# Diagnosing Vulnerability of Diabetic Patients to Heart Diseases using Support Vector Machines

G. Parthiban
Research Scholar,
Dr. MGR Educational Research and Institute,Maduravoyal,
Chennai, India.

A. Rajesh
Professor, Dept of CSE
C.Abdul Hakkeem College of Engineering and Technology,
Melvishram, Vellore, India.

S. K. Srivatsa, PhD.
Sr. Professor, Dept of E & I,
St.Joseph's College of Engineering,
Chennai, India.

## ABSTRACT
Data mining is the analysis step of the Knowledge Discovery in Databases process (KDD). While data mining and knowledge discovery in databases are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. Diabetes is a chronic disease that occurs when the pancreas does not produce enough insulin, or when the body cannot effectively use the insulin it produces. Most of these systems have successfully employed Support Vector Machines for the classification purpose. On the evidence of this we too have used SVM classifier using radial basis function kernel for our experimentation. The results of our proposed system were quite good. The system exhibited good accuracy in predicting the vulnerability of diabetic patients to heart diseases.

## Keywords
Data Mining, Diabetes, Heart Diseases, Knowledge Discovery, Support Vector Machines.

## 1. INTRODUCTION
Knowledge discovery in databases (KDD) also termed as Data mining aims to find useful information from large collection of data. This process consists of iterative sequence of data cleaning, data selection, data mining pattern recognition and knowledge presentation. Data mining technology is useful for extracting non trivial information from medical databases. [1], [2] It is a interdisciplinary field closely connected to data warehousing, statistics, machine learning, and neural networks.

Data mining is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses [3]. Data mining tools predict future trends and behaviours, help organizations to make proactive knowledge-driven decisions [4]. There are various data mining techniques available with their suitability dependent on the domain application. Data mining application in health can have tremendous potential and usefulness. It automates the process of finding predictive information in large databases. The classification model used training data set to build classification prediction model and testing data used for testing the classification efficiency.

The term "diabetes mellitus" describes a metabolic disorder of multiple aetiology characterized by chronic hyperglycaemia with disturbances of carbohydrate, fat and protein metabolism resulting from defects in insulin secretion, insulin action, or both. Diabetes is classified into two main types. Type 1 (T1B) diabetes is due to deficient insulin production. It develops during childhood and adolescence. In this case, patients require lifelong insulin injection for survival. Type 2 (T2B) diabetes is due to body's ineffective use of insulin. [5] Diabetes is a chronic disease which causes serious health complications including heart disease, kidney failure and blindness. [5], [6]

Heart disease is a term for variety of disease that affecting the heart such as chest pain, shortness of breath, heart attack and other symptoms. It encompasses the diverse diseases that affect the heart. [29] Chest pains arise when the blood received by the heart muscles is inadequate.

It is the more common type – accounts for 90% of diabetic cases worldwide. It develops during adulthood. It is related to obesity, lack of physical activity and unhealthy diets. Treatment involves lifestyle changes, weight loss, or oral medications or even insulin injection in some cases.

Hyperglycemia in the long term may cause damage to eyes (leading to blindness), damage to kidneys (leading to impotence and foot disorders/ amputation), increases the risk of heart disease (stroke) and insufficiency in blood flow to legs [9]. Around 366 million people have diabetes world wide according to statistics taken in the year 2011. Also it has been projected that the people with diabetes will increase to around 552 million by the year 2030. The number of people with type 2 diabetes is increasing in every country. [7], [8]

Diabetes is a major risk factor for cardiovascular disease (disease of the heart and circulatory system). It is the main cause of death in people with diabetes (around 50%). People with type 2 diabetes are likely to die 5 to 10 years earlier than people without diabetes. Most of these deaths is due to cardiovascular disease [10]. People with type 2 diabetes are more prone to have a heart attack or stroke – twice as likely as those without diabetes [11]. It has been found that a large part of the costs attributable to type 2 diabetes is due to the treatment of cardiovascular diseases [12]. Changes in lifestyle, weight loss, dietary changes and increased physical activity can greatly reduce the risks due to cardiovascular diseases [12]. Timely detection of these people will result in reduced mortality of diabetics as well as eliminating the cost due to the treatment of cardiovascular diseases. Automatic intelligent diagnosis systems can help greatly in identifying vulnerable sections of the diabetic patients. There are several systems for diagnosis and management of diabetes [13] – [17]. However these systems are designed to predict the chances of a person getting diabetes not the vulnerability of diabetic patients to heart disease.

Likewise, there are systems to predict the chances of a person getting cardiovascular disease [20], [21]. The utility of such systems in health care has been found to be quite high [18].It has been found that Support Vector Machines (SVM's) have been quite successfully employed in such systems [22]. Hence we have used an SVM classifier for our experimentation.

This research paper is the extension of our previous work, [30] diagnosis of heart disease for diabetic patients using Naïve bayes method. Here we are using Support vector machine and it is organized as follows in the subsequent sections – section 2 gives a brief background of support vector machines, section 3 gives our experimentation methodology, section 4 gives the results of our experiments and section 5 concludes this paper.

## 2. SUPPORT VECTOR MACHINES (SVM) BACKGROUND

A Support Vector Machine (SVM) is a concept in statistics and computer science for a set of related supervised learning methods that analyze data and recognize pattern introduced by Corinna Cortes and Vladimir Vapnik used for classification and regression analysis. SVM have shown good performance in a number of application areas. It constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. [23] SVM's are very much useful in data classification. SVM's classify data by finding an optimal hyper plane separating the d – dimensional data into its two classes with a maximum interclass margin. SVM's use so called kernel functions to cast data into a higher dimensional space where the data is separable. [24], [25]

SVM is a learning machine that plots the training vectors in high dimensional space and labels each vector by its class. [28] SVM based on the principle of risk minimization which aims to, minimize the error rate. [26], [27] SVM uses a supervised learning approach for classifying data. That is, SVM produces a model based on a given training data which is then used for predicting the target values of the test data. Given a labelled training set $(x_i,y_i)$, SVM require the solution of the following optimization problem to perform classification [17].

$$\min_{w,b,\varepsilon} {}^{1}/_{2}\, W^T W + C \sum_{i=1}^{l} \varepsilon_i \qquad 1$$

Subject to,

$$y_i(W^T \phi(x_i) + b) \geq 1 - \varepsilon_i \qquad 2$$

where,

$\varepsilon_i \geq 0$, a slack variable to allow for errors in the classification

$x_i$ – training vectors, $x_i \in R^n$

$\phi$ - function mapping $x_i$ into a higher dimension space,

C – penalty parameter of the error term (usually C>0),

$y_i$ – Class label, $y_i \in \{1,-1\}^l$

## 3. EXPERIMENTATION METHODOLOGY

The methodology described in this paper is diagnosing vulnerability of diabetic patients to heart diseases and we had collected 500 records of diabetic patients to perform the experimentation. The attributes making up each record is shown in Table 1.

**Table 1. Attributes used for the diagnosis**

| Attribute Role | Attribute Name | Attribute Type | Description |
|---|---|---|---|
| Regular | Sex | binominal | Sex of the patient. Takes the following values: Male, Female |
| Regular | Age | integer | Age of the patient |
| Regular | Fam/Heri | polynomial | Indicates whether the patient's parents were affected by diabetes. Takes the following values: Father, Mother, Both |
| Regular | Weight | numeric | Weight of the patient |
| Regular | BP | polynomial | Blood Pressure of the patient |
| Regular | Fasting | integer | Fasting Blood Sugar |
| Regular | PP | integer | Post Prondial Blood Glucose |
| Regular | A1C | numeric | Glycosylated Hemoglobin Test |
| Regular | LDL | integer | Low Density Lipoprotein |
| Regular | VLDL | integer | Very Low Density Lipoprotein |
| Label | Vulnerability | nominal | Indicates the vulnerability of the patients to heart disease. Takes the following values: High, Low |

Out of the 500 records, 142 records were pertaining to patients highly vulnerable to heart diseases. The remaining 358 records were pertaining to patients less vulnerable to heart disease. Since SVM processes only numeric attributes, the nominal were converted to numeric attributes by replacing each value by a unique integer. For example, the attribute Sex values are converted as follows: Male – 1 and Female – 0.

The values of the attributes were then normalized to the range 0 to 1. These records were then given as input to the SVM classifier.

SVM uses kernel functions to map the data set to a high dimensional data space for performing classification. The different types of kernel functions are as follows [17]:

Linear: $K(x_i,x_j) = x_i^T x_j$ 3

Polynomial:

$$K(x_i,x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0 \qquad 4$$

Radial Basis Function:

$$K(x_i, x_j) = \exp\left(\gamma \| x_i - x_j \|^2\right), \gamma > 0 \quad (5)$$

Sigmoid:

$$K(x_i, x_j) = \tanh\left(\gamma x_i^T x_j + r\right) \quad (6)$$

where γ, r, d are kernel parameters. The choice of the kernel depends on whether the relationship between the class labels and attributes are linear or nonlinear. For nonlinear relationships, the radial basis function (RBF) kernel has been found to be a good choice as it has lesser number of hyper parameters than other nonlinear kernels. Also RBF kernel has fewer numerical difficulties [19]. Hence we have used RBF kernel in our SVM classifier.

## 4. RESULT ANALYSIS

The data set used for training the classifier comprises of 500 diabetic patient records out of which 142 records are of those having heart disease (positive cases) and the remaining 358 records are of those not having heart disease (negative cases). These records after sufficient pre-processing was given as input to train the SVM classifier.

The SVM classifier was trained for different values of the RBF kernel parameters, C and γ. The models thus obtained for each of the values of C and γ where then tested for accuracy. A good classifier should be able to exhibit high accuracy for datasets unseen rather than the training data. Hence we have used 10 fold cross validation for testing the accuracy of the classifier.

In 10-fold cross-validation, we first divide the training set into 10 subsets of equal size. Sequentially one subset is tested using the classifier trained on the remaining 9 subsets. Thus, each instance of the whole training set is predicted once so the cross-validation accuracy is the percentage of data which are correctly classified. The cross validation tests prevents overfitting problem. Based on the exhaustive trials conducted, we found that for C = 5.0 and γ = 1.0 the classifier exhibited the best accuracy of 94.60%. The accuracy obtained for a few values of C and γ in our trials is shown in the Table 2.

**Table 2. Partial results of the trials conducted**

| C Value | γ Value | Accuracy of the |
|---|---|---|
| . | . | . |
| 2 | 0.125 | 89.60% |
| 2 | 0.75 | 92.40% |
| 4 | 2.5 | 93.20% |
| 4 | 2 | 93.60% |
| 4 | 1.5 | 93.80% |
| 4 | 1 | 94.20% |
| **5** | **1** | **94.60%** |
| 6 | 1.25 | 94% |
| . | . | . |

The ROC curve for the classifier characteristics is shown in Fig. 1

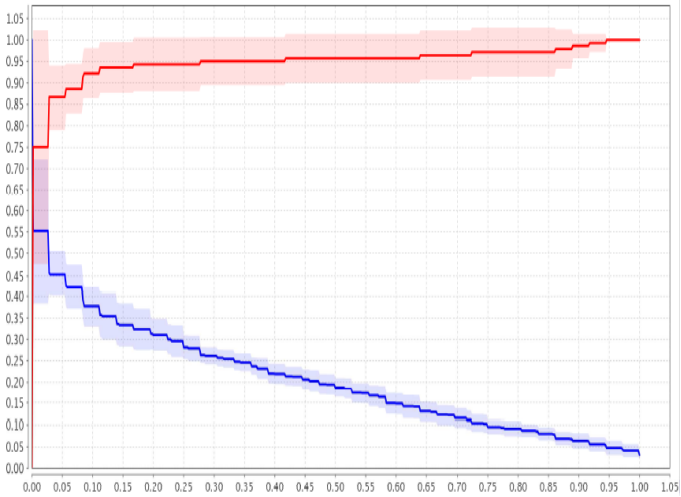

**Fig 1: ROC curve for the classifier characteristics**

The confusion matrix indicating the accuracy of the SVM classifier for the given data set is shown in Table 3.

**Table 3. The confusion matrix of the classifier**

| | True low | True high | Class |
|---|---|---|---|
| pred. low | 355 | 24 | 93.67% |
| pred. high | 3 | 118 | 97.52% |
| class recall | 99.16% | 83.10% | |
| Overall accuracy: 94.60% +/- 2.01% (mikro: 94.60%) | | | |

From the results obtained, it can be seen that the classifier exhibits a very high classification accuracy i.e 94.60% overall. It also shows a very high precision for the positive class (97.52%) and also the recall of the positive class is quite good (83.10%). In the case of negative classes, the classifier exhibits high precision (93.67%) as well as high recall (99.10%).

## 5. CONCLUSIONS

In this paper, we have shown that it is possible to diagnose heart disease vulnerability in diabetic patients with reasonable accuracy. Classifiers of this kind can help in early detection of the vulnerability of a diabetic patient to heart disease. There by the patients can be forewarned to change their lifestyle. This will result in preventing diabetic patients from being affected by heart disease, there by resulting in low mortality rates as well as reduced cost on health for the state. SVM's have proven to be a classification technique with excellent predictive performance and also been investigated with the help of ROC curve for both training and testing data. Hence this SVM model can be recommended for the classification of the diabetic dataset.

## 6. ACKNOWLEDMENTS

We are grateful to Dr.V.Shesiah, Chairman and Managing director of Dr.V.Shesiah Diabetic Research Institute, Chennai for providing an access to medical diabetic data and for his involvement in this domain.

## 7. REFERENCES

[1] J. Han Kamber, M. 2006. Data Mining: Concepts and Techniques, 2nd ed. San Francisco: Morgan Kaufman.

[2] U. Fayyad, G.Piatetsky-Shapiro, and P.Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine, Vol.17, pp.37-54, 1996.

[3] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation 10500 Falls Road, Potomac, MD 20854 (U.S.A.),1999.

[4] L.A.Rose, D.T., "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN O-471-66657-2, ohn Wiley & Sons, Inc, 2005.

[5] World Health Organization. Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia: Available: http://www.who. int/ diabetes/en

[6] World Health Organization. Available: http: // www.who.int/topics/ diabetes mellitus/en/

[7] International Diabetes Federation (IDF), Available: http://www.idf .org/ about-diabetes

[8] American Diabetes Association Available: http: // www.diabetes.org/

[9] NJ Morrish, Wang , Stevens , Fuller JH, Keen H. "Mortality and causes of death in the WHO Multinational Study of Vascular Disease in Diabetes". Diabetologia 2001; 44 Suppl 2:s14 – s21 Available: http:// citeulike.org /user/mpgrayer/article/3496572

[10] R.Emslie – AM Smith Gardner ID, Morris AD. "Vascular Complications of Diabetes". British Medical Journal 2000; 320:1062 – 6 Available:http://bmj.com/BMJ 2000; 320 doi: 10.1136/bmj.320.7241.1062

[11] S.M. Haffner Lehto S, Ronnemaa T, Pyoyala K, Laakso M. "Mortality from Coronary heart disease in subjects with Type 2 Diabetes and Nondiabetic Subjects with and without prior Myocardial Infarction". The New England Journal of Medicine 1998; 339:229 – 34 Available: http:// biomedcentral.com/BMC Health Services Research 2011, 11:180 doi: 10.1186/1472-6963-11-180

[12] Diabetes with cardiovascular disease- Available: www. idf.org/fact-sheets/diabetes-cvd

[13] Y. Huang, P. McCullagh, N. Black, and R. Harper, "Feature selection and classification model construction on type 2 diabetic patient's data," in Lecture Notes in Artificial intelligence, vol. 3275, ICDM 2004, P. Perner, Ed. Berlin: Springer-Verlag, 2004, pp. 153-162.

[14] R. Bellazzi, C. Larizza, P. Magni, S. Montani, and M. Stefanelli, "Intelligent analysis of clinical time series: an application in the diabetes mellitus domain," Artificial Intelligence in Medicine, vol. 20 pp. 37-57, 2000.Available: http://dx.doi.org/10.1016/S0933-3657(00)00052-X

[15] R. Bellazzi, "Telemedicine and Diabetes Management: Current Challenges and Future Research Directions," Journal of Diabetes Science and Technology, vol. 2, no. 1, pp. 98-104, 2008.

[16] R. Goel, A. Misra, D. Kondal, R. M. Pandey, N. K. Vikram, J. S. Wasir, V. Dhingra, and K. Luthra, "Identification of insulin resistance in Asian Indian adolescents:classification and regression tree (CART) and logisticregression based classification rules," Clinical Endocrinology, vol. 70 pp. 717-724, 2009.

[17] K. E. Heikes, B. Arondekar, D. M. Eddy, and L. Schlessinger, "Diabetes Risk Calculator,A simple tool for detecting undiagnosed diabetes and pre-diabetes," Diabetes Care, vol. 31, no. 5, pp. 1040-1045, 2008

[18] N. Lavrac, E. Keravnou, and B. Zupan, "Intelligent Data Analysis in Medicine," in Encyclopedia of Computer Science and Technology, vol. 42, K. e. al., Ed. New York: Dekker, 2000, pp. 113-157.

[19] W. Kong, L. Tham, K. Y. Wong, and P.Tan, "Support vector machine approach for cancer detection using amplified fragment length polymorphism (AFLP) method," Proc. the 2nd Asia-Pacific Bioinformatics Conference (APBC2004), Dunedin, New Zealand, 2004.

[20] A. Karaolis, Joseph A. Moutiris, Demetra Hadjipanayi, and Constantinos S. Pattichis. "Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees". IEEE Transaction on Information Technology in Biomedicine, Vol. 14, No. 3, May2010.Available: http://ieeexplore.ieee.org//xpls/abs_all.jsp?arnumber=5378501DigitalObject Identifier: 10.1109/TITB.2009.2038906

[21] K. Srinivas, Dr. G. Raghavendra Rao, and Dr. A. Govardhan. "Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques". The 5th International Conference on Computer Science & Education Hefei, China. August 24–27, 2010.

[22] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. "A Practical Guide to Support Vector Classification". Available: http:// www.csie .ntu.edu.tw/~cjlin.

[23] Introduction to Support Vector Machine Available: http://en. wikipedia. org/wiki/Support_vector_machine

[24] Colin Campbell and Yiming Ying, Learning with Support Vector Machines, 2011, Morgan and Claypool. Available: http://www. morganclaypool.com/doi/abs/10.2200/S00324ED1V01Y201102AIM010?journalCode=aim

[25] H.Barakat, Andrew P.Bradley and Mohammed Nabil H.Barakat (2009) "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus", IEEE Transactions on Information Technology in Bio Medicine, Volume 14, Issue 4, pp 1-7, 2009. Available: http://ieeexplore. ieee.org /xpls/ abs_ all.jsp?arnumber=5378519 Digital Object Identifier: 10.1109 /TITB. 2009.2039485

[26] N. Barakat and A.P.Bradley, "Rule Extraction from Support Vector Machines: A Sequential Covering Approach " IEEE Transactions on Knowledge and Data Engineering, Volume 19,no.6,pp 729-741, 2007. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=04161896 Digital Object Identifier no. 10.1109/TKDE.2007.1023.

[27] S.Balakrishnan, R.Narayanaswamy, N.Savarimuthu, R.Samikannu "SVM Ranking with Backward Search for Feature Selection in Type II Diabetes Databases" 2008

IEEE International Conference on Systems, Man and Cybernetics. Available: http://ieeexplore. ieee.org /xpls /abs_all.jsp?arnumber=4811692 Digital Object Identifier: 10.1109 /IC SMC.2008.4811692

[28] G.Suganya, D.Dhivya "Extracting Diagnostic rules from SVM" , Journal of Computer Applications (JCA), 2011.

[29] K.Srinivas et al. / "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", International Jounal on Computer Science and Engineering (IJCSE) Vol..02, No.02, 2010, 250-255. Available:http://www.enggjournals.com/ijcse/doc /IJCSE10-02-02-25.pdf

[30] G.Parthiban, A.Rajesh, S.K.Srivatsa, "Diagnosis of Heart Disease for Diabetic Patients using Naïve Bayes Method", International Journal of Computer Applications (IJCA) Volume 24-No.3, June 2011, 0975-8887.Available: http://www.ijcaonline.org/archives/volume24/number3/2933-3887 doi 10.5120/2933-3887