Heart Attack Detection Using Bag Of Words With Machine Learning

Research Paper

Surbhi.Kadu[1], Dr.G.R.Bamnote[2]

Surbhi.Kadu[1], PRMIT&R Badnera , Amravati, India.

G.R.Bamnote[2] , PRMIT&R Badnera , Amravati, India.

Abstract

According to recent survey by WHO organization 17.5 million people dead each year. It will increase to 75 million in the year 2030. Medical professionals working in the field of heart disease have their own limitation; they can predict chance of heart attack up to 67% accuracy, with the current epidemic scenario doctors need a support system for more accurate prediction of heart disease. Machine learning algorithm opens new door opportunities for precise predication of heart attack. The Bag of Words model (BOW) originated in natural language processing. It makes the simplifying assumption that the order of the words in a sentence or text document is of negligible importance for classifying it. This paper highlights the important role played by the machine learning algorithm in analyzing huge volumes of healthcare related data in prediction and diagnosis of disease.

Keywords: *Heart Diseases, Data mining, k-mean clustering, bag of words*

1. INTRODUCTION

Healthcare organizations are facing with challenges to give cost-effective and high quality patient care. Administrators and clinicians need to examine wealth of data that is easily available in the databases of healthcare information systems so that they can discover knowledge and to make informed decisions. This is basic specifically to improve the viability of sickness treatment and preventions. It has becomes of more important in case of heart disease (HD) that is regarded as the primary reason behind death in adults. Data mining serves as an analysis tool to discover unapparent or hidden relationships and patterns in HD medical.

There are five models constructed of single and combined data mining techniques to support clinical decisions in (HD) diagnosis and prediction. The five systems give automatic pattern recognition and attempt to reveal relationships among different parameters and symptoms of HD. Each system reveals set of strengths and restrictions in terms of the type of data it is handling, accuracy, reliability and generalization ability. Weak generalization ability is still a big open issue for data mining in healthcare mainly because of the lack of input data and cost of re-processing. Data can be a great benefit to healthcare organizations, but they also have to be transformed into some information". Many demands are placed on using this information to build knowledge that allows the strategy of healthcare organizations: maximize patient care and cost. The healthcare environment is considered as being "information rich" but "knowledge poor" There is a wealth of clinical and administrative data available within healthcare systems, however, there is also a absence of effectual analysis tools to find knowledge contained in the databases of these systems . Knowledge Discovery in database (KDD) is mention as the "non trivial extraction of implicit previously unknown and potentially useful information about data". Data mining is the core of KDD and which is mention as "a process of selection, exploration and modeling of huge quantities of data to find regularities or relations that are at first unknown with the aim of obtaining a useful results for the owner of database". The knowledge that is discovered in healthcare databases can also be used by healthcare administrators to improve operations and quality of service. It can be useful in the healthcare professionals to improve their medical practice . The objective is to provide an appraisal of current state-of-the art applications of knowledge discovery in medical databases using data mining techniques to predict heart conditions. There are several of applications that use a single

or combination of predictive data mining techniques to improve prediction accuracy. The bag of words technique using k-mean clustering and machine learning to make predictions over available medical data.

A bag-of-words model, or BOW for short is a way of extracting features from text for use in modeling, such as with machine learning algorithm. The approach is very easy and flexible, and it can be used in a myriad of ways to pull out features from documents. A bag-of-word is a representation of text that describes the occurrence of words within a document. It is called bag-of-words because any information about the order or structure of words in the document is discarded. The model is only worrying whether words that are known occur in the document, not where in the document. The bag-of-word model is a way of representing text data when modeling text with machine learning algorithms. The bag-of-model is easy to understand and implement and has great success in many problem such as language modeling and document classification.

2. RELATED WORK

Asha Rajkumar [13] suggested a prototype of IHDPS (Intelligent Heart Disease Prediction System) implementing data mining algorithms, like Naive-Bayes, Decision Trees and Neural Network. The final output of these algorithm describe that each method has its different capabilities in the purpose of the described mining goals. Intelligent Heart Disease Prediction System is simple, easy to use, scalable, expandable and reliable web based predication system that can give output of difficult queries and the traditional decision support system fail to do. In IHDPS author's use the medical profile attributes like age range, gender, high blood pressure and high blood-sugar to discover the symptoms of patients. It is implemented on the .NET platform

Das and Turkoglu [16] introduce a technique that uses Statistical Analysis System (SAS) software version 9.1.3 for investigation of the heart base disease. A neural networks ensemble based techniques generate fresh models by linking the posterior possibilities and discovered values against multiple predecessor models for creating high accurate models. The research exercise were done on the heart diseases dataset to predict heart disease in an entirely automatic way using 3 independent neural networks models to develop the ensemble models. The authors obtained 89.01% classification accuracy, 95.91% specificity and 80.95% sensitivity values on the data drawn from Cleveland heart diseases database.

Srinivas [14] intended application of data mining techniques in discovery of heart diseases. The author used the tanagra tool for implementation of data mining, statistical and machine learning algorithms. Author use the training data set with 14 different attributes and 3000 instances. The experiments were performed on the training data set to measure algorithm performance, in terms of time taken and accuracy. The instances and attributes in the data set were describing the outcomes of various kind of experiments to calculate the efficiency of heart disease. Author divide the data set into two different parts, 30% of data was used as testing and 70% was used as training. The comparison was done on the bases of 10 fold cross validations. The proposed work best performance algorithm results were 52.33% accuracy using Naive Bayes among of these classification algorithms on heart diseases data set.

Shouman [2] presented K-means clustering using the decision tree technique to measure accuracy the heart disease data set. To boost the efficiency of K-means clustering they proposed various kind of centroid selection technique.

Cleveland Clinic Foundation Heart diseases. An accuracy and sensitivity were measured with several centroids selection technique and several bunch of clusters. The ten different runs were performed for the random attribute and random row techniques and the average and best for each technique were measured. Finally author compare the performance of previously used decision tree implemented formerly on the same data with the combine implementation K-mean clustering and decision tree approach. The integrating k-means clustering with decision tree improve resultant efficiency of decision tree to predict heart diseases of patients. The accuracy improved by the enabler technique with 2 clusters was 83.9%. It has been identified from Literature survey, that there are certain issues which are still needs improvement. Classification and association suffers from inefficiency, due to the evidence that it usually produce huge number of rules in associative rule mining. So it very difficult to select best suitable and effective rules from among them. Many associative classifiers create rules in a level wise manner with low support pruning. Often that ahead to generation of a large amount of insignificant rules and at the same time good rules with relatively low support are not produced. In most of the case associative classification algorithms support the exhaustive search technique used in the foremost a prior method to finding the rules and need many

number of passes over the large data sets. Moreover, they discover frequent items in single phase and create the different rules in different phase exhausting more efforts, storage and processing time.

In case of heart diseases diagnosis the accuracy of heart data set is calculated on the basis of classification methods like Naive Bayes, IBk, Neural Network, Decision tree etc. Accuracy of correctly classified instances is not sufficient to predict heart diseases on the basis of training data set and need implementation of association classification approach for better accuracy.

3. PROPOSED WORK

A methodology to detect a heart attack on the basis of available patient dataset which contains various symptoms and other health related parameters like B.P, H.R, etc. Various techniques for pre-processing of collected data are applied such as applying stop words removing, stemming, feature reduction and feature selection techniques to fetch the keywords from all the attributes and finally using different classifiers to decide whether their is heart condition or not. Numeric feature representation technique for feature extraction and K-mean clustering algorithm for clustering feature vectors, in various clusters and machine learning technique to make predictive analysis of heart condition.

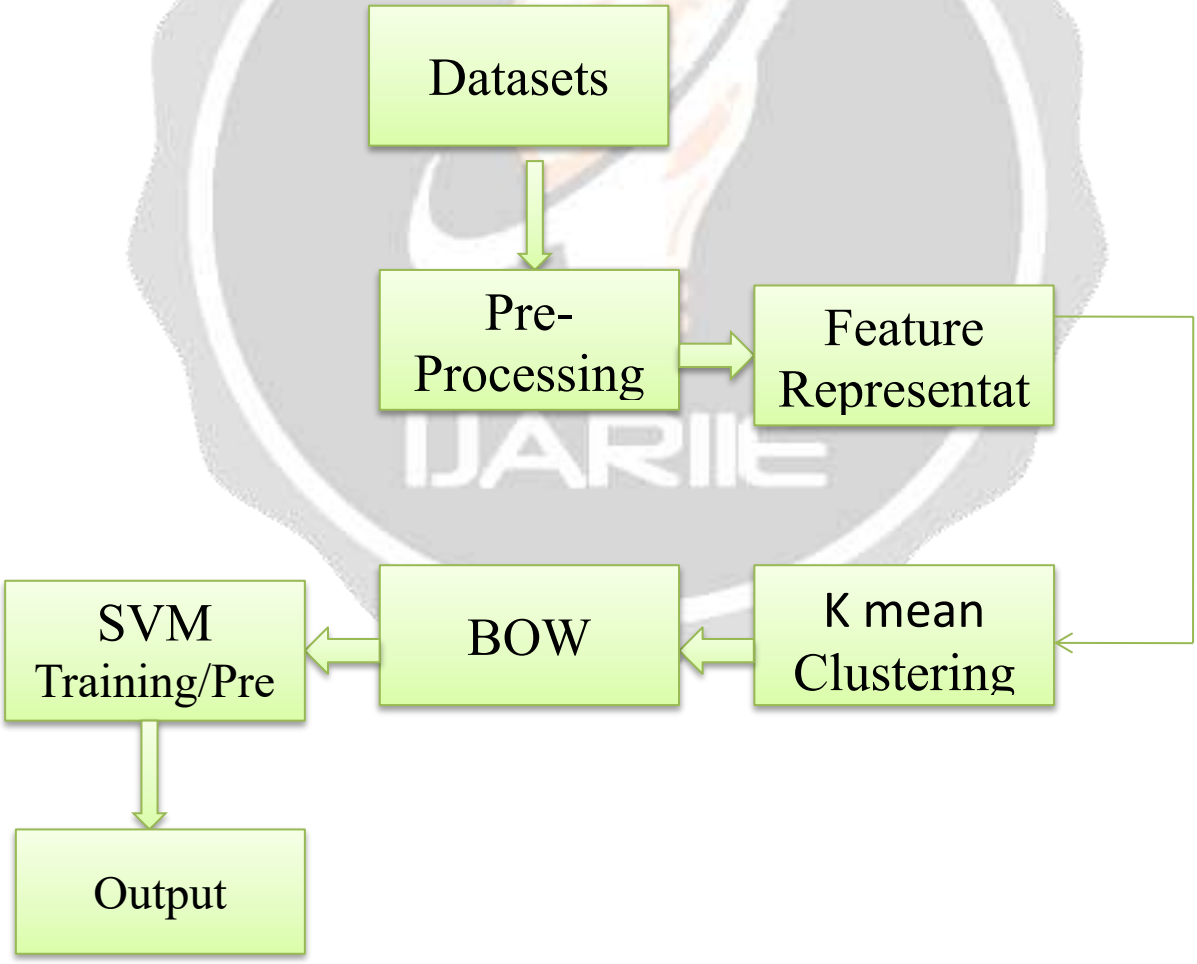

Fig 1: Proposed system architecture

Pre-processing:

A single record containing the gathered works of a creator in spite of the fact that are just keen on a solitary work. Or, on the other hand there might be given a vast work separated into volumes (this is the situation for Les Misèrables, as we will see later) where the division into volumes is not imperative to us.

If a long text is break up into (such as a book-length work) smaller chunks so we can get a sense of the variability in an author's writing. If comparing one group of writers to a second group, to sum a particular information about writers belonging to the same group. There is need for merging documents or other information that are initially separate. The section shows two common pre-processing step: splitting large texts into smaller "chunks" and aggregating texts together.

Another important pre processing step is tokenization. The process of splitting a text into individual words or sequences of words (*n-grams*). Decisions regarding tokenization will rely on the language(s) that are studied and the research question. For example, should the phrase "her father's arm-chair" be tokenized as as ["her", "father", "s", "arm", "chair"] or ["her", "father's", "arm-chair"]. Tokenization patterns that work for one language may not be appropriate for another (What is the appropriate tokenization of "Qu'est-ce que c'est?"?). The segment begins with a substantial discourse of tokenization before covering splitting and merging text.

Tokenizing:

Tokenization when put on to data security is the method of replacing a sensitive data element with a non-sensitive equivalent, mention to as a token , that has no extrinsic or exploitable meaning or value.

Stemming:

To count inflected forms of a word together. This procedure is referred to as *stemming*. Stemming a German text treat the given words as instances of the word "Wald": "Wald", "Walde", "Wälder", "Wäldern", "Waldes", and "Walds". Similar in English the words would be counted as "forest": "forest", "forests", "forested", "forest's", "forests'". As stemming lowers the number of unique vocabulary items that are needed to be tracked, it speeds up a variety of computational operations. For some kinds of analyses, such as authorship attribution or fine-grained stylistic analyses, stemming may obscure differences among writers. For example, an author might be distinguished by the use of a plural form of a word.

Chunking:

Splitting a long text into small samples is a very common task in text analysis. Most of the kinds of quantitative text analysis are take as inputs an unordered list of words, breaking a text into little pieces allows one to preserve context that would otherwise be removed; observing two words together in a paragraph-sized chunk of text tells us more about the relationship between those two words than observing two words occurring together in an 100,000 word book. As we will be using a selection of tragedies as examples, we may also consider the difference between knowing that two character names occur in the similar scene versus knowing that those two names occur in the ; same play.

Stopping:

Removal of stop words – Stop words like "and", "the", "of", etc are very usual in all English sentences and are not very relevant in deciding spam or legitimate status, so these words have been taken out from the emails.

Feature extraction process:

Once the dictionary is ready, we can remove word count vector (our feature here) of 3000 dimensions for every email of training set. Each **word count vector** has the frequency of 3000 words in the training file. Of course you might have guessed by now that most of them will be zero. Let us take an example. Suppose we have 500 words in our dictionary. Each word count vector contains the frequency of 500 dictionary words in the training file. Suppose text in training file was "Get the work done, work done" then it will be

[0,0,0,0,0,…….0,0,2,0,0,0,……,0,0,1,0,0,…0,0,1,0,0,……2,0,0,0,0,0]. Here, the word counts are placed at 296th, 359th, 415th, 495th index of 500 length word count vector and the remaining are zero. The python code will create a component vector grid whose lines signify 700 documents of preparing set and sections indicate 3000 expressions of word reference. The value at index '*ij'* will be the number of occurrences of jth word of dictionary in ith file.

K-means clustering :

K-means clustering is a method of vector quantization, originally from signal processing. K-means clustering aim is to partition observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a separate of the data space into Voronoi cells.

The problem is computationally hard however, there are efficient heuristic algorithms that are commonly employed and converge quickly to a local optimum .These are usually equal to the expectation-maximization for mixtures of Gaussian distribution via an iterative refinement approach employed by the both algorithms. K-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes. The algorithm has a free relationship to the k-nearest neighbor classifier, a famous machine learning technique for classification that is often confused with *k*-means because of the *k* in the name. One can put in the 1-nearest neighbor classifier on the cluster centers acquire by *k*-means to classify new data into the present clusters. This is known as nearest centered classified or Rocchio algorithm.

Training the classifiers:

The scikit-learn ML library for training classifiers. It is an open source python ML library which comes pack up in 3rd party distribution anaconda or can be used by separate installation following . Once it is installed, then it is imported in our program.

There are Naive Bayes classifier and Support Vector Machines (SVM). Naive Bayes classifier is a conventional and very popular method for document classification problem. It is a supervised probabilistic classifier based on Bayes theorem assuming independence between each pair of features. SVMs are supervised binary classifiers which are very structured when you have big number of features. The main goal of SVM is to separate some subset of training data from rest called the support vectors (boundary of separating hyper-plane). The decision function of SVM model that predicts the class of the test data is based on support vectors and makes use of a kernel trick. Once the classifiers are trained, check the performance of the models on test-set

4. RESULT

Experiment was carried out on dataset, which is publicly available. It is given below

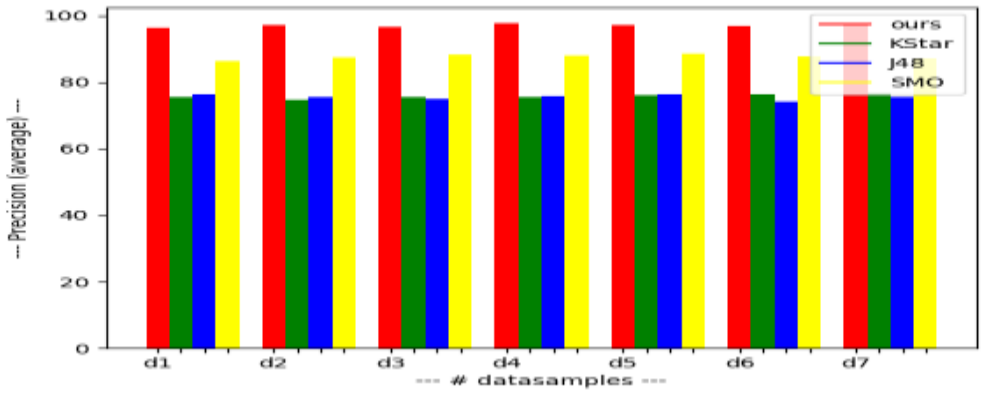

Chart 1 : Average precision graph

Datasets	**Methods**			
	OURS	**Kstar**	**J48**	**SMO**
d1	95	75	79	85
d2	96	76	78	86
d3	97	78	79	87
d4	98	77	78	88
d5	99	78	79	89
d6	98	76	74	88
d7	99	76	74	85

In this experiment we apply different classifier and gain more accurate classified instances by using the combination of clustering and classification algorithm. We produced more than 95% accurate result by implementing our proposed method. Figure represents the resultant values of above classified datasets using machine learning algorithm which was compared. It may differ according to it attributes of heart disease.

5. CONCLUSION

Heart diseases are complicated and take away lots of lives every year .When the early symptoms of heart diseases are ignored, the patient might end up with drastic consequences in a short span of time. Sedentary lifestyle and excessive stress in today's world have worsened the situation. If the disease is noticed first then it can be kept under control. However, it is always advisable to exercise daily and discard unhealthy habits at the earliest. Tobacco consumption and unhealthy diets increase the chances of stroke and heart diseases. Eating at least 5 helpings of fruits and vegetables a day is a good practice. For heart disease patients, it is advisable to restrict the intake of salt to one teaspoon per day.

Healthcare organizations are facing challenges to give cost-effective and high quality patient care. The administrators and clinicians both need to examine a wealth of data in the databases of healthcare information systems to find knowledge and to make informed decisions. This is basic specifically to improve the viability of sickness treatment and preventions. It becomes of most important in case of heart disease (HD) that is regarded as the main reason behind death in adults. Data mining serves as an analysis tool to discover unapparent or hidden relationships and patterns in HD medical.

Classification techniques are accomplished of processing a large amount of data. It is one of the most widely used methods of Data Mining in Healthcare organization. The widespread classification techniques used in risk prediction of heart disease are Bayesian Networks, Artificial Neural Network, Nearest Neighbour method, Fuzzy logic, Fuzzy based Neural Networks, Decision trees, Genetic Algorithms and Support Vector Machines18. Also, applying hybrid data mining techniques has revealed promising results in the diagnosis of heart disease. There is being required to have reliable model for predicting the existence or absence of heart disease with known and unknown risk factors. Some time poor clinical decisions lead to mortality. And all clinicians are not equally good in predicting the heart disease. In the case of heart disease time is precious, proper risk identification at the right time saves life of many patients.

Various techniques has been proposed to predict heart conditions using different mining techniques as well as using different datasets, but after in depth analysis , it seems that bag of words approach using k-mean clustering along with machine learning algorithm like , S.V.M, Random forest, Logistic Regression, etc, will give different and efficient results. Online datasets are collecting for exactingly monitored symptoms and various other health related parameters like B.P, pulse rate, diabetic condition etc, and pre-process and filter the data as needed and apply the appropriate machine learning algorithm for predictive analysis of testing data.

A methodology to detect an heart attack on the basis of available patient dataset which contains various symptoms and other health related parameters like B.P, H.R, etc. Numeric feature representation technique for feature extraction and K-mean clustering algorithm for clustering feature vectors, in various clusters and machine learning technique to make predictive analysis of heart condition.

6. REFRENCES

[1] Indira S. Fal Dessai,"Intelligent Heart Disease Prediction System Using Probabilistic Neural Network", International Journal on Advanced Computer Theory and Engineering (IJACTE),2013,Vol. 2, Issue 3, pp. 38-44.

[2] M. Shouman, T. Turner and R. Stocker, "Integrating decision tree and k-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients," in Proc. of Int. Conf. on Data Mining, Australian Defence Force Academy Northcott Drive, Canberra, 2012, pp. 1-7.

[3] G. Purusothaman, and P. Krishnakumari ,"A Survey of Data Mining Techniques on Risk Prediction: Heart Disease", Indian Journal of Science and Technology**,** June 2015,Vol. 8(12), DOI:10.17485/ijst/2015/v8i12/58385, pp. 1-5.

[4] Shaikh Abdul Hannan, A.V. Mane, R. R. Manza,and R. J. Ramteke, Dec 2010,"Prediction of Heart Disease Medical Prescription using Radial Basis Function",IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), DOI: 10.1109/ICCIC.2010.5705900 ,28-29 .

[5] Mrudula Gudadhe, Kapil Wankhade, and Snehlata Dongre, Sept 2010,"Decision Support System for Heart Disease Based on Support Vector Machine and Artificial Neural Network", International Conference on Computer and Communication Technology (ICCCT),DOI:10.1109/ICCCT.2010.5640377, 17-19.

[6] K. S. Kavitha, K. V. Ramakrishnan, and Manoj Kumar Singh, September "Modelling and Design of Evolutionary Neural Network for Heart Disease Detection" ,International Journal of Computer Science Issues (IJCSI),2010, Vol. 7, Issue 5, pp. 272-283

[7] Prajakta Ghadge,Vrushali Girme, Kajal Kokane, and Prajakta Deshmukh, "Intelligent Heart Attack Prediction System Using Big Data", International Journal of Recent Research in Mathematics Computer Science and Information Technology,Vol. 2, Issue 2, pp.73-77, October 2015–March.

[8] Shantakumar B. Patil,and Dr. Y. S. Kumaraswamy, "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction", IJCSNS International Journal of Computer Science and Network Security, February 2009,Vol. 9 , No. 2,pp. 228-235.

[9] Sairabi H. Mujawar, and P. R. Devale, "Prediction of Heart Disease using Modified k-means and by using Naive Bayes", International Journal of Innovative Research in Computer and Communication Engineering(An ISO 3297: 2007 Certified Organization) October 2015,Vol. 3, Issue 10, pp. 10265-10273.

[10] S. Suganya, and P.Tamije Selvy, "A Proficient Heart Disease Prediction Method using Fuzzy-Cart Algorithm", International Journal of Scientific Engineering and Applied Science (IJSEAS), January 2016,Vol. 2, Issue1,ISSN: 2395-3470.

[11] S.Florence, N. G. Bhuvaneswari Amma, G. Annapoorani, and K. Malathi, "Predicting The Risk of Heart Attacks using Neural Network and Decision Tree", International Journal Of Innovative Research In Computer And Communication Engineering, ISSN (Online): 2320-9801, November 2014, Vol. 2, Issue 11, pp. 7025-7028.

[12] K Cinetha, and Dr. P. Uma Maheswari, "Decision Support System for Precluding Coronary Heart Disease using Fuzzy Logic.", International Journal of Computer Science Trends and Technology (IJCST), Mar.-Apr. 2014, Vol. 2, Issue 2, pp. 102-107

[13] Asha Rajkumar, G.Sophia Reena, "Diagnosis Of Heart Disease Using Data mining Algorithm", Global Journal of Computer Science and Technology 38 Vol. 10 Issue 10 Ver.1.0 Septembe2010.

[14] K. Srinivas, B.K. Rani and D.A. Govrdhan, "Application of data mining techniques in healthcare and prediction of heart attacks." International Journal on Computer Science and Engineering,2011, vol. 2, no. 2, pp. 250-255.

[15] Dani Yogatama and Noah A. Smith, "Making the Most of Bag of Words: Sentence Regularization with Alternating Direction Method of Multipliers", Proceedings of the 31st International Conference on Machine Learning, Beijing, China, JMLR: W&CP, 2014,volume 32.

[16] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart diseases through neural network ensembles".Expert Syastem with Applications,vol.36,pp.7675-7680,2009.