# Heart Disease Diagnosis Using Machine Learning Algorithm

Shashikant U. Ghumbre[1] and Ashok A. Ghatol[2]

[1] Computer Engineering Department,
College of Engineering Pune, Pune, Maharashtra, India
shashi.ghumbre@gmail.com
[2] Dr. B.A.T.University, Lonere, Maharshtra, India
vc_2005@rediffmail.com

**Abstract.** Recent advances in computing and developments in technology have facilitated the routine collection and storage of medical data that can be used to support medical decisions. However, in most countries, there is a first need for collecting and organizing patient's data in digitized form. Then, the collected data are to be analyzed in order for a medical decision to be drawn, whether this involves diagnosis, prediction, course of treatment, or signal and image analysis. In this paper, India centric dataset is used for Heart disease diagnosis. The correct diagnosis performance of the automatic diagnosis system is estimated by using classification accuracy, sensitivity and specificity analysis. The study shows that, the SVM with Sequential Minimization Optimization learning algorithm have better choice for medical disease diagnosis application.

**Keywords:** Decision Support System, Support Vector Machine, Heart Disease, Machine Learning.

## 1 Introduction

The application of machine learning methods in medical field is the subject of considerable ongoing research, which mainly concentrates on modeling some of the human actions or thinking processes and recognizing diseases from a variety of input sources. Other application areas are knowledge discovery [10] and biomedical systems, which include genetics and DNA analysis [1, 12]. The design may also be influenced by the desired performance on one or more specific classes of the problem instead of the overall performance. This is usual in most medical tasks as a different degree of significance may be required for the system's performance on each class. The computer programs or machine learning techniques can be used to reduce the mortality rate, improve the accuracy in disease diagnosis and mainly reduce the diagnosis time. The advancement in computer technology and communication encourages health-care providers to work using the Internet or Telemedicine technology [9,13,14].

In a medical diagnosis problem, what is needed is a set of examples or attributes that are representative of all the variations of the disease. The examples need to be selected very carefully if the system is to perform reliably and efficiently. The fact that there is no need to provide a specific algorithm on how to identify the disease, presents a major advantage over the application of machine learning methods to this

S.C. Satapathy et al. (Eds.): Proceedings of the InConINDIA 2012, AISC 132, pp. 217–225.
springerlink.com

type of problems. However, development of artificial intelligence systems for medical decision making problems is not a trivial task. Difficulties include the acquisition, collection and organization of the data that will be used for training the system. This becomes a major problem especially when the system requires large data sets over long periods of time, which in most cases are not available due to the lack of an efficient recording system. The above mentioned problems or the existing procedures involved in the medical task may not be the only factors affecting the design of a Decision Support System (DSS). The design may also be influenced by the desired performance on one or more specific classes of the problem instead of the overall performance. This is usual in most medical tasks as a different degree of significance may be required for the system's performance on each class. For example, in a heart disease diagnosis task, it is necessary for the accuracy on healthy patients to be as high as possible, as a misclassification in this category may result in a healthy patient going under treatment for no reason. The balance of the system's performance between different classes could vary and is largely dependent on the medical problem itself and the collected data. In addition, in most of the countries, insufficient numbers of medical specialist have increased the mortality of patients suffering from various diseases. Heart diseases have emerged as the number one killer in both urban and rural areas in most of the countries. As of 2010, it is the leading cause of death in the U.S., England and Canada, accounting for 25.4% of the total deaths in the United States. Similar situation is found rest of the countries all over the world. In case of heart disease time is very crucial to get correct diagnosis in early stage[37]. It is observed that, in many cases due to wrong diagnosis or trial/error procedure for diagnosis leads to patient health compromise. The dearth of medical specialists and/or wrong diagnosis procedure will never be overcome within a short period of time [3,4]. Patient having chest pain complaint may undergo unnecessary treatment or admitted in the hospital. In most of the developing countries specialists are not widely available for the diagnosis. Hence, such automated system can help to medical community to assist doctor for the accurate diagnosis well in advance.

The rest of the paper is organized as follows: Section 2 briefly reviews some prior works on machine learning techniques in medical Diagnosis. Section 3 briefly describes the heart disease diagnosis and the proposed Decision Support System (DSS) and its techniques used are discussed. Section 4 details the use of Support Vector Machine in medicine. The experimental results are given in Section 5. Section 6 concludes the paper.

## 2 Literature Review

Zhi-Hua Zhou and Yuan Jiang [4] have proposed an approach named C4.5 Rule-PANE, which gracefully combines the advantages of artificial neural network ensemble and rule induction. A specific rule induction approach, i.e. C4.5 Rule, is used to learn rules from the new training data set. Case studies on diabetes, hepatitis, and breast cancer show that C4.5 Rule-PANE could generate rules with strong generalization ability, which profits from artificial neural network ensemble, and strong comprehensibility, which profits from rule induction.

Leung et al., [32] have presented a data mining framework for biological data sets. And it has been applied to the Hepatitis B Virus DNA data sets which are real world data. Their method has good performance using the fuzzy measure and the nonlinear integral, since the non additivity of the fuzzy measure reflects the importance of the feature attributes, as well as their inherent interactions.

Saangyong Uhmn et al., [34] have presented the machine learning techniques, SVM, decision tree, and decision rule to predict the susceptibility of the liver disease, chronic hepatitis from single nucleotide polymorphism (SNP) data. The experimental results have shown that decision rule is able to distinguish chronic hepatitis from normal with the maximum accuracy of 73.20%, whereas SVM is with 67.53% and decision tree is with 72.68%. It is also shown that decision tree and decision rule is potential tools to predict the susceptibility to chronic hepatitis from SNP data.

Ozyilmaz, L. and Yildirim, T [35] have presented three neural network algorithms for diagnosis of hepatitis diseases. The results were compared with some statistical methods used in a previous work. Their results have shown that using a hybrid network CSFNN that combines MLP and RBF is more reliable for the diagnosis. Thus the compared results have shown that neural networks can be used in the problem of diagnosis for hepatitis diseases as efficiently as statistical methods.

# 3 Heart Disease Diagnosis

## 3.1 Heart Disease

The heart, shown in Figure 1, is actually two separate pumps: a *right heart* that pumps blood through the lungs, and a *left heart* that pumps blood through the peripheral organs. In turn, each of these hearts is a pulsatile two-chamber pump composed of an *atrium* and a *ventricle*. Each atrium is a weak primer pump for the ventricle, helping to move blood into the ventricle. The ventricles then supply the main pumping force that propels the blood either through the pulmonary circulation by the right ventricle or through the peripheral circulation by the left ventricle.

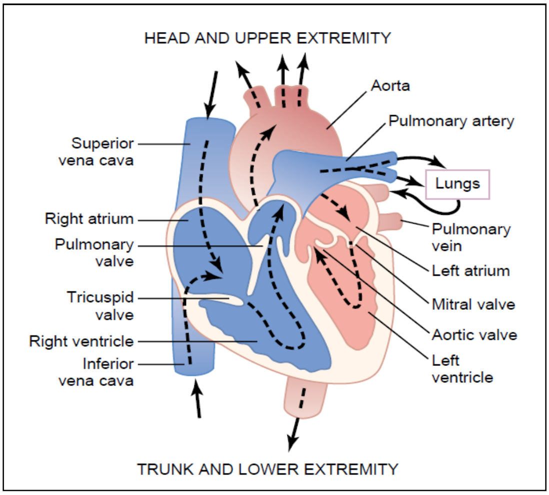

**Fig. 1.** Structure of the heart, and course of blood flow through the heart chambers and heart valves

The cardiac events that occur from the beginning of one heartbeat to the beginning of the next are called the *cardiac cycle.* The cardiac cycle consists of a period of relaxation called *diastole,* during which the heart fills with blood, followed by a period of contraction called *systole.* Blood normally flows continually from the great veins into the atria; about 80 per cent of the blood flows directly through the atria into the ventricles even before the atria contract. Then, atrial contraction usually causes an additional 20 per cent filling of the ventricles. Therefore, the atria simply function as primer pumps that increase the ventricular pumping effectiveness as much as 20 per cent. When the atria fail to function, the difference is unlikely to be noticed unless a person exercises; then acute signs of heart failure occasionally develop, especially shortness of breath. The *P, Q, R, S,* and *T waves in* electrocardiogram are electrical voltages generated by the heart and recorded by the electrocardiograph from the surface of the body [36]. The choice of which tests to perform depends on several things. These include a patient's risk factors, history of heart problems, current symptoms and the doctor's interpretation of these factors. The cardiovascular problems may experienced with ordinary physical activity causes symptoms: Undue fatigue, Palpitations-the sensation that heart is skipping a beat or beating too rapidly, Dyspnea-difficult or labored breathing, Anginal pain -chest discomfort from increased activity. The doctor diagnose the heart attack with, review the patient's complete medical history, physical examination, use an electrocardiogram (ECG) to discover any abnormalities caused by damage to the heart, sometimes use a blood test to detect abnormal levels of certain enzymes in the bloodstream. Blood tests confirm (or refute) suspicions raised in the early stages of evaluation that may occur in an emergency room, intensive care unit or urgent care setting. These tests are sometimes called heart damage markers or cardiac enzymes [37].

## 3.2 Automated Heart Disease Diagnosis

The first stage of this diagnosis study decodes the real and typical doctor-based diagnosis procedure for the particular Heart disease. During data collection, number of visits to the hospitals and discussions with medical practitioners has been carried out. Table: 1 shows the Heart disease symptoms and tests which are usually observed during the diagnosis.

**Table 1.** Heart Disease Attributes

| Symptoms and signs | Test results |
|---|---|
| Chest pain Types(Left and Right), Arm pain, backache, Sweating, Breathlessness, addiction, Diabetic, MAP, Pulse rate | ECG: ST Elevation, ST Depression, T Elevation, T Depression, Q waves, BSL, CK-MB |

Based on medical records 214 instances in total are selected for the automated diagnosis. Dataset consists of 19 different attributes with four class distribution: 0-Myalgia, 1- Myocardial Infarction (MI), 2- Ischemic Heart Disease (IHD), 3-Unstable Angina (UA). For this study binary classification problem is considered for

Heart disease diagnosis, in which 78 instances are belongs to 0 i.e. Myalgia (Normal), and 139 are considered as 1 i.e. Patient having Heart disease. In order to provide physicians with both structured questions and structured responses within medical domains of specialized knowledge or experience [17,25,26] medical expert systems have been developed. The advice of one or more medical experts, who also suggest the optimal questions to be considered, and provide the most accurate conclusions to be drawn from the answers the physician chooses, is used to embody the structure in the program. A trained Support Vector Machine (SVM), Multilayer Perceptron (MLP), Radial Basis Function Neural Network (RBF) and other methods are used to assume the evolution of the biological indicators. Once the patients' personal data is presented along with the results of the tests taken at the onset of the treatment and the postulated code of reaction, the evolution in time of the illness can be specified by the expert system.

# 4 Support Vector Machine

Support Vector Machine (SVM) is a category of universal feed forward networks like Radial-basis function networks, pioneered by Vapnik. SVM can be used for pattern classification and nonlinear regression. More precisely, the support vector machine is an approximate implementation of the method of structural risk minimization. This principle is based on the fact the error rate of a learning machine on test data is bounded by the sum of the training-error rate and term that depends on the Vapnik-Chervonenkis (VC) dimension [33]. The support vector machine can provide good generalization performance on pattern classification problem.

## 4.1 Optimal Hyperplane for Patterns

Consider the training sample $\{(\mathrm{xi}, yi)\}_{i=1}^{N}$ where xi is the input pattern for the ith instance and $y_i$ is the corresponding target output. With pattern represented by the subset $y_i= +1$ and the pattern represented by the subset $y_i= -1$ are linearly separable. The equation in the form of a hyperplane that does the separation is

$$w^T x + b = 0 \tag{1}$$

where, x is an input vector, w is an adjustable weight vector, and b is a bias. Thus,

$$w^T x_i + b \geq 0 \quad \text{for} \quad y_i = +1 \tag{2}$$

$$w^T x_i + b < 0 \quad \text{for} \quad y_i = -1 \tag{3}$$

for a given weight vector w and a bias b, the separation between the hyperplane defined in eq. 1 and closest data point is called the margin of separation, denoted by ρ as shown in figure 2, the geometric construction of an optimal hyperplane for a two-dimensional input space.

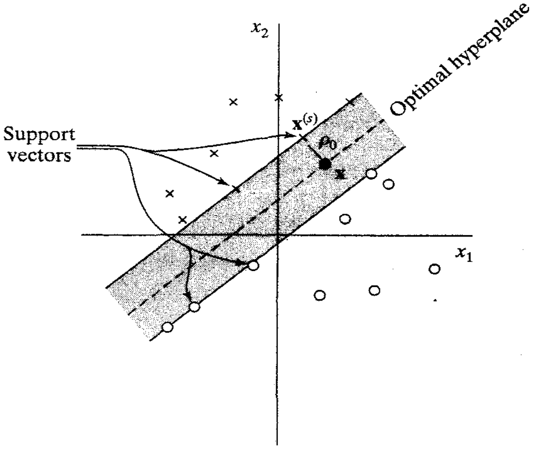

**Fig. 2.** Optimal Hyperplane for a two dimensional input space

The discriminant function gives an algebraic measure of the distance from x to the optimal hyperplane for the optimum values of the weight vector and bias, respectively.

$$g(x) = w_o^T x + b_o \tag{4}$$

## 5 Experimentations and Results

The diagnosis of the disease for a new patient to be performed on basis of dataset is facilitated by the first stage of data preparation and feature extraction. Initially, network size, sample size, model selection, and feature extraction are considered as key parameters for the study of a practical network design issues related to learning and generalization. During experimentation, it is observed that for such type of nonlinear classification problem unsupervised learning procedure for feature extraction is not appropriate because of nonlinear correlation structure. It is also observed that, correct and complete data collection procedure is the proper route for the selection of best classifier. Heart disease dataset is used for diagnosis with 214 instances of 3 types of predicted heart diseases. For binary classification problem with limited data size it is necessary to validate networks model with cross validation, hence 5-fold and 10-fold cross validation techniques are used on heart disease dataset. Table 2 and 3 shows comparative results of SVM, MLP, RBF, BayesNet, J48 and Rule, in which SVM gives promising results using 5-fold and 10-fold cross validation.

**Table 2.** Generalization performance using 5-fold cross validation

| Classifier | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| SVM | 85.51% | 84.60% | 88.50% |
| MLP | 82.71% | 85.30% | 78.20% |
| RBF | 82.24% | 82.40% | 82.10% |
| BayesNet | 79.90% | 77.20% | 84.60% |
| J46 | 71.43% | 73.54% | 70.10% |
| Rule | 68.90% | 72.81% | 69.94% |

**Table 3.** Generalization performance using 10-fold cross validation

| Classifier | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| SVM | 85.05% | 84.60% | 85.90% |
| MLP | 84.11% | 87.50% | 78.20% |
| RBF | 82.71% | 83.10% | 82.10% |
| BayesNet | 80.37% | 77.20% | 85.90% |
| J48 | 76.65% | 73.80% | 74.10% |
| Rule | 71.16% | 67.90% | 72.80% |

## 6 Conclusions

Medical diagnosis has become highly attributed with the development of technology lately. Furthermore the computer and communication tools have improved the medical practice implementation to a greater extent. Though the Artificial neural network ensemble is a powerful learning technique that could aid in the remarkable improvement in the generalization ability of neural learning systems its lucidity is worse than that of a single artificial neural network thereby deterring its wide recognition among the medical practitioners. In this paper we have projected a Decision Support System for the diagnosis of Heart disease by means of radial basis function network structure and Support Vector Machine. Therefore the diagnosis of Heart disease is carried out utilizing different data samples from diverse patients and the results have denoted that SVM with Sequential minimize optimization is equivalently good as the ANN and other models in the diagnosis of Heart disease. The classification accuracy, sensitivity, and specificity of the SVM has been found to be high thus making it a good option for the diagnosis.

## References

1. Brause, R.W.: Medical Analysis and Diagnosis by Neural Networks. In: Proceedings of Medical Data Analysis, October 8-9, vol. 20, pp. 1–13. Springer, Heidelberg (2001)
2. Gerard Wolff, J.: Medical diagnosis as pattern recognition in a framework of information compression by multiple alignment, unification and search. Decision Support Systems 42(2), 608–625 (2006)

3. Steimann, F., Adlassnig, K.P.: Fuzzy medical diagnosis. In: Rupini, E., Bonissone, P., Pedrycz, W. (eds.) Handbook of Fuzzy Computation. Oxford University Press and Institute of Physics Publishing, Bristol (1998)
4. Zhou, Z.-H., Jiang, Y.: Medical Diagnosis with C4.5 Rule Proceeded by Artificial Neural Network Ensemble. IEEE Transactions on Information Technology in Biomedicine 7(1), 37–42 (2003)
5. Richards, G., Rayward-Smith, V.J., Sönksen, P.H., Carey, S., Weng, C.: Data mining for indicators of early mortality in a database of clinical records. Artificial Intelligence in Medicine 22(3), 215–231 (2000)
6. Ćosić, D., Lončarić, S.: Rule-Based Labeling of CT Head Image. In: Proceedings of the 6th European Conf. of AI in Medicine Europe AIME 1997, pp. 453–456 (1997)
7. Duch, W., Adamczak, R., Grąbczewski, K., Żal, G., Hayashi, Y.: Fuzzy and crisp logical rule extraction methods in application to medical data. In: Szczepaniak, P.S., Lisboa, P.J.G., Kacprzyk, J. (eds.) Fuzzy systems in Medicine, pp. 593–616. Physica - Verlag, Springer, Heidelberg (1999)
8. Stevens, A., Lowe, J.S., Young, B.: Wheater's Basic Histopathology: a color atlas and text, 4th edn., p. 315. Churchill Livingstone (2003) ISBN-10 0-4430-7001-6
9. Manickam, S., Abidi, S.S.R.: Experienced Based Medical Diagnostics System Over The World Wide Web (WWW). In: Proceedings of The First National Conference on Artificial Intelligence Application In Industry, Kuala Lumpur, pp. 47–56 (1999)
10. Alexopoulos, E., Dounias, G.D., Vemmos, K.: Medical Diagnosis of Stroke Using Inductive Machine Learning. In: Machine Learning and Applications: Machine Learning in Medical Applications, Chaina, Greece, pp. 20–23 (1999)
11. Shortliffe, E.H.: Computer Programs to Support Clinical Decision Making. Readings in Uncertain Reasoning, 161–166 (1990) ISBN:1-55860-125-2
12. Neves, J., Alves, V., Nelas, L., Romeu, A., Basto, S.: An Information System, That Supports Knowledge Discovery and Data Mining in Medical Imaging. In: Machine Learning and Applications: Machine Learning in Medical Applications, Chaina, Greece, pp. 37–42 (1999)
13. Cunningham, P., Carney, J., Jacob, S.: Stability problems with artificial neural networks and the ensemble solution. Artificial Intelligence in Medicine 20(3), 217–225 (2000)
14. Kononenko: Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine 23(1), 89–109 (2001)
15. Zhou, Z.-H., Jiang, Y., Yang, Y.B., Chen, S.F.: Lung cancer cell identification based on artificial neural network ensembles. Artificial Intelligence in Medicine 24(1), 25–36 (2002)
16. Pranckeviciene, E.: Finding Similarities Between An Activity of the Different EEG's by means of a Single layer Perceptron. In: Machine Learning and Applications: Machine Learning in Medical Applications, Chaina, Greece, pp. 49–52 (1999)
17. Luger, G.F., Stubblefield, W.A.: Artificial intelligence and the design of expert systems. Benjamin/Cummings Publ. Co., Redwood City (1989)
18. Dimitrios Siganos: Neural Networks in Medicine (1995), from `http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol2/ds12/article2.html`
19. Wolff, J.G.: Medical Diagnosis as Pattern Recognition in a Framework of Information Compression by Multiple Alignment, Unification and Search. Decision Support Systems 42(2), 608–625 (2006)
20. Sasikala, K.R., Petrou, M., Kittler, J.: Fuzzy classification with a GIS as an aid to decision making. EARSeL Advances in Remote Sensing 4, 97–105 (1996)
21. Tou, J.T., Gonzalez, R.C.: Pattern Recognition Principles. Addison – Wesley, Reading (1974)
22. Moody, J., Darken, C.J.: Fast learning in networks of locally tuned processing units. Neural Computation 2, 281–294 (1989)

23. Hanson, S.J., Burr, D.J.: Minkowski-r back propagation: learning in connectionist models with non-Euclidean error signals. In: Neural Information Processing Systems, pp. 348–357. American Institute of Physics, New York (1988)
24. Poggio, T., Girosi, F.: Regularization algorithms for learning that are equivalent to multilayer networks. Science 247, 978–982 (1990)
25. Poggio, T., Girosi, F.: Networks for approximation and learning. Proceedings of the IEEE 78(9), 1481–1497 (1990)
26. Hartman, E.J., Keeler, J.D., Kowalski, J.M.: Layered neural networks with Gaussian hidden units as universal approximators. Neural Computation 2, 210–215 (1990)
27. Park, J., Sandberg, I.W.: Universal approximation using radial basis function networks. Neural Computation 3, 246–257 (1991)
28. Park, J., Sandberg, I.W.: Approximation and radial basis function networks. Neural Computation 5, 305–316 (1993)
29. Venkatesan, P., Anitha, S.: Application of a radial basis function neural network for diagnosis of diabetes mellitus. Current Science 91(9), 1195–1199 (2006)
30. Lin, T.-C., Kuo, M.-J., Chen, Y.-C.: Frequency Domain Analog Circuit Fault Diagnosis Based on Radial Basis Function Neural Network. In: International Conference on Communications, Circuits and Systems, ICCCAS 2004, June 27-29, vol. 2, pp. 1183–1185 (2004)
31. Bhatia, S., Prakash, P., Pillai, G.N.: SVM based Decision Support System for Heart Disease Classification with Integer-coded Genetic Algorithm to select critical features. In: Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA, pp. 34–38 (2008)
32. Leung, K.S., Ng, Y.T., Lee, K.H., Chan, L.Y., Tsui, K.W., Mok, T., Tse, C.H., Sung, J.: Data Mining on DNA Sequences of Hepatitis B Virus by Nonlinear Integrals. In: Proceedings Taiwan-Japan Symposium on Fuzzy Systems & Innovational Computing, 3rd Meeting, Japan, pp. 1–10 (August 2006)
33. Haykin, S.: Neural Networks, 2nd edn. Prentice-Hall (2003)
34. Uhmn, S., Kim, D.-H., Kim, J., Cho, S.W., Cheong, J.Y.: Chronic Hepatitis Classification Using SNP Data and Data Mining Techniques. In: Frontiers in the Convergence of Bioscience and Information Technologies, FBIT 2007, October 11-13, pp. 81–86 (2007)
35. Ozyilmaz, L., Yildirim, T.: Artificial neural networks for diagnosis of hepatitis disease. In: Proceedings of the International Joint Conference on Neural Networks, July 20-24, vol. 1, pp. 586–589
36. Gyton, C., Hall, J.E.: Textbook of Medical Physiology, 11th edn. Elsevier (2006)
37. http://www.heart.org