Decision Support System for Medical Diagnosis Using Data Mining

D.Senthil Kumar[1], G.Sathyadevi[2] and S.Sivanesh[3]

**[1] Department of Computer Science and Engineering, Anna University of Technology,
Tiruchirappalli, Tamil Nadu, India**

**[2,3] Department of Computer Science and Engineering, Anna University of Technology,
Tiruchirappalli, Tamil Nadu, India**

Abstract

The healthcare industry collects a huge amount of data which is not properly mined and not put to the optimum use. Discovery of these hidden patterns and relationships often goes unexploited. Our research focuses on this aspect of Medical diagnosis by learning pattern through the collected data of diabetes, hepatitis and heart diseases and to develop intelligent medical decision support systems to help the physicians. In this paper, we propose the use of decision trees C4.5 algorithm, ID3 algorithm and CART algorithm to classify these diseases and compare the effectiveness, correction rate among them**.**

Keywords: *Active learning, decision support system, data mining, medical engineering, ID3 algorithm, CART algorithm, C4.5 algorithm.*

1. Introduction

The major challenge facing the healthcare industry is the provision for quality services at affordable costs. A quality service implies diagnosing patients correctly and treating them effectively. Poor clinical decisions can lead to disastrous results which is unacceptable. Even the most technologically advanced hospitals in India have no such software that predicts a disease through data mining techniques. There is a huge amount of untapped data that can be turned into useful information. Medical diagnosis is known to be subjective; it depends on the physician making the diagnosis. Secondly, and most importantly, the amount of data that should be analyzed to make a good prediction is usually huge and at times unmanageable. In this context, machine learning can be used to automatically infer diagnostic rules from descriptions of past, successfully treated patients, and help specialists make the diagnostic process more objective and more reliable.

The decision support systems that have been developed to assist physicians in the diagnostic process often are based on static data which may be out of date. A decision support system which can learn the relationships between patient history, diseases in the population, symptoms, pathology of a disease, family history and test results, would be useful to physicians and hospitals. The concept of Decision Support System (DSS) is very broad because of many diverse approaches and a wide range of domains in which decisions are made. DSS terminology refers to a class of computer-based information systems including knowledge based systems that support decision making activities. In general, it can say that a DSS is a computerized system for helping make decisions. A DSS application can be composed of the subsystems. However, the development of such system presents a daunting and yet to be explored task. Many factors have been attributed but inadequate information has been identified as a major challenge. To reduce the diagnosis time and improve the diagnosis accuracy, it has become more of a demanding issue to develop reliable and powerful medical decision support systems (MDSS) to support the yet and still increasingly complicated diagnosis decision process. The medical diagnosis by nature is a complex and fuzzy cognitive process, hence soft computing methods, such as decision tree classifiers have shown great potential to be applied in the development of MDSS of heart diseases and other diseases.

The aim is to identify the most important risk factors based on the classification rules to be extracted. This section explains how well data mining and decision support system are integrated and also describes the datasets undertaken for this work. In the next section relevant related works referred to the exploitation of classification technology in the medical field are surveyed. Section III outlines the results, explaining the decision tree algorithms devised for the purposes outlined above. Section IV illustrates conclusions.

Decision support systems are defined as interactive computer based systems intended to help decision makers utilize data and models in order to identify problems, solve problems and make decisions. They incorporate both data and models and they are designed to assist decision makers in semi-structured and unstructured decision making processes. They provide support for decision making, they do not replace it. The mission of decision support systems is to improve effectiveness, rather than the efficiency of decisions [19]. Chen argues that the use of data mining helps institutions make critical decisions faster and with a greater degree of confidence. He believes that the use of data mining lowers the uncertainty in decision process [20]. Lavrac and Bohanec claim that the integration of dm can lead to the improved performance of DSS and can enable the tackling of new types of problems that have not been addressed before. They also argue that the integration of data mining and decision support can significantly improve current approaches and create new approaches to problem solving, by enabling the fusion of knowledge from experts and Knowledge extracted from data [19].

2. Overview of related work

Up to now, several studies have been reported that have focused on medical diagnosis. These studies have applied different approaches to the given problem and achieved high classification accuracies, of 77% or higher, using the dataset taken from the UCI machine learning repository [1]. Here are some examples:

Robert Detrano's [6] experimental results showed correct classification accuracy of approximately 77% with a logistic-regression-derived discriminant function.

The John Gennari's [7] CLASSIT conceptual clustering system achieved 78.9% accuracy on the Cleveland database.

L. Ariel [8] used Fuzzy Support Vector Clustering to identify heart disease. This algorithm applied a kernel induced metric to assign each piece of data and experimental results were obtained using a well known benchmark of heart disease.

Ischemic -heart:-disease (IHD) -Support .Vector Machines serve as excellent classifiers and predictors and can do so with high accuracy. In this, tree based: classifier uses non-linear proximal support vector machines.(PSVM).

Polat and Gunes [18] designed an expert system to diagnose the diabetes disease based on principal component analysis. Polat *et al.* also developed a cascade learning system to diagnose the diabetes.

Campos-Delgado *et al.* developed a fuzzy-based controller that incorporates expert knowledge to regulate the blood glucose level.Magni and Bellazzi devised a stochastic model to extract variability from a self-monitoring blood sugar level time series [17].

Diaconis,P. & Efron,B. (1983) developed an expert system to classify hepatitis of a patient. They used Computer-Intensive Methods in Statistics.

Cestnik,G., Konenenko,I, & Bratko,I. designed a Knowledge-Elicitation Tool for Sophisticated Users in the diagnosis of hepatitis.

3. Analysis and results

3.1 About the Datasets

The Aim of the present study is the development and evaluation of a Clinical Decision Support System for the treatment of patients with Heart Disease, diabetes and hepatitis. According to one survey, heart disease is the leading cause of death in the world every year. Just in the United States, almost 930,000 people die and its cost is about 393.5 billion dollars. Heart disease, which is usually called coronary artery disease (CAD), is a broad term that can refer to any condition that affects the heart. Many CAD patients have symptoms such as chest pain (angina) and fatigue, which occur when the heart isn't receiving adequate oxygen. Nearly 50 percent of patients, however, have no symptoms until a heart attack occurs.

Diabetes mellitus is a chronic disease and a major public health challenge worldwide. According to the International Diabetes Federation, there are currently 246 million diabetic people worldwide, and this number is expected to rise to 380 million by 2025. Furthermore, 3.8 million deaths are attributable to diabetes complications each year. It has been shown that 80% of type 2 diabetes complications can be prevented or delayed by early identification of people at risk. The American Diabetes Association [2] categorizes diabetes into type-1 diabetes [17], which is normally diagnosed in children and young adults, and type-2 diabetes, i.e., the most common form of diabetes that originates from a progressive insulin secretory defect so that the body does not produce adequate insulin or the insulin does not affect the cells. Either the fasting plasma glucose (FPG) or the 75-g oral glucose tolerance test (OGTT [19]) is generally appropriate to screen diabetes or pre-diabetes.

Hepatitis, a liver disorder requires continuous medical care and patient self-management education to prevent acute complications and to decrease the risk of long-term complications. This is caused due to the condition of anorexia (loss of appetite) and increased level of alkaline phosphate. The disease can be classified in to Hepatitis a, b, etc,. All these datasets used in this study are taken from UCI KDD Archive [1].

3.2 Experimental Data

We have used three medical datasets namely, heart disease, diabetes and hepatitis datasets. All these datasets are obtained from UC-Irvine archive of machine learning datasets [1]. The aim is to classify the diseases and to compare the attribute selection measure algorithms such as ID3, C4.5 and CART. The heart disease dataset [1] of 473 patients is used in this experiment and has 76 attributes, 14 of which are linear valued and are relevant as shown in table 1. The hepatitis disease dataset [1] has 20 attributes, and there are 281 instances and 2 classes which are described in table 2. The diabetic dataset [1] of 768 patients with 9 attributes is as shown in table 3.

Table 1: Description of the features in the heart disease dataset

No	Name	Description
1	Age	age in years
2	Sex	1 = male ; 0 = female
3	Cp	chest pain type (1 = typical angina; 2 = atypical angina ; 3 = non-anginal pain; 4 = asymptomatic)
4	Trestbps	resting blood pressure(in mm Hg on admission to the hospital)
5	Chol	serum cholestoral in mg/dl
6	Fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7	Restecg	resting electrocardiographic results (0 = normal; 1 = having ST-T wave abnormality; 2 = showing probable or define left ventricular hypertrophy by Estes' criteria)
8	Thalach	maximum heart rate achieved
9	Exang	exercise induced angina (1 = yes; 0 = no)
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	the slope of the peak exercise ST segment (1 = upsloping; 2 = flat ; 3= downsloping)
12	Ca	number of major vessels (0-3) colored by flourosopy
13	Thal	(3 = normal; 6 = fixed defect; 7 = reversible defect)
14	Num	Diagnosis classes (0 = healthy; 1 = patient who is subject to possible heart disease)

Table 2: Description of the features in the hepatitis dataset

1	Class	DIE, LIVE
2	Age	10, 20, 30, 40, 50, 60, 70,80
3	Sex	male, female
4	Steroid	no, yes
5	Antivirals	no, yes
6	Fatigue	no, yes
7	Malaise	no, yes
8	Anorexia	no, yes
9	Liver Big	no, yes
10	Liver Firm	no, yes
11	Spleen Palpable	no, yes
12	Spiders	no, yes
13	Ascites	no, yes
14	Varices	no, yes
15	Bilirubin	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
16	Alk Phosphate	33, 80, 120, 160, 200, 250
17	SGOT	13, 100, 200, 300, 400, 500,
18	Albumin	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
19	Protime	10, 20, 30, 40, 50, 60, 70, 80, 90
20	Histology	no, yes

Table 3: description of the features in the diabetes dataset

No	Attribute Name	Description
1	Number of times pregnant	Numerical values
2	Plasma glucose concentration	glucose concentration in a 2 hours in an oral glucose tolerance test
3	Diastolic blood pressure	In mm Hg
4	Triceps skin fold thickness	Thickness of skin in mm
5	2-Hour serum insulin	Insulin (mu U/ml)
6	Body mass index	(weight in kg/(height in m)^2)
7	Diabetes pedigree function	A function – to analyse the presence of diabetes
8	Age	Age in years
9	Class	1 is interpreted as “tested positive for diabetes and 0 as negative

3.3 Attributes Selection Measures

Many different metrics are used in machine learning and data mining to build and evaluate models. We have implemented the ID3, C4.5 CART algorithm and tested them on our experimental datasets. The accuracy of these algorithms can be examined by confusion matrix produced by them. We employed four performance measures: precision, recall, F-measure and ROC space [5]. A distinguished confusion matrix (sometimes called contingency table) is obtained to calculate the four measures. Confusion matrix is a matrix representation of the classification results. It contains information about actual and predicted classifications done by a classification system. The cell which denotes the number of samples classifies as true while they were true (i.e., TP), and the cell that denotes the number of samples classified as false while they were actually false (i.e., TN). The other two cells denote the number of samples misclassified. Specifically, the cell denoting the number of samples classified as false while they actually were true (i.e., FN), and the cell denoting the number of samples classified as true while they actually were false (i.e., FP). Once the confusion matrixes were constructed, the precision, recall, F-measure are easily calculated as:

$$\text{Recall}= TP/ (TP+FN) \qquad (1)$$

$$\text{Precision} = TP/ (TP+FP) \qquad (2)$$

$$\text{F\_measure} = (2*TP)/ (2*TP+FP+FN) \qquad (3)$$

Less formally, precision measures the percentage of the actual patients (i.e. true positive) among the patients that got declared disease; recall measures the percentage of the actual patients that were discovered; F-measure balances between precision and recall. A ROC (receiver operating characteristic [5]) space is defined by false positive rate (FPR) and true positive rate (TPR) as x and y axes respectively, which depicts relative tradeoffs between true positive and false positive.

$$TPR= TP/ (TP+FN) \qquad (4)$$

$$FPR= FP/ (FP+TN) \qquad (5)$$

ID3 Algorithm

Itemized Dichotomozer 3 algorithm or better known as ID3 algorithm [13] was first introduced by J.R Quinlan in the late 1970's. It is a greedy algorithm that selects the next attributes based on the information gain associated with the attributes. The information gain is measured by entropy, ID3 algorithm [13] prefers that the generated tree is shorter and the attributes with lower entropies are put near the top of the tree. The three datasets are run against ID3 algorithm and the results generated by ID3 are as shown in tables 4, 5, 6 respectively.

Table 4: Confusion matrix of id3 algorithm- heart disease dataset

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.686	0.281	0.66	0.686	0.673	0.68	No
0.719	0.314	0.742	0.719	0.73	0.719	Yes

Table 5: Confusion matrix of id3 algorithm- hepatitis dataset

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.686	0.281	0.66	0.686	0.673	0.68	No
0.719	0.314	0.742	0.719	0.73	0.719	Yes

Table 6: confusion matrix of id3 algorithm- diabetes dataset

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.582	0.154	0.67	0.582	0.623	0.767	Yes
0.846	0.418	0.791	0.846	0.817	0.767	No

C4.5 Algorithm

At each node of the tree, C4.5 [15] chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. C4.5 [16] made a number of improvements to ID3. Some of these are:

a. Handling both continuous and discrete attributes –creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it.
b. Handling training data with missing attribute values
c. Handling attributes with differing costs.
d. Pruning trees after creation – C4.5 [16] goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with leaf nodes.

When the three medical datasets are run against the C4.5 algorithm and the results are indicated in the tables 7, 8, 9 respectively.

Table 7: confusion matrix of c4.5 algorithm- heart disease dataset

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.596	0.364	0.586	0.596	0.591	0.636	No
0.636	0.404	0.646	0.636	0.641	0.636	Yes

Table 8: Confusion matrix of c4.5 algorithm-hepatitis dataset

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.97	0.615	0.89	0.97	0.929	0.669	Live
0.385	0.03	0.714	0.385	0.5	0.669	Die

Table 9: Confusion matrix of c4.5 algorithm-diabetes dataset

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.597	0.186	0.632	0.597	0.614	0.751	Yes
0.814	0.403	0.79	0.814	0.802	0.751	No

CART Algorithm

Classification and regression trees (CART [14]) is a non-parametric technique that produces either classification or regression trees, depending on whether the dependent variable is categorical or numeric, respectively. Trees are formed by a collection of rules based on values of certain variables in the modelling data set. Rules are selected based on how well splits based on variables' values can differentiate observations based on the dependent variable Once a rule is selected and splits a node into two, the same logic is applied to each "child" node (i.e. it is a recursive procedure). Splitting stops when CART detects no further gain can be made, or some pre-set stopping rules are met. The basic idea of tree growing is to choose a split among all the possible splits at each node so that the resulting child nodes are the "purest". In this algorithm, only univariate splits are considered. That is, each split depends on the value of only one predictor variable. All possible splits consist of possible splits of each predictor. CART innovations include:

a. solving the "how big to grow the tree"- problem;
b. using strictly two-way (binary) splitting;
c. incorporating automatic testing and tree validation, and;
d. Providing a completely new method for handling missing values.

The result of CART algorithm for the medical datasets are described in the following tables 10, 11, 12 respectively

Table 10: Confusion matrix of CART algorithm-heart disease dataset

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.702	0.258	0.702	0.702	0.702	0.726	No
0.742	0.298	0.742	0.742	0.742	0.726	Yes

Table 11: Confusion matrix of CART algorithm- hepatitis dataset

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.91	0.769	0.859	0.91	0.884	0.541	Live
0.231	0.09	0.933	0.831	0.273	0.541	Die

Table 12: Confusion matrix of CART algorithm- diabetes dataset

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.534	0.132	0.884	0.934	0.6	0.727	Yes
0.868	0.466	0.776	0.868	0.82	0.727	No

3.4 Classification Rules

Significant rules [20] are extracted which are useful for understanding the data pattern and behaviour of experimental dataset. The following pattern is extracted by applying CART decision tree algorithm [14]. Some of the rules extracted for heart disease dataset are as follows,

1. Heartdisease(absence):- Thal=fixed_defect,Number_Vessels=0, Cholestoral =126-213.
2. Heart_disease(presence):- Thal=normal,Number_Vessels=0, Old_Peak=0-1.5, Max_Heart_Rate=137-169, Cholestoral=126-213.
3. Heart_disease(absence):- Thal=normal,Number_Vessels=0, Old_Peak=0-1.5, Max_Heart_Rate=137-169,Cholestoral=214-301, Rest=0, Pressure=121-147.

The rules for Hepatitis datasets are extracted and some of them are as follows

1. Ascites = Yes AND Histology = No: Live (46.0/1.0)
2. Anorexia = Yes ANDProtime > 47 AND Fatigue = No: Live (8.0)
3. Anorexia = Yes AND Malaise = Yes AND Ascites = Yes: Live (10.0/2.0)
4. Anorexia = Yes: Die (10.0) : Live (6.0)

Some classification rules for diabetes datasets are as follows,

1. Age <= 28 AND Triceps skin fold thickness > 0 AND Triceps skin fold thickness <= 34 AND Age > 22 AND No.timespreg <= 3 AND Plasma gc(2) <= 127: No (61.0/7.0)
2. Plasma gc(2) <= 99 AND 2-Hour serum insulin <= 88 AND 2-Hour serum insulin <= 18 AND Triceps skin fold thickness <= 21: No (26.0/1.0)
3. Age <= 24 AND Triceps skin fold thickness > 0 AND Body MI <= 33.3: No (37.0) Diastolic blood pressure <= 40 AND Plasma gc(2) > 130: Yes (10.0)
4. Plasma gc(2) <= 107 AND Diabetespf <= 0.229 AND Diastolic blood pressure <= 80: No (23.0)
5. No.timespreg <= 6 AND Plasma gc(2) <= 112 AND Diastolic blood pressure <= 88 AND Age <= 35: No (44.0/8.0)
6. Age <= 30 AND Diastolic blood pressure > 72 AND Body MI <= 42.8: No (41.0/7.0)

3.5 Comparison Of ID3, C4.5 and CART Algorithm

Algorithm designers have had much success with greedy, divide-and-conquer approaches to building class descriptions. It is chosen decision tree learners made popular by ID3, C4.5 (Quinlan1986) and CART (Breiman, Friedman, Olshen, and Stone 1984 [14]) for this survey, because they are relatively fast and typically they produce competitive classifiers. On examining the confusion matrices of these three algorithms, we observed that among the attribute selection measures C4.5 performs better than the ID3 algorithm, but CART performs better both in respect of accuracy and time complexity. When compared with C4.5, the run time complexity of CART is satisfactory.

Table 13: Prediction accuracy table

S.No	Name of algorithm	Accuracy %
1	CART Algorithm	83.2
2	ID3 Algorithm	64.8
3	C4.5 Algorithm	71.4

We have done this research and we have found 83.184% accuracy with the CART algorithm which is greater than previous research of ID3 and C4.5 as indicated in the table XVIII.

4. Conclusions

The decision-tree algorithm is one of the most effective classification methods. The data will judge the efficiency and correction rate of the algorithm. We used 10-fold cross validation to compute confusion matrix of each model and then evaluate the performance by using precision, recall, F measure and ROC space. As expected, bagging algorithms, especially CART, showed the best performance among the tested methods. The results showed here make clinical application more accessible, which will provide great advance in healing CAD, hepatitis and diabetes. The survey is made on the decision tree algorithms ID3, C4.5 and CART towards their steps of processing data and Complexity of running data. Finally it can be concluded that between the three algorithms, the CART algorithm performs better in performance of rules generated and accuracy. This showed that the CART algorithm is better in induction and rules generalization compared to ID3 algorithm and C4.5 algorithm. Finally, the results are stored in the decision support repository. Since, the knowledge base is currently focused on a narrow set of diseases. The approach has been validated through the case study, it is possible to expand the scope of modeled medical knowledge. Furthermore, in order to improve decision support, interactions should be considered between the different medications that the patient is on.

References

[1] UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html .

[2] American Diabetes Association, “Standards of medical care in diabetes—2007,” *Diabetes Care*, vol. 30, no. 1, pp. S4 S41, 2007.

[3] J. Du and C.X. Ling, “Active Learning with Generalized Queries,” Proc. Ninth IEEE Int’l Conf. Data Mining, pp. 120-128, 2009

[4] Jiawei Han and Micheline Kamber, “*Data Mining Concepts and techniques*”, 2nd ed., Morgan Kaufmann Publishers, San Francisco, CA, 2007.

[5] H.W. Ian, E.F., "Data mining: Practical machine learning tools and techniques," 2005: Morgan Kaufmann.

[6] R. Detrano, A.J., W. Steinbrunn, M. Pfisterer, J.J. Schmid, S. Sandhu, K.H.Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," American Journal of Cardiology,1989. 64: p. 304-310.

[7] G. John, "Models if incremental concept formation," Journal of Atificial Intelligence, 1989: p. 11-61.

[8] A. L. Gamboa, M.G.M., J. M. Vargas, N. H. Gress, and R. E. Orozco, "Hybrid Fuzzy-SV Clustering for Heart Disease Identification," in Proceedings of CIMCA-IAWTIC'06. 2006.

[9] D. Resul, T.I., S. Abdulkadir, "Effective diagnosis of heart disease through neural networks ensembles," Elsevier, 2008.

[10] Z. Yao, P.L., L. Lei, and J. Yin, "R-C4.5 Decision tree modeland its applications to health care dataset, in roceedings of the 2005 International Conference on Services Systems and Services Management," 2005. p. 1099-1103.

[11] K. Gang, P.Y., S. Yong, C. Zhengxin, "Privacy-preserving data mining of medical data using data separation-based techniques," Data science journal, 2007. 6.

[12] L. Cao, “*Introduction to Domain Driven Data Mining*,” Data Mining for Business Applications, pp. 3-10, Springer, 2009.

[13] Quinlan, J.R., "Induction of Decision Trees," Machine Learning. Vol. 1. 1986. 81-106.

[14] L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth Int. Group, 1984.

[15] S. R. Safavin and D. Landgrebe. A survey of decision tree classifier methodology. IEEE Trans. on Systems, Man and Cybernetics, 21(3):660-674, 1991.

[16] Kusrini, Sri Hartati, ”Implementation of C4.5 algorithm to evaluate the cancellation possibility of new student applicants at stmik amikom yogyakarta.” Proceedings of the International Conference on Electrical Engineering and Informatics Institut Technologic Bandung, Indonesia June 17-19, 2007.

[17] P. Magni and R. Bellazzi, “A stochastic model to assess the variability of blood glucose time series in diabetic patients self-monitoring,” *IEEE Trans. Biomed. Eng.*, vol. 53, no. 6, pp. 977–985, Jun. 2006.

[18] K. Polat and S. Gunes, “An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease,” *Dig. Signal Process.*, vol. 17, no. 4, pp. 702–710, Jul. 2007.

[19] J.Friedman, “Fitting functions to noisy data in high dimensions”, in Proc.20th Symp. Interface Amer. Statistical .Assoc. , E.J.Wegman.D.T.Gantz, and I.J. Miller.Eds.1988 pp.13-43

[20] T.W.simpson, C.Clark and J.Grelbsh ,”Analysis of support vector regression for appreciation of complex engineering analyses “, presented as the ASME 2003.

[21] L. B. Goncalves, M. M. B. R. Vellasco, M. A. C. Pacheco, and F. J. de Souza, “Inverted hierarchical neuro-fuzzy BSP system: A novel neuro-fuzzy model for pattern classification and rule extraction in LEE AND WANG: FUZZY EXPERT SYSTEM FOR DIABETES DECISION SUPPORT

APPLICATION 153 databases," IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 36, no. 2, pp. 236–248, Mar. 2006.

First Author D. Senthil Kumar is an Assistant Professor in the Department of Computer Science and Engineering in Anna University of Technology, Tiruchirappalli, India. He has completed 10 years of Teaching in various courses in the Undergraduate and Postgraduate Engineering & MBA program. He received a Master of Science in Mathematics from Presidency College, University of Madras and Master of Engineering in Systems Engineering And Operations Research from College of Engineering, Anna University (both located in Chennai, India). He received Prof. T.R. Natesan Endowment Award (Instituted by Operational Research Society Of India – Chennai Chapter). He is a member of IEEE and his research interest includes Optimization, Security and Data Mining.

Sathyadevi received the B.E degree in computer science and Engineering from Coimbatore Institute of Engineering and Information Technology in 2009. She is currently a M.E. candidate in the Department of Computer Science at Anna University of Technology, Tiruchirappalli. Her research interests include data mining, machine learning, and related real-world applications.

Third Author S.Sivanesh is an Assistant Professor in Computer Science and Engineering in Anna University of Technology, Tiruchirappalli, India. His research interests include Internet routing, routing security, network management and measurement.