrella term used to refer to any type of illness that affects the heart and/or circulatory system of a living organism. Young adults between the ages of 20 and 30 are in fact becoming afflicted by heart issues. Young people may be more likely to experience cardiac problems due to poor eating habits, restlessness, lack of sleep, depression, and a variety of other variables, including rotundity, family history, poor diet, high blood pressure, inactivity, high blood cholesterol, and family history, smoking and hypertension. The opinion of heart conditions is an important and is the most complicated task in medical field. When assessing and comprehending cases by medical professionals, all the specified variables are taken into serious consideration and examined with homemade tests at frequent intervals. The heart complaint symptoms greatly depend upon which of discomfort felt by an existent. Some symptoms are not generally linked by the common people. Still, common symptoms include chest pain, breathlessness, and heart pulsations. Angina pectoris, commonly referred to as angina, is a type of chest pain associated with various types of heart problems, and happens when the heart is not receiving enough oxygen. Angina is started by stressful events/physical exertion and typically lasts under tentwinkles.

The symptoms of a heart attack are comparable to angina, however, they occur when a person is not active and are more serious. In some cases, they may even feel like indigestion. Heart attacks can also be caused by various types of heart conditions. Signs and symptoms of a heart attack are heartburn, a stomach ache, a feeling of heaviness in the chest, discomfort that radiates from the chest to the arms, neck, back, stomach, or jaw, lightheadedness, dizziness, nausea, vomiting, and heavy sweating. Heart failure is the result of other heart conditions, and it is caused by the heart becoming too weak to pump enough blood throughout the body, resulting in difficulty breathing. Certain heart ailments may not display any warning signs, particularly in elderly people and those with diabetes. The phrase 'natural heart ailment' applies to a variety of maladies, with typical indications being perspiration, considerable levels of exhaustion, rapid heartbeat and respiration, shortness of breath, and chest discomfort.

Although, these indications may not show up until someone is over thirteen years old. In these types of cases, the opinion becomes an intricate task taking great experience and high skill. Those with a risk of a heart attack or existing heart condition should take precautionary steps and be prepared to act independently if the condition is identified beforehand. Recently, the healthcare industry has been producing vast amounts of data regarding cases and their medical diagnosis records are especially being used for the examination of. Classification is a data mining technique that can be used to anticipate future developments based on existing data. Medical data mining enabled the integration of different approaches and automated the training process on the dataset, which then led to discovering patterns in the medical datasets for the prediction of a case's future condition. Therefore, by using medical data booby-trapping it's possible to give perceptivity on a case's history and is suitable to give clinical support through the analysis. Clinical analysis of cases requires the use of bracket algorithms, which are essential for predicting the probability of a heart attack. Medical data mining utilizes these algorithms in a straightforward way.

This research looks into supervised machine learning to make predictions about the possibility of a person being affected by a heart condition, and examines the efficacy of Random Forest, Decision Tree, and Naïve Bayes data mining techniques to do so. The classification algorithms are trained and evaluated for accuracy. The analysis is done at several situations of cross confirmation and several chance of chance split evaluation styles independently. This research project uses the Statlog dataset from the UCI machine learning repository to construct a model for predicting heart conditions. The model is created by employing bracket algorithms when the heart condition dataset is used for training. This model can then be utilized to assess any kind of heart ailment.

2. RELATED WORKS

This paper aimed to apply data mining techniques to identify the key criteria for predicting patient survival, develop a method for computing the probability of survival, and determine the most suitable healthcare approach for individuals with heart failure, as the outcome of life is often dismal. Five hundred and thirty three cardiac arrest cases were included in the analysis.

They performed classical i) statistical analysis and ii) data mining analysis using substantially Bayesian networks. The mean age of 533 cases was 63 (± 17) and sample was composed of 390 (72) men and 143 (28) women. Cardiac arrest was observed at home for 412 (77) cases, in public place for 62 (12) cases and on public trace for 60 (11)cases. The belief network of variables showed that the remaining alive probability after heart failure is directly associated to 5 variables coitus, age, the original cardiac meter, where heart failure originated and the techniques used for resuscitation.

By utilizing data booby-trapping techniques, clinicians are able to forecast the survival odds of cases and adjust their treatment methods. This research was conducted for each medical procedure and medical issue, which enabled the development of a decision tree rapidly with the data of a service. A comparison of classic analysis and data mining analysis revealed the efficacy of the data mining system in organizing variables and recognizing the importance or consequence of data and variables on the research objective. The main limit of system is knowledge accession and necessity to obtain sufficient data to yield an applicable model.

The sudden cardiac death has a high impact on public health, as the survival rate for cardiac arrest cases is only 1-20%. Each year, cardiac arrest leads to over 350,000 fatalities in the United States and more than 20,000 in France. It is characterized as a sudden, irreversible disruption of the heart's normal rhythm, and the lack of palpable femoral pulses for a period of five seconds or longer. If not treated through resuscitation, it leads to sudden cardiac death. The profiles of the cases are now well known since it generally concern men from about 40 to 75 times. It is estimated by the American Heart Association that heart disease is the leading cause of death in the United States annually, with approximately people suffering from a heart attack. The method of assigning culpability in these heart attacks is inconsistent, and some studies show a preference for certain strategies depending on the source of the cardiac arrest. Therefore, hospitalization should be swift and efficient.

Almost 65,000 people in the United States pass away from heart issues every year, and 95 percent of those who suffer from sudden cardiac arrest do not make it to a medical facility. These facts demonstrate the importance of predicting mortality risk and examining the occurrences throughout treatment in order to provide prognostic information. Furthermore, this is higher than the combined death toll from the top six causes of death, which are chronic lower respiratory conditions, cancer, accidents, influenza, diabetes mellitus, and pneumonia. Previous statistical investigations have shed light on the epidemiology of heart failure and the root causes of this condition. The use of probability in a statistical method was examined in this paper to demonstrate heart failure in case studies and forecast the consequences of treatments in the care process. The authors of the paper proposed that the machine literacy research conducted over the past decade should be widely applied in the upcoming decade. The authors of the AAAI Press book Knowledge Discovery in the Databases held a positive outlook for the possibilities that this research could offer.

The editors of AI Magazine hope that this enthusiasm will be reflected in their archives. Both scientific and government databases are expanding rapidly. The National Aeronautics and Space Administration had far more data than they could analyze. Earth observation satellites, planned for 1990s, are anticipated to induce the one terabyte (1015 bytes) of data daily — further than all previous operations combined. At a rate of 1 pictureeach alternate, it would take the person several times (working nights/weekends) just to look at the film land generated in each day. The Human Genome Project, which is backed by the government, will save the thousands of bytes for each of the billions of heritable bases in biology.

Closer to everyday lives, the 1990 U.S. tale data of a million million bytes render the patterns that in retired ways describe the cultures and mores of moment's United States. What are we supposed to do with this deluge of theraw data? Easily, little of it'll ever be seen by the mortal eyes. Despite the fact that basic techniques for data analysis have been around for a while, more advanced methods for intelligent data analysis are yet to be perfected. Nevertheless, if this is to be understood at all, it must be examined by computers.

As a result, there is an increasing gap between the amount of data generated and its understanding. It is expected that when data is thoroughly analyzed and portrayed in a useful manner, it can become a valuable asset that can give a competitive edge. The computer wisdom community is responding to the scientific and practical challenges presented by need to get the knowledgeadrift in the deluge of data. In assessing the eventuality of the AI technologies, Michie (1990), the leading European expert on machineliteracy, prognosticated that "the coming area that is goingto explode is machine literacy tools usage as an element of large-scale data analysis." The recent National Science Foundation factory on the database exploration future ranked data mining among the most promising exploration motifs for 1990s.

American Airlines utilizes its frequent flyer database to recognize the customers of greatest value, allowing them to be presented with special marketing promotions. Farm Journal analyzes its subscribers' databases and uses advanced printing technology for custom-figuring hundreds of editions acclimatized to specific groups. Several banks have been able to provide better loans based on the patterns identified in previous loan and credit histories. The General Motors also use a database of machine trouble reports to decide the individual expert systems for their colorful models. Manufacturers of packaged goods analyze supermarket scanner data to examine the sales of their products and to identify consumer buying habits.

The demand for, and effort to provide, tools and methods to explore databases has been heightened due to an amalgamation of business research interests. This book is the first of its kind to bring together the latest research from all over the world on these topics. It spans various different approaches to discovery, including inductive literacy, Bayesian statistics, knowledge accession for expert systems,semantic query optimization information proposition, and fuzzy sets.This book is intended to provide information and motivation for those interested in gaining knowledge about computers and data processing and to encourage further exploration. It will be of particular interest for professionals working in databases and operation information systems and to persons applying machine literacy to real-world problems.

3.METHODOLOGY

3.1 BIG DATA ANALYTICAL MODEL

The utilization of big data analytical tools in cardiovascular care can result in improved care and cost savings. The utilization of big data analytics as the basis of many new digital health platforms and health tools that are driven by Artificial Intelligence illustrates this.The implementation of significant predictive models is a complex process of large data analytics. In the past, prediction models have relied on a restricted amount of particularized variables manually entered to figure out the risk score. These models usually work well when they are used at the population level, but not at the individual patient level. A multitude of risk models related to cardiovascular diseases are utilized to make medical judgments, despite their number.

Analyzing big data can provide more meaningful insights into topics ranging from mortality to case-reported issues to resource usage, making it more useful in a clinical context. For example, machine intelligence can identify patterns emerging directly from the data, rather than relying on predetermined variables. The traditional statistical models were assessed alongside a full range of associations and relations between data. A training process was utilized by Machine literacy which included various data sets being fed to the model repeatedly to analyze a large variety of predictive characteristics for enhanced examination.

Deep phenol-typing, or phenol-mapping, is a promising approach when it comes to big data. Complaint groups, or phenotypes, tend to be fuzzy and varied; big data analytics can detect similar case clusters, leading to the development of multiple phenotypes within each complaint domain. A more exact phenol-mapping of grievances, countries and areas may help to give more customized medical advice. Given the need to incorporate larger and more intricate datasets, big data plays a significant role in seeking to achieve excellence in healthcare. These datasets are supported by the combination of various data sources from extensive case populations, to estimate the implicit advantages of ICD's for individual cases. Furthermore, advanced analytical software is a hallmark of big data analytics.

The Big data analysis guide programs to addressthe certain case member by particular interventions. The success of policy is critically dependent on quality of the underpinning exploration and quality (effectiveness) of the interventions. For various interventions (for case in social/ internal health sphere) widely accepted styles to validate success were still lacking. There are several challenges result regarding the Big Data and populate heart complaint similar as

- Data protection regulations make it delicate to dissect datafrom various providers and services of heart conditions combined.
- Significant parts of population health records are unshapedheart complaint textbook.
- There are data quality, interoperability and limitations regarding data integration.

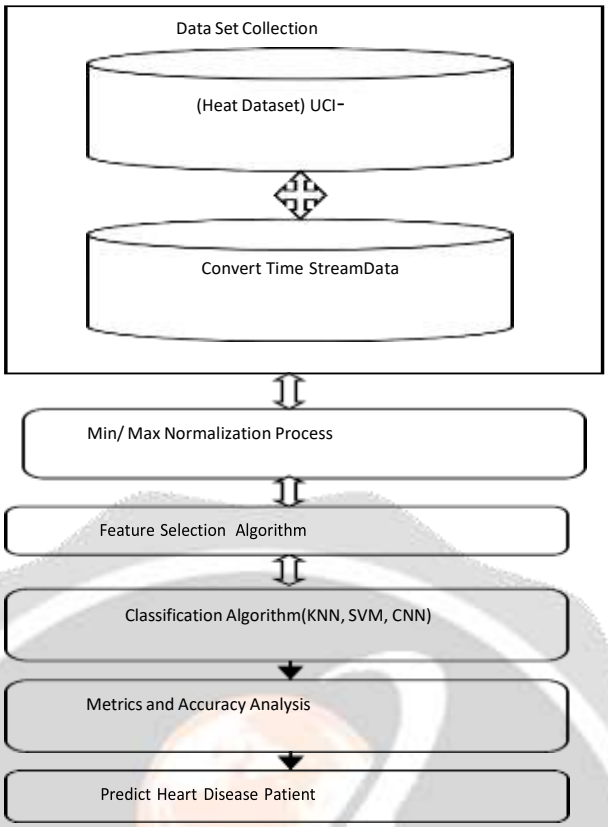

Fig 3.1 Proposed System Architecture Diagram

3.2 NORMALIZATION MODEL

Data normalization is a preprocessing step used in data mining systems to modify the values of an attribute in the dataset to fit within a specified range, for example between 0.0 and 1.0. This is especially beneficial for classification algorithms, such as those involving neural networks, or for distance measurements that include nearest neighbor classification and clustering. Other normalization methods include min-max normalization, normalization by decimal scaling, and z-score normalization, with min-max normalization performing particularly well in linear transformations of the original dataset. This type of normalization maps the value d of P to d′ within the range [new_min (P), new_max (P)].Min max normalization preserves relationship among the actual heart dataset values**.** The table 4.1 describes the sample normalized heat disease dataset model details which show the following details.

Attribute	**Original Values**	**Normalized dataset**
age	70.0	35.0
chest pain type	1.0	0.0
resting blood pressure	130.0	140.0
maximum heart rate achieved	109.0	79.0
exercise induced angina	0.0	1.0

Table 3.1 Normalized Heart Dataset

3.3 GREEDY FEATURE EXTRACTION MODEL

Despite the fact that the selection of significant features without class markers is still a challenge, feature selection, one of the techniques for dimension reduction, is often employed to improve comprehension of the data and enhance the performance of other analyses. This is particularly so for supervised learning, which has been extensively studied.

In this paper, an innovative unsupervised point selection system was devised that carefully selects features. Additionally, a useful criterion to evaluate the reconstruction error of the matrix was set up that was contingent upon the chosen subset of features. Furthermore, an algorithm was established to reduce the reconstruction error based on the given features. This algorithm employed a recurrence equation to estimate the reconstruction error.

The greedy algorithm, which is more computationally and memory efficient than other advanced techniques for unsupervised point selection (both forward and backward), selects the most indicative point from the remaining features, thus removing the impact of the previously selected features from the data matrix. This decreases the possibility of analogous features to the ones already identified being chosen, thus reducing the redundancy of the named features.

3.4 CLASSIFICATION ANALYTICAL MODEL

The field of Machine Literacy is growing quickly, and it focuses on how computers can take in data and use it to become more effective. Included in this field are algorithms that can detect patterns and make predictions from the data. Classic machine learning tasks connected to data mining are some examples of this.

• Supervised bracket literacy

The supervised learning model requires that all the data is labeled and the algorithm will be taught to predict the outcome from the training dataset.
E.g. SVM.

• Unsupervised bracket learning

In this session, an unsupervised algorithm is used to cluster a grounded input dataset to find the essential structure without including the entire data.
E.g. K- means, KNN Neural Networks

•Semi-supervised bracket Learning

Semi-supervised literacy is combination of both supervised & unsupervised literacy. In this technique, some data are labeled while some remain unlabeled. This approach entails constructing class models with a labeled dataset and establishing the separating lines between the classes with an unlabeled dataset.

4. RESULTS AND DISCUSSIONS

4.1 DATASET DESCRIPTION

This paper used Indian Heart Case (ILPD) Data Set (table 4.1) to prepare the database and carry out the results.

Attributes Type	Description	Gender Categorical
age	age given in years	Real number
Sex	sex (Value 1 : male;Value 0 : female)	String
Cp	chest pain type(1:typical angina ; 2:atypical angina 3: non-anginal pain ;4: asymptomatic)	Real number
Trestbps	resting blood pressure (in mm Hg on admission to thehospital)	Real number
Chol	Cholestoral(Serumcholestoral) in mg/dl	Real number
Fbs	Fasting blood sugarin mg/dl (>120) Value 1 = true; Value 0 = false)	Real number
Restecg	Resting electrocardiographic results	Real number
Thalach	Heart rate achievedat maximum	Integer
Exang	Exercise inducedangina (Value 1 : yes; Value 0 : no)	Integer
Oldpeak	ST depression originated by exercise relative torest	Integer
Slope	Slope of the peak exercise ST segment (Value 1:upsloping ; Value 2: flat ; Value 3:downsloping)	Integer
Ca	Major vessels (0-3)colored by flouroscopy	Integer
Thal	Result of thalium stress test (Value 3 = normal; Value 6 = fixed defect; Value 7 = reversable defect)	Integer
Num	status of heartdisease (angiographicstatus) Value 0: < 50% diameter narrowing Value 1: > 50% diameter narrowing	Binary

Table 4.1 Dataset Attribute

Table 4.1 describes the attribute type, description and Gender Categorical values.

Fig 4.1 Performance of SVM Classification

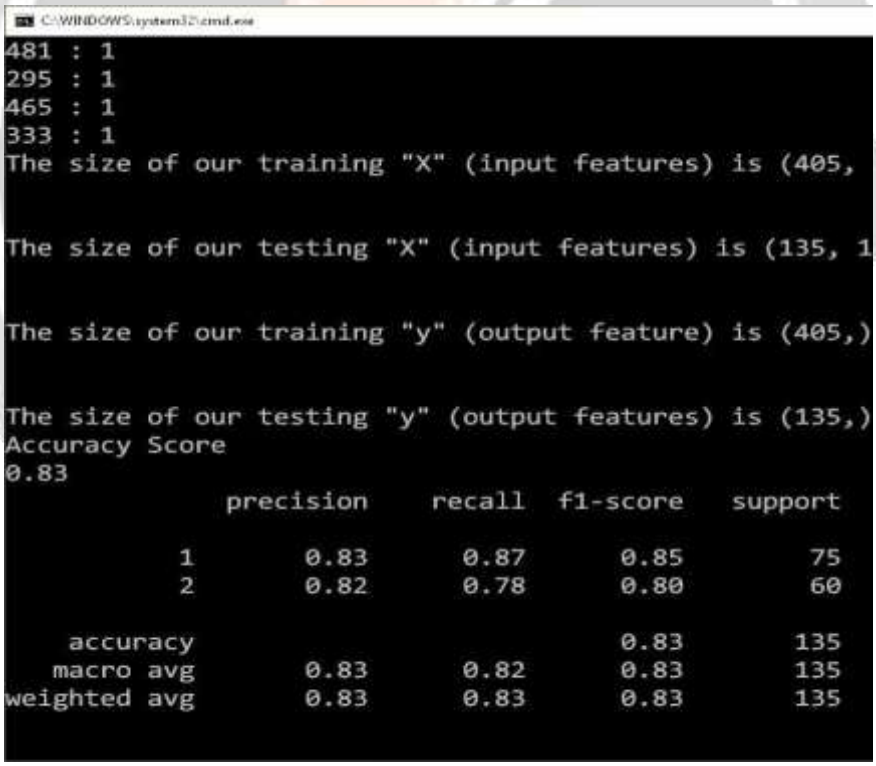

Fig 4.2 Performance of KNN Classification

Fig 4.1, Fig 4.2 describes the performance analysis of SVMand KNN classification algorithms respectively.

5. CONCLUSIONS

This project aimed to accurately classify the different types of heart disease, based on data from the UCI repository, through the use of SVM classification algorithms. Preprocessed like zero values, N/A values and unicode character removal are also carried out here. Important attributes are extracted out for better classification. The confusion matrix is generated by computing an accuracy score value, and KNN classification algorithms and neural networks are used to identify different risk types accurately. Before this, the dataset is taken and preprocessed, for example by removing Unicode. Critical features are extracted for better classification. The confusion matrix is created with the accuracy score calculation. Accuracy prediction is then carried out. A Convolutional Neural Network-based prediction model was constructed to evaluate the efficacy of the algorithm. 510 training records and 240 test records were utilized to train the CNN. This investigation into the classifications is only in the early stages; various avenues of research and exploration are open.

6. REFERENCES

[1] Liaqat Ali, Atiqur Rahman, Aurangzeb Khan, Mingyi Zhou, Ashir Javeed,Javed Ali Khan, "An Automated Diagnostic System For Heart Disease Prediction Based On X^2 Statistical Model And Optimally Configured Deep Neural Network",IEEE ,Vol.7, No. 3,pp. 34938 – 34945, 13 Mar 2019.

[2] Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques",IEEE Access,vol.7pp. 81542 - 81554, 2019.

[3] S. M. M. Hasan, M. A. Mamun, M. P. Uddin, M. A. Hossain, "Comparative Analysis Of Classification Approaches For Heart Disease Prediction",IEEE Access, No. 4, Feb 9 2018.

[4] B. Jin, C. Che, Z. Liu, Shulong Zhang, Xiaomeng Yin, X.P. Wei, "Predicting The Risk Of Heart Failure With Ehr Sequential Data Modeling",IEEE Access,Vol. 6,No. 3,pp. 9256 – 9261,2018.

[5] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, ``Prediction of heart disease using machine learning," in Proc. 2nd Int. Conf. Electron.,Commun. Aerosp. Technol. (ICECA), Mar. 2018, pp. 1275,1278.

[6] S. N. Rao, P. Shenoy M, M. Gopalakrishnan, and A. Kiran B, ``Applicability of the Cleveland clinic scoring system for the risk prediction of acute kidney injury after cardiac surgery in a South Asian cohort," Indian Heart J., vol. 70, no. 4, pp. 533,537, 2018. doi: 10.1016/j.ihj.2017.11.022.

[7] T. Karay_lan and Ö. K_l_ç, ``Prediction of heart disease using neural network," in *Proc. Int. Conf. Comput. Sci. Eng. (UBMK)*, Antalya, Turkey, Oct. 2017, pp. 719_723.

[8] T. Mahboob, R. Irfan, and B. Ghaffar, ``Evaluating ensemble prediction of coronary heart disease using receiver operating characteristics," in *Proc.Internet Technol. Appl. (ITA)*, Sep. 2017, pp. 110_115.

[9] Marjia Sultana, Afrin Haider, Mohammad ShorifUddin," Analysis Of Data Mining Techniques For Heart Disease Prediction", IEEE,pp. 641-648, Vol. 32,2016.

[10] M. Durairaj and V. Revathi, ``Prediction of heart disease using back propagation MLPalgorithm,"*Int. J. Sci. Technol. Res.*, vol. 4, no. 8, pp. 235_239,2015.

[11] Shantakumar B. Patil and Y.S. Kumaraswamy, "Intelligent and Effective Heart Attack Prediction System using Data Mining and Artificial Neural Network", European Journal of Scientific Research, Vol.31, No. 4, pp. 642-656, 2009.

[12] Latha Parthiban and R. Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Biological, Biomedical and Medical Sciences, Vol. 3, No. 3, pp. 1-8, 2008.

[13] Sellappan Palaniappan and Rafiah Awang, "Intelligent Heart Disease Prediction System using Data Mining Techniques", International Journal of Computer Science and Network Security, Vol. 8, No. 8, pp.1-6, 2008.

[14] Kiyong Noh, HeonGyu Lee, Ho-Sun Shon, Bum Ju Lee and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", Intelligent Computing in Signal Processing and Pattern Recognition, Vol. 345, pp. 721-727, 2006.

[15] Franck Le Duff, CristianMunteanb, Marc Cuggiaa and Philippe Mabob, "Predicting Survival CausesAfter Out of Hospital Cardiac Arrest using Data Mining Method", Studies in Health Technology and Informatics, Vol. 107, No. 2, pp. 1256-1259, 2004.