# Performance of SVM Classifiers in Predicting Diabetes

Rahul Samant,
*SVKM'S NMIMS, Shirpur Campus, India;*

Srikantha Rao,
*TIMSCDR, Mumbai University, Kandivali, Mumbai, India,*

## Abstract

*This paper investigates the ability of several models of Support Vector Machines (SVMs) with alternate kernel functions to predict the probability of occurrence of Diabetes (DT) in a mixed patient population. To do this a SVM was trained with 13 inputs (symptoms) from the medical dataset obtained from a university hospital. Different kernel functions, such as Linear, Quadratic, Polyorder (order three), Multi Layer Perceptron (MLP) and Radial Basis Function kernel (RBF) were coded and tested to build the medical diagnosis system (MDS). A detailed database, comprising of healthy and diabetic patients from a university hospital was used for training the SVM for prediction of diabetes. All kernel functions for SVM models showed reasonably good accuracy in prediction of disease (s), with linear kernel structure showing best prediction in 3 out of 4 datasets and Polyorder in one database. Thus the best choice appears to be situation specific.*



## 1. Introduction

Diabetes is caused by the inability of body to produce or respond properly to insulin which is needed to regulate glucose. [1]. Besides contributing to heart disease, diabetes also increases the risks of developing kidney disease, blindness, nerve damage, and blood vessel damage. Diabetes disease diagnosis via proper interpretation of the symptoms data is an important classification problem. In the practice of medicine, Support Vector Machines (SVM) are now being vigorously applied in areas as diverse as cardiology, analysis of images (ECG, EEG and SPECT), cytology, genetics and clinical chemistry. [2] However the major problem here is to diagnose a disease with high reliability. Although earlier identification of this disease is gaining importance in clinical research, the investigation of factors for prevention and intervention are also crucial issues in preventive medicine. Modifiable factors such as life-style variables and body measurements, for reducing risk of the disease are especially interesting for public health professionals. [3]. Human experts make mistakes because of their diverse experiences, training and other limitations. Brause highlighted that almost all physicians are confronted during their training by the imposing task of learning to diagnose. Here, they have to solve the problem of deducing certain diseases or formulating a treatment based on more or less specified observations and knowledge. For this task, certain basic difficulties have to be taken into account. Principally, human thinker does not resemble statistical modeling as done with computers but it acts like pattern recognition systems. In particular, it has been repeatedly noted that Humans can recognize patterns or objects very easily but fail when probabilities have to be assigned to different outcomes, here diagnosis. In reality, the quality of diagnosis is totally depended on the physician talent and experience. Further, emotional problems and fatigue degrade the doctor's performance. The training procedure of doctors, in particular specialists, is a lengthy and expensive one. So even in developed countries one may sometimes feel the lack of sound medical advice. Medical science is one of the most rapidly growing and changing fields of science. Therefore like many other endeavors of modern enterprise, today knowledge-based technology is being increasingly engaged in the field of medical diagnosis [4]. It has also been shown that employing computer aided diagnostic systems (CAD) as a "second opinion" has lead to improved diagnostic decisions and support vector machines (SVMs) have shown remarkable success in this area [5].

## 2. Literature Review

Cheng et al. [6] investigate clinical diagnosis through case study. SVM classifier with a radial basis inner function was established to predict and discriminate some unknown patients in the study. At the same time, Bayes angle discriminated model and Logistic regression, which are traditional statistical classification approaches, are set up to compare with SVM methods in POAG diagnosis. In the end, they conclude that SVM method is reliable and superior in many respects to statistical classification methods in the POAG recognition.
Ghumbre et al. [7] presented an intelligent system based support vector machine along with a radial basis function network for the diagnosis. Expert system based on clinical symptoms is used to

decide what type of heart disease is possible to appear for a patient, whether it is heart attack or not. The support vector machine with sequential minimal optimization algorithm is applied to India based patients' data set. Then, the Radial Basis Function (RBF) network structure trained by Orthogonal Least Square (OLS) algorithm is applied to same data set for predictions. Results obtained show that support vector machine can be successfully used for diagnosing heart disease.

Peng et al. [8] designed a cancer predicting medical diagnosis system. They used univariate analysis to analyze the relationship between the six image indicators. A SVM model was built with these indicators as input index. The output index was that lymph node metastasis of the patient was positive or negative. It was confirmed by the surgery and histopathology. A standard machine-learning technique called k-fold cross-validation (5-fold) was used to train and test SVM models. The diagnostic capability of the SVM models in lymph node metastasis was evaluated with the receiver operating characteristic (ROC) curves.

Kampourakia et al. [9] proposed-- e-doctor; a web-based application that makes automatic diagnoses about health problems using SVMs. The system was trained by symptoms and diagnosis of a health problem. The system then made an automatic diagnosis/prediction by means of answering if the patient has (or may have in the future) a specific health problem. The application can be used in cases where statistical information plays a vital role on deciding about a patient's condition. A prototype was developed and the system trained and tested for the case of heart symptoms. The results were satisfactory.

Huang et al. [10] presented a computer aided diagnosis (CAD) system with textural features for classifying benign and malignant breast tumors on medical ultrasound systems. A series of pathological proven breast tumors were evaluated using the SVMs in the differential diagnosis of the breast tumors. The proposed system used facial textural features. The proposed CAD system showed high accuracy in detecting the breast tumors.

Priya et al. [11] developed two models like Probabilistic Neural network (PNN) and Support vector machine (SVM) for the diagnosis of Diabetic Retinopathy. Experimental results were compared for accuracy and proven that SVM model outperforms the other model.

Huang et al. [12] provided a survey using Support Vector Machine (SVM) to predict and assess metabolic functions of diabetes based on bio-heat transfer theory and infrared thermal imaging technology. Two metabolic characteristic values, metabolic function parameter and blood perfusion rate, are extracted from thermography data of cold water stimulation experiment as inputs of SVM to set up models by different kernel functions. The system reported satisfactory classification accuracy.

Bagchi et al. [17] extends the utility of asymmetric soft margin support vector machines—by marshalling decision theoretic notions. It was argued that this could be a more realistic approach to set the asymmetric classification boundaries of a soft margin SVM. The commonly used assumption of equal proportion of misclassifications in each class basis of assigning penalty costs for margin errors and misclassifications were prejudiced. Additionally, the effect of input data quality on the performance of SVM is analytically investigated**.**

## 3. The SVM Classification Methodology

Since 1980 as the power of computing began to grow, automated learning aimed at modeling and understanding relationships among a set of variables derived from objects drew much interest [13]. The goal became that of using supervised learning to model the relationship between some selected inputs and outputs. Artificial Neural Nets (ANN) and Support Vector Machines (SVM) are two such devices created in that period and these continue even today as state-of-the-art classification methods. Of late, in the last twenty or so years, SVM has been extensively used to target problems of classification where an input-output training dataset is presented to the algorithm, which in turn, when its learning is complete, becomes capable of classifying *yet new* input data. The first such procedure was designed and presented by Vapnik and his colleagues [5, 13]. Most work on SVM and its applications have focused on the two-class pattern classification problem [14].

Briefly, the two-class SVM classifier may be described as follows, though comprehensive references on it are already extensive [13, 15]. This summary is based on book by Vapnik, *Statistical Learning Theory* [13].

Let vector **x** of inputs be a pattern that we need to classify and let y (a scalar) denote its assigned class label, ±1. Let $\{(\mathbf{x}_i, y_i), i = 1, 2, \ldots l\}$ be the training examples based on *l* patterns classified earlier by examining each example and tagging or labeling it as "+1" or "-1" earlier. The SVM's learning task then becomes constructing the classifier or a decision function $f(\mathbf{x})$ that would be able to correctly classify a new input pattern **x** not included in the training set. Such classifiers may be linear, or nonlinear.

If the training dataset *is* linearly separable, there will exist a linear function or hyperplane of the form

$$f(\mathbf{x}) = \boldsymbol{w}^T\mathbf{x} - \gamma \qquad \text{----------(1)}$$

such that for each training example $\mathbf{x_i}$ the function yields $f(\mathbf{x}) \geq 0$ whenever $y_i = +1$, and $f(x) < 0$ when $y_i = -1$. Thus the training data are separated by a

function f(**x**) = $\boldsymbol{w}^{\mathrm{T}}\mathbf{x} - \gamma = 0$, the equation representing the hyperplane in the **x** space. While there may be many such hyperplanes existing that can achieve such separation of **x**, SVM aims at locating the hyperplane that maximizes the separation between the two classes of **x** it creates. Mathematically, this is achieved by finding unit vector **w** that minimizes a *cost function* $\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} = \frac{1}{2}\| \boldsymbol{w} \|^2$ subject to the separability constraints

$y_i(\boldsymbol{w}^T\boldsymbol{x}_i - \gamma) \geq 1; i = 1, 2, 3, \dots, l.$ ----------(2)

Sometimes the training data is not completely separable by a hyperplane. In such situations a slack variable $\xi_i$ is added to relax the strict separability constraints in (2) as follows:

$y_i(\boldsymbol{w}^T\boldsymbol{x}_i - \gamma) \geq 1 - \xi_i;\ \xi_i \geq 0;\ i = 1, 2, 3, \dots, l.$ ---------- (3)

The new cost function that now must be minimized becomes

$J(\boldsymbol{w}, \boldsymbol{\xi}) = \frac{1}{2}\| \boldsymbol{w} \|^2 + C\sum_{i=1}^{l}\xi_i$ ------- (4)

Vapnik called $C$ a user-specified, positive "regularization" parameter. In the general sense, not all situations comprising training examples {($\mathbf{x}_i$, $y_i$), i = 1, 2, … $l$} can be effectively modeled by the linear relationship (1), for the relationship may be nonlinear. To handle these SVM utilizes *kernels*—functions that can easily compute dot products of two vectors, a key requirement to achieve computational efficiency [14].

In (1) $\boldsymbol{w}$ is a weight vector and b is the bias. The hyperplane {**x**: f(**x**) = $\boldsymbol{w}^{\mathrm{T}}\mathbf{x} - \gamma$} devides the input space of **x** into two and the sign of f(**x**), the discriminant function of the classifier, denotes the side of the hyperplane a point **x** is on. The decision boundary is the demarcation between the two regions classified as positive and negative. When the decision boundary is a linear function of the input examples, it is called a liner classifier. In general, this boundary can be nonlinear. If we assume that the input data space spanned by **x** *is* linearly separable, a linear decision boundary (a hyperplane) exists in it. Indeed, many such hyperplanes may exist. The goal of SVM learning is to use the input data to design an optimum hyperplane (f(**x**)) that will maximize the geometric distance (the "margin") between the examples in the two classes. This is achieved as stated earlier by finding unit vector **w** that minimizes the *cost function* $\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} = \frac{1}{2}\| \boldsymbol{w} \|^2$ subject to the separability constraints

$y_i(\boldsymbol{w}^T\boldsymbol{x}_i - \gamma) \geq 1; i = 1, 2, 3, \dots, l.$

These constraints here ensure that the classifier f(**x**) classifies each example $\mathbf{x}_i$ correctly. Under the just stated assumption of linear separability being possible, the ***hard margin SVM*** (Figure 1, source Stackoverflow.com 2013) can be constructed to help classify unseen examples. Note that γ is computed once (4) has been minimized [13]. Mathematically this problem is one of optimization:

$minimize\ \frac{1}{2}\| \boldsymbol{w} \|^2\ using\ \boldsymbol{w}, \gamma\ subject\ to\ y_i(\boldsymbol{w}^T x_i - \gamma) \geq 1\ i = 1, 2, 3, \dots, n$ ------ (5)

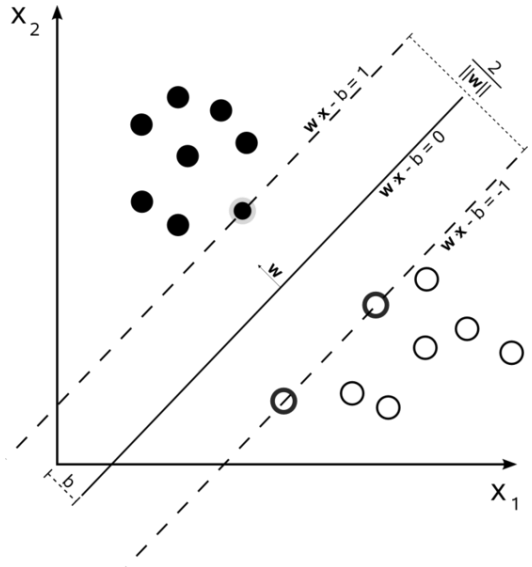


**Figure 1. The Hard Margin SVM in the ($X_1$, $X_2$) feature space**

The different kernel functions are listed below [16].

1] *Polynomial:* A polynomial mapping is a popular method for non-linear modeling. Intuitively, the polynomial kernel considers given features and combination of these features to determine their similarity. The second kernel is usually preferable as it avoids problems with the hessian becoming Zero.

$K(x, x') = (s\langle x, x' \rangle + c)^d$ ------- (6)

$K(x, x') = (\langle x, x' \rangle + 1)^d$ ------ (7)

2] *Gaussian Radial Basis Function*: Radial basis functions most commonly with a Gaussian form

$K(\mathrm{x}, \mathrm{x}') = \exp(-\frac{\|x - x'\|^2}{2\sigma^2})$ -------- (8)

3] Quadratic : This kernel function is used with non-linearly separable data.

$K(x, x') = (s\langle x, x' \rangle + c)^2$ --------- (9)

4] *Multi-Layer Perceptron*: The long established MLP, with a single hidden layer, also has a valid kernel representation.

$K(x, x') = \tanh(\langle x, x' \rangle + C)$ --------- (10)

5] Linear kernel : This kernel function is used to classify linearly separable data.

$K(x, x') = \langle x, x' \rangle$ ----------(11)

where s, c and $\sigma$ are kernel-specific parameters.

## 4. Experiments

The database used for analysis in this study has been compiled as a part of an earlier study entitled Early Detection Project (EDP) conducted at the Hemorheology Laboratory of the erstwhile Inter-Disciplinary Programme in Biomedical Engineering at the School (now Department) of Biosciences and Bioengineering, Indian Institute of Technology Bombay (IITB), Mumbai, India. Spanning over a period from January 1995 to April 2005, it compiled 981 records, each with 30 parameters, which encapsulated the biochemical, hemorheological and clinical status of the individuals. We note that the Hemorheology Laboratory has pioneered the research in the field of Clinical Hemorheology by conducting the baseline hemorheological studies in the Indian population and correlating various hemorheological parameters with several disease conditions.

In all, 13 parameters were noted for each respondent after using PCA (principal component analysis) technique for data dimension reduction [19]. Table 1 describes the symptom (input) variables used for the present study. They include age, health indicators (e.g. systolic blood pressure (BP1), diastolic blood pressure (BP2)) and biochemical parameter like Serum Proteins (SP), Serum Albumin (SALB), Hematocrit (HCT), Serum Cholesterol (SC), Serum Triglycerides (STG), along with various hemorheological (HR) parameters (e.g.; Whole Blood Viscosity(CBV), Plasma Viscosity(CPV), using a Contraves 30 viscometer, and Red Cell Aggregation (RCA).We used this database to develop and validate SVM models for four classification schemes: Classification Scheme I ( healthy vs. diabetic) , Classification Scheme II (diabetic vs. hypertensive), classification scheme III ( diabetic vs. diabetic and hypertensive) and classification scheme IV ( healthy vs. diabetic) with KNN-imputed data for missing values . The SVM models were used to select thirteen input variables that would yield the best classification of individuals into these diabetes categories.

For inputs to the SVM model, the first 13 columns of data represent the patient's health parameters. The 14$^{th}$ column represented the diagnosis made by the doctor for the patient. Dataset DS1 is a mixed data set, having samples of diabetic and healthy patients. Dataset DS2 is a dataset which stores data about hypertensive and diabetic patients. DS3 is a dataset, having diagnosis information about patients who are diabetic and hypertensive as well as diabetic. Dataset DS4 is KNN imputed dataset for missing values having information about patients who are diabetic and healthy. [18]

**Table 1. Symptom variables of datasets used in the study**

| #. | Symptom variable name | Data Type |
|---|---|---|
| 1. | AGE | Numeric, Range(19-73) |
| 2. | BSF | Numeric, Range(48,311) |
| 3. | BSP | Numeric, Range(61,383) |
| 4. | SC | Numeric, Range(90,389) |
| 5. | STG | Numeric, Range(41,456) |
| 6. | SALB | Numeric, Range(3.1,6.45) |
| 7. | SP | Numeric, Range(0.83,10.68) |
| 8. | CPV | Numeric, Range(1.069,1.785) |
| 9. | CBV | Numeric, Range(2.448,8.695) |
| 10. | HCT | Numeric, Range(22,60) |
| 11. | RG | Numeric, Range(1.374,6.174) |
| 12. | BP1 | Numeric, Range(98,240) |
| 13. | BP2 | Numeric, Range(60,116) |

## 5. Results and Discussion

Kernel functions such as *linear, quadratic, Polyorder, MLP and RBF* were evaluated in terms of their discriminative classification accuracy. The liner kernel function performed best in Classification Scheme -- I, III and IV, and the quadratic linear kernel function performed best in Classification Scheme II. Performance parameters such as the accuracy, sensitivity and specificity were presented in Table. The SVM was train using four datasets. For the first dataset, DS1—SVM with linear kernel recorded the best classification accuracy of 85.2% with sensitivity 87.3 % and specificity 83.2%. The best classification accuracy of 84.8% for DS2 was shown by quadratic kernel function model. The sensitivity and specificity were 79.3 % and 82.1% respectively. For datasets, DS3 and DS4 all the kernel functions displays satisfactory level of accuracy, linear kernel function was a better choice due to slightly better accuracy level. The classification accuracy for first three datasets was higher than that of fourth dataset due to the fact that the fourth dataset was KNN-imputed for missing values and first three datasets were cleaned datasets.

**TABLE 2. Experimental results of SVM classifier accuracy (sensitivity, specificity).**

| # | Dataset | SVM classification accuracy with kernel functions | | | | |
|---|---|---|---|---|---|---|
| | | Linear | Quadratic | Poly (3) | RBF | MLP |
| 1 | DS 1 | 85.2 (87.3, 83.2) | 82.7 (80.8, 82.6 | 76.4 (80.1, 83.7) | 75.1 (84.3, 85.6) | 67.3 (71.7, 62.6) |
| 2 | DS 2 | 83.8 (81.1, 85.1) | 84.8 (79.3, 82.1) | 74.5 (81.3, 88.6) | 72.2 (80.4, 88.6) | 68.5 (72.3, 68.1) |
| 3 | DS 3 | 88.4 (88.5, 84.4) | 86.2 (84.2, 80.4) | 75.1 (79.3, 88.5) | 63.4 (71.3, 78.6) | 64.1 (63.3, 68.6) |
| 4 | DS 4 | 81.2 (75.6, 78.2) | 80.5 (71.1, 68.9) | 68.37 (71.5, 72.8) | 66.13 (61.3, 71.6) | 63.5 (65.3, 68.6) |

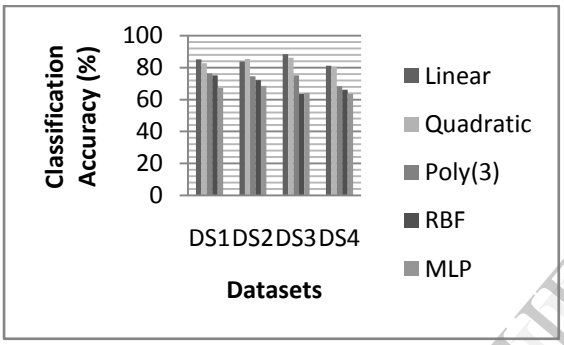


**Figure 2. Classification accuracy for different SVM kernel function models**

## 6. Conclusion

Support vector machine modeling is a promising classification approach for detecting persons with common diseases such as diabetes and pre-diabetes in the population. In this study we implemented SVM with five different kernel functions and investigated the appropriate choice of kernel function for the prediction of diabetes.

SVM is a model-free method that provides efficient solutions to classification problems without any assumption regarding the distribution and interdependency of the data. In epidemiologic studies and population health surveys, the SVM technique has the potential to perform better than traditional statistical methods like logistic regression, especially in situations that include multivariate risk factors with small effects (e.g., genome-wide association data and gene expression profiles), limited sample size, and a limited knowledge of underlying biological relationships among risk factors. This is particularly true in the case of common complex diseases where many risk factors, including gene-gene interactions and gene-environment interactions, have to be considered to reach sufficient discriminative power in prediction models. Our work provides a promising proof of principle by demonstrating the predictive power of the SVM with just a small set of variables. This approach can be extended to include large data sets, including many other variables, such as genetic biomarkers, as data from different domains become available.

## 7. References

[1] WHO/IDF 2006 (2007, Jan.). *Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia*, World Health Organization [Online]. Available: http://www.who.int/diabetes/publications/Definition%20and%20diagnosis%20of%20diabetes_new.pdf

[2] Ban Hyo-Jeong, Jee Yeon Heo, Kyung-Soo Oh and Keun-Joon Park (2010). Identification of Type 2 Diabetes-associated combination of SNPs using Support Vector Machine, *BMC Genetics*, 11:26 http://www.biomedcentral.com/1471-2156/11/26

[3] M. Uusitupa, "Lifestyle matter in prevention of type 2 diabetes," *Diabetes Care*, vol. 25, no. 9, pp. 1650–1651, 2002.

[4] Rudiger W. Brause,(2000) Medical Analysis and Diagnosis by Neural Networks , URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1.2901&rep=rep1&type=pdf.

[5] Boser, BE. I M Guyon and V N Vapnik (1992). *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, ed. D Haussler, ACM Press.

[6] Lijun Cheng, Yongsheng Ding , SVM and statistical technique method applying in Primary Open Angle Glaucoma diagnosis, Intelligent Control and Automation (WCICA), 2010 8th World Congress, pp- 2973 - 2978

[7] Shashikant Ghumbre, Chetan Patil, and Ashok Ghatol,(2011) , Heart Disease Diagnosis using Support Vector

Machine , International Conference on Computer Science and Information Technology (ICCSIT'2011) Pattaya Dec. 2011 ,pp 84-88.,

[8] Xiao-Peng Zhang*, Zhi-Long Wang, Lei Tang, Ying-Shi Sun, Kun Cao and Yun Gao, (2011) Support vector machine model for diagnosis of lymph node metastasis in gastric cancer with multidetector computed tomography: a preliminary study URL: http://www.biomedcentral.com/1471-2407/11/10

[9] A. Kampourakia, D. Vassisa, P. Belsisb, C. Skourlasa, (2013), e-Doctor: A Web based Support Vector Machine for Automatic Medical Diagnosis, *Procedia - Social and Behavioral Sciences* Volume 73, 27 February 2013, Pages 467–474

[10] Yu-Len Huang , Kao-Lun Wang , Dar-Ren Chen (2005) , Diagnosis of breast tumors with ultrasonic

texture analysis using support vector machines URL: http://web.thu.edu.tw/ylhuang/www/publish/NCA200604.pdf

[11] R.priya and P Aruna. Article: SVM and Neural Network based Diagnosis of Diabetic Retinopathy. *International Journal of Computer Applications* 41(1):6-12, March 2012

[12] Chunquan Huang The Research on Evaluation of Diabetes Metabolic Function Based on Support Vector Machine 2010 3rd International Conference on Biomedical Engineering and Informatics (BMEI 2010)

[13] Vapnik, VN (1998). *Statistical Learning Theory*, John Wiley.

[14] El-naqa, Isham (2012). Machine learning methods for predicting tumor response in lung cancer, Wiley Interdisciplinary Reviews: *Data Mining and Knowledge Discovery*, Vo. 2 (2), 173-181.

[15] Han, J and M Kamber (2006). *Data Mining Concepts and Techniques*, 2nd ed., Morgan Kaufman.

[16] Nello Cristianini and John Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods", Cambridge University Press, 2000.

[17] Tapan Bagchi, Rahul Samant, Milan Joshi (2013), "SVM Classifiers Built Using Imperfect Training Data", International Conference on Mathematical Techniques In Engineering Applications, ICMTEA 2013-BM-003

[18] Rahul Samant, Srikantha Rao . " *Effects of Missing Data Imputation on Classifier Accuracy* ", Vol.2 - Issue 11 (November - 2013), International Journal of Engineering Research & Technology (IJERT) , ISSN: 2278-0181 , pp 264-266, URL: www.ijert.org

[19] Rahul Samant, Srikantha Rao . " *A study on Feature Selection Methods in Medical Decision Support Systems* ", Vol.2 - Issue 11 (November - 2013), International Journal of Engineering Research & Technology (IJERT) , ISSN: 2278-0181 , pp 615-619, URL: www.ijert.org