An Empirical Study on the Performance of Integrated Hybrid Prediction Model on the Medical Datasets

Sarojini Balakrishnan
Professor, Department of Computer Applications,
K.L.N. College of Information Technology, Madurai, India

Ramaraj Narayanaswamy
Principal
G.K.M. College of Engineering & Technology
Chennai, India

Ilango Paramasivam
Professor
School of Computing Science & Engineering
VIT University, Vellore. India

ABSTRACT
The medical data are multidimensional and hundreds of independent features in these high dimensional databases need to be considered and analyzed, for valuable decision-making information in medical prediction. Most data mining methods depend on a set of features that define the behavior of the learning algorithm and directly or indirectly influence the complexity of the resulting models. Hence, to improve the efficiency and accuracy of mining task on high dimensional data, the data must be preprocessed. Feature selection is a preprocessing step which aims to reduce the dimensionality of the data by selecting the most informative features that influence the diagnosis of the disease. We propose a feature selection embedded Hybrid Prediction model that combines two different functionalities of data mining; the clustering and the classification. The F-score feature selection method and k-means clustering selects the optimal feature subsets of the medical datasets that enhances the performance of the Support Vector Machine classifier. The performance of the SVM classifier is empirically evaluated on the reduced feature subset of Diabetes, Breast Cancer and Heart disease data sets. The proposed model is validated using four parameters namely the Accuracy of the classifier, Area Under ROC Curve, Sensitivity and Specificity. The results prove that the proposed feature selection embedded hybrid prediction model indeed improve the predictive power of the classifier and reduce false positive and false negative rates. The proposed method achieves a predictive accuracy of 98.9427% for diabetes dataset, 99% for cancer dataset and 100% for heart disease dataset, the highest predictive accuracy for these datasets, compared to other models reported in the literature.

General Terms
Data Mining, Dimensionality Reduction, Feature selection, Prediction Model

Keywords
Medical Data Mining, F-score, Support Vector Machine Classifier, Accuracy, Sensitivity, Specificity, Area Under ROC Curve

1. INTRODUCTION
Medical data mining is becoming increasingly important in healthcare. The diversity of medical data collected/stored for diagnosis and prognosis and the availability of widespread data mining techniques to process these data place medical data mining in a unique position to truly impact patient care using these stored data. The application of data mining in medicine has proved successful in the areas of diagnosis, prognosis and treatment [1]. The discovered patterns may represent valuable knowledge that could lead to medical discoveries, for example identification of combinations of features that lead to diagnosis of the disease. Studies show that improved medical diagnosis and prognosis may be achieved through automatic analysis of patient data stored in medical records i.e. by learning from past experiences [2]. The medical databases are characterized by the particular constraints and difficulties of the privacy-sensitive, heterogeneous, but voluminous data of medicine [3]. Hundreds of independent features (parameters) in these high dimensional databases need to be simultaneously considered and analyzed, for valuable decision-making information in medical prediction. Medical databases may contain data with characteristics such as in-completeness (missing values), incorrectness (noise in data), sparseness (few and/or non-representable patient records), and inexactness (inappropriate selection of parameters for a given task). Research shows that the inclusions of redundant, irrelevant features cause the predictive performance of data mining algorithms to decline [4]. Data pre-processing is required to prepare the data for data mining and machine learning to increase the predictive accuracy. The application of efficient and sound data preprocessing procedures could reduce the amount of data to be analyzed without losing any critical information, improve the quality of the data, enhance the performance of the actual data mining algorithms and reduce the execution time of mining algorithms [5]. Feature subset selection, a robust pre-processing technique, based on the principle of parsimony (or Occam's razor) [6], chooses the feature subset that maximizes the accuracy of prediction. Feature selection is an important issue in building a better classification system. It is advantageous to use the feature selection process in the classification problems to limit the number of input features in order to improve the performance and the computation cost of the classifier [7, 8]. The aim of the feature selection is to identify the dataset containing the smallest number of non-redundant features which gave the best result. Feature selection involves searching through various feature subsets and evaluating each of these subsets using some criterion [9, 10, 11]. Feature selection in medical data mining is appreciable as the diagnosis of the disease could be done in this patient-care activity with minimum number of features that is to say with minimum number of clinical testings thereby reducing the cost and time. Though a number of feature selection methods that enhance the performance of the classifier are available, still the research goes on to identify more informative

features of a dataset. Also, the discriminating ability of the classifier measured as Sensitivity and Specificity which is very much important in medical domain is considered for evaluation. The objective of this research work is to show that selecting the more significant features for medical diagnosis helps the physician to make accurate diagnosis. The focus is on aggressive dimensionality reduction with an increase in the prediction accuracy. The empirical results show that the feature selection embedded hybrid prediction model achieves significant dimensionality reduction in the medical datasets viz., Pima Indian Diabetes dataset, Breast Cancer dataset and Cleveland Heart Disease dataset obtained from the UCI Machine Learning repository [12]. The accuracy of the SVM classifier is the highest predictive accuracy for these datasets compared to other models in literature for this problem.

2. LITERATURE REVIEW ON PREDICTIVE CLASSIFICATIONS IN MEDICINE

Majority of research papers published in the area of feature selection and predictive classification deal with the goal of improving accuracy. A Lot of research has been done on Pima Indian Diabetes disease diagnosis, Breast Cancer diagnosis and Cleveland heart disease diagnosis. The accuracy of the classification is used as the criteria for measuring the performance of the classifier. When the studies in the literature related with this classification application are examined, it can be seen that a great variety of methods were used which reached high classification accuracies using the dataset taken from UCI machine learning repository.

A number of different classification algorithms using the Pima Indian diabetes dataset have achieved accuracy in the range of 59.5–77.7%. The classification accuracy of 82.05% was attained by Polat, Gunes, and Aslan who presented a cascade learning system based on Generalized Discriminate Analysis (GDA) and Least Square Support Vector Machine (LSSVM) to the diagnosis of Pima Indian diabetes disease [13]. A Hybrid Prediction Model proposed by Patil B.M. et al., attained a predictive of accuracy of 92.38% [14]. The classification accuracies using various methods for Pima Indians Diabetes dataset are discussed by Polat et.al, 2008 and Patil. B.M. et.al., 2010.

A new decision making system based on combining fuzzy weighted pre-processing for feature selection and Artificial Immune Recognition System(AIRS) classifier proposed by Polat et al. to classify the heart disease dataset has achieved 92.59% classification accuracy with 50–50% training–testing split [15]. Polat et al. achieved 87% classification accuracy using a hybrid system based on Artificial Immune Recognition System (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting pre-processing on the diagnosis of heart disease [16]. Polat et al. have diagnosed the heart disease using a hybrid expert system combining AIRS classifier and fuzzy weighted pre-processing and obtained 96.39% classification accuracy with 10-fold cross-validation [17]. Özsen et al. proposed a novel classification algorithm called Artificial Immune System (AIS) with Hybrid feature vectors and applied to heart disease diagnosis. They obtained 83.95% classification accuracy [18]. Kahramanli et al. obtained 86.8% classification accuracy using fuzzy neural network on the diagnosis of Cleveland heart disease [19].

As for other clinical diagnosis problems, classification systems have been used for breast cancer diagnosis problem, also. A detailed literature survey by Polat et.al clearly depict the accuracy obtained by using various classification algorithms [20]. Among these, Setiono achieved a maximum accuracy of 98.1% using neuro-rule method [21]. Polat et.al achieved the highest accuracy of 98.53% for Breast cancer diagnosis using LSSVM [20].

3. DATASETS

3.1 Pima Indian Diabetes Dataset

The Pima Indian Diabetes data set was selected from a larger data set held by the National Institutes of Diabetes and Digestive and Kidney Diseases. All patients in this database are Pima Indian women at least 21 years old and living near Phoenix, Arizona, USA.

There are eight clinical findings: 1.Number of times pregnant 2. Plasma glucose concentration a 2 h in an oral glucose tolerance test 3. Diastolic blood pressure (mm Hg) 4. Triceps skin fold thickness (mm) 5. Two hour serum insulin (mu U/ ml) 6. Body mass index 7. Diabetes pedigree function 8. Age (years). The binary response variable takes the values ‘0’ or ‘1’, where ‘1’ means a positive test for diabetes and ‘0’ is a negative test for diabetes. There are 268 (34.9%) cases in class ‘1’ and 500 (65.1%) cases in class ‘0’. A preliminary analysis of the Pima Indians Diabetes dataset indicates usage of zero for missing data. Also, the numbers of missing values for the features serum–insulin and triceps skin fold are very high (374 and 227, respectively out of total 768 instances). So, both these features are removed and the instances which have missing values for other features are also eliminated. After removing all the above said values and features, only 625 instances and six features are taken up for further study.

3.2 Wisconsin Breast Cancer Dataset

The WBC (Wisconsin breast cancer) dataset consists of 699 samples that were collected by Dr.W.H.Wolberg at the University of Wisconsin–Madison Hospitals taken from needle aspirates from human breast cancer tissue. The WBCD database consists of nine features obtained from fine needle aspirates, each of which is ultimately represented as an integer value between 1 and 10. The features are (1)clump thickness, (2) uniformity of cell size, (3) uniformity of cell shape, (4) marginal adhesion, (5) single epithelial cell size, (6) bare nucleoi, (7) bland chromatin, (8) normal nuclei, (9) mitoses. The dataset is preprocessed to remove missing values. After removing 16 instances with missing values, 444 instances of benign, and 239 instances of malignant are taken up for further study.

3.3 Heart Disease Dataset

This database is taken from the Cleveland Clinic Foundation and was supplied by Robert Detrano, M.D., Ph.D. of the V.A. Medical Center, Long Beach, CA. It is part of the collection of databases at the University of California, Irvine collected by David Aha [22]. The aim of the dataset is to classify the presence or absence of heart disease given the results of various medical tests carried out on a patient. This database consists of 13 features. These features are (1) age: (in the range of 29 and 77), (2) sex: (Male, Female), (3) chest pain type (4 values: 1, 2,

3, 4), (4) resting blood pressure: (in the range of 94 and 200), (5) serum cholesterol in mg/dl (in the range of 126 and 564), (6) fasting blood sugar >120 mg/dl (in the range of 126 and 564), (7) resting electrocardiograph results (values 0, 1, 2), (8) maximum heart rate achieved (in the range of 71 and 202), (9) exercise induced angina (either 0 or 1), (10) old peak = ST depression induced by exercise relative to rest (in the range of 0 and 6.2), (11) the slope of the peak exercise (ST segment: 1, 2, and 3), (12) number of major vessels (0–3) colored by fluoroscopy: 0–3, (13) thal: 3 = normal; 6 = fixed defect; 7 = reverse defect. The database originally contained 303 examples but 6 of these contained missing values and so were discarded leaving 297. The main criterion that physicians use to determine the diagnosis of heart disease is the narrowing in diameter of any major blood vessel. The diagnosis was considered to be positive (presence of heart disease) if the diameter of any major vessel was narrowed by more than 50%; and negative otherwise. The dataset contains 137 positive cases and 160 negative cases, after removing the instances with missing attribute values.

4. PROPOSED SYSTEM

The Proposed system consists of three stages: Selection of informative features that contribute to the performance of the classification algorithm, Extraction of pattern using clustering; and Classification using Support Vector Machine.

The three steps of the proposed model is

1. The informative features of the datasets are selected using F-Score method
2. Patterns extraction is performed using K-means clustering
3. The performance of the SVM classifier using extracted data is empirically evaluated

The performance of the SVM classifier is validated in terms of Accuracy, Sensitivity and Specificity and Area Under the ROC Curve. The proposed model is evaluated by comparing the performance of the classifier before and after feature selection.

4.1 Feature Selection using F-score Method

One way of selecting the relevant features from the data available is to estimate their influence in diagnosis decision. The F-score method proposed by Chen and Lin, 2005 uses the F-score values to measure the discriminating power of individual features in the database in respect to class labels [23]. The F-score values of each feature in dataset are computed and the features with relatively high F-scores are considered as "informative". The F-Score values are estimated using the following equation (Eq. (1)).

Let X_k, k = 1, . . . ,m, be the training vectors and n^+ and n^- be the number of positive and negative instances respectively, the F-score F(i) of the i^{th} feature is defined as:

$$F(i) = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 - (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n^+ - 1}\sum_{k=1}^{n^+}(x_{k,i}^{(+)} - \bar{x}^{(+)})^2 + \frac{1}{n^- - 1}\sum_{k=1}^{n^-}(x_{k,i}^{(-)} - \bar{x}^{(-)})^2}$$

Where $\bar{x}_i, \bar{x}_i^{(+)}, \bar{x}_i^{(-)}$ are the average of the i^{th} feature of the whole, positive and negative datasets respectively. $x_{k,i}^{(+)}$ is the i^{th} feature of the k^{th} positive instance and $x_{k,i}^{(-)}$ is the i^{th} feature of the k^{th} negative instance. The numerator indicates the discrimination between the positive and negative sets, and the denominator indicates the one within each of the two sets. The estimated F-score values of the features are arranged in decreasing order of importance. The least rank features are eliminated one at a time from backwards and the performance of the clustering algorithm is observed. The contribution of the features in accurate clustering is used as the criteria to select the optimal feature subset. The feature subset that enhances the performance of the cluster is the optimal feature subset of the dataset. Tables 1, 2 and 3 show the F-Score values of the three medical datasets.

Table 1. F-score values of six features of Pima Indians Diabetes Dataset

Features	F-scores
f_2: Glucose tolerance test	0.278279
f_6:Body mass index	0.093697
f_8:Age	0.060236
f_1: No. of times pregnant	0.051789
f_7: Diabetes pedigree function	0.031164
f_3: Diastolic blood pressure	0.004252

From Table 1 it is evident that the features No. of times pregnant, Age, Body mass index and Glucose tolerance test have yielded F-score values above the threshold value of 0.05.

Table 2. F-score values of nine features of Breast Cancer Dataset

Features	F-scores
f_6: Bare Nuclei	2.094332
f_3: Uniformity of Cell Shape	2.081709
f_2: Uniformity of Cell Size	2.064805
f_7: Bland Chromatin	1.352438
f_8:Normal Nucleoli	1.068239
f_1:Clump Thickness	1.044675
f_4:Marginal Adhesion	0.995416
f_5: Single Epithelial Cell Size	0.913594
f_9:Mitoses	0.218484

The feature Mitoses of the dataset has comparatively low F-score compared to other features of the dataset.

Table 3. F-score values of the thirteen features of Heart Dataset

Features	F-scores
f_{13}:Thal	0.380541
f_{12}:Ca	0.261561
f_9:Exang	0.213319
f_8: Thalach	0.212347
f_{10}: Oldpeak	0.211676
f_3:Chest pain type	0.211025
f_{11}:Slope	0.128648
f_2:Sex	0.097258
f_1:Age	0.047209
f_7:Restecg	0.034294
f_4:Trestbps	0.024741
f_5:Chol	0.014126
f_6:Fasting blood sugar	0.000266

The features Age, Restecg, trestbps, cholesterol, fasting blood sugar have F-score values below the threshold value of 0.05. These features with low F-score values are removed one at a time and the influence of the feature in reducing the clustering error is used as the performance indicator to derive the optimal feature subset.

4.2.Pattern Extraction by k-means Clustering

Clustering is used to find hidden patterns in data and choosing subsets of features that contain good patterns remains a challenging research problem. Clustering can also be used as a reduction technique by storing the characteristics of the clusters instead of the individual data [24]. In our proposed approach we use simple k-means clustering algorithm implemented in the Weka tool [25] to group the similar instances. K-means is one of the popular partitioning algorithms which use Euclidean distance as the dissimilarity method. The features with low F-scores are removed one at a time and the clustering error is used as the performance indicator to determine the optimal feature subset. The reduced feature subset that gives the minimal clustering error is considered to be the optimal feature subset. The features in the optimal feature subset represent the cluster patterns for class 'yes' and 'no'. The empirical results of clustering prove that by applying feature selection, the clustering error is reduced. This method of feature selection achieves a feature reduction of 50% for Pima dataset, 33.3% for breast cancer dataset and 36% for Cleveland dataset.

Table 4. Confusion matrix for Breast Cancer dataset

Data sets	Pima Indian Diabetes Dataset			Breast Cancer Dataset			Cleveland Heart Disease Dataset		
	Feature subset	Number of instances incorrectly clustered	Error in Clustering (%)	Feature Subset	Number of instances incorrectly clustered	Error in Clustering (%)	Feature Subset	Number of instances incorrectly clustered	Error in Clustering (%)
Before Feature Selection	6 {F_1,F_2,F_3 ,F_6,F_7,F_8 }	192	30.7692	9 {F_1,F_2,F_3, F_4,F_5,F_6,F_7,F_8, F_9}	27	3.9531	14 {F_1,F_2,F_3, F_4,F_5,F_6,F_7 ,F_8,F_9,F_{10}, F_{11},F_{12},F_{13}, F_{14}}	32	10.7383
After Feature Selection	3 {F_2,F_6,F_8 }	159	25.4808	6 { F_1,F_2,F_3, F_6,F_7,F_8 }	26	3.8067	9 {F_2,F_7,F_8, F_9,F_{10},F_{11}, F_{12},F_{13},F_{14} }	5	1.8797

The results of clustering are shown in Table 4. The 466 instances and the three informative features of Diabetes dataset, 657 instances and six features of Breast cancer dataset and 497 instances and nine features of Heart disease dataset determine the pattern for diagnosis for the presence/ absence of the disease.

4.3.Supervised Classification using SVM Classifier

The optimal feature subset derived by F-score feature selection and k-means clustering algorithm determine the pattern for diagnosis. The RBF kernel of SVM is used as Classifier for our proposed approach as RBF kernel function can analyze higher-dimensional data and requires only two parameters, C and γ to be defined [26]. The parameter selection tool of the LibSVM classifier, Cross validation via parallel grid search [27] is used to find the best values of C and γ. The grid points are chosen on a logarithmic scale and classifier accuracy is estimated for each point on the grid. Grid search is done by specifying the parameter space, the range of C, γ and the stopping tolerance. To avoid bias in the estimation of accuracy, 10-fold cross validation is used. The training data is separated into 10 folds. Sequentially a fold is considered as the validation set and the rest are for training. The average accuracy of prediction across the validation sets is the cross validation accuracy. The performance

of the SVM classifier for the optimal feature subset derived is observed to validate the approach. The selected feature subset is validated in terms of increase in the accuracy of the classifier, an improved Area Under Receiver Operating Characteristic curve and Specificity and Sensitivity values.

4.4.Empirical Results and Performance Analysis

The empirical results of SVM classification on the optimal feature subsets of Pima Indian Diabetes, Wisconsin Breast Cancer and Cleveland Heart Disease datasets are given in Table 5, 6 and 7 in the form of confusion matrix. The confusion matrix is a visualization tool commonly used to present performances of classifiers in classification tasks [27]. It shows the relationships between real class attributes and that of predicted classes. The confusion matrix depicts the prediction ability of the classifier in correctly classifying positive and negative classes of the dataset. The True positives and True Negatives represent the correct classifications of "yes' 'no' class and False Negative and False Positive represent the incorrect classifications. A False positive (FP) occurs when the outcome is incorrectly predicted as yes (or positive) when it is actually no (negative). A False Negative (FN) occurs when the outcome is incorrectly predicted as no when it is actually 'yes'. The level of effectiveness of the classification model is calculated with the number of correct and incorrect classifications in each possible value of the variables being classified in the confusion matrix [28]. Out of 466 correctly clustered instances of Pima Indian Diabetes dataset 461 instances are correctly classified. Out of 657 correctly clustered instances of Breast cancer dataset 651 instances are correctly classified. In Heart disease dataset all the 497 correctly clustered instances are correctly classified.

Table 5. Confusion matrix for Diabetes dataset

Actual Class	Predicted Class	
	Yes	No
Yes = 150	149 (True Positive)	1(False Negative)
No = 316	4 (False Positive)	312(True Negative)

Table 6. Confusion matrix for Breast Cancer dataset

Actual Class	Predicted Class	
	Yes	No
Yes = 436	430 (True Positive)	6(False Negative)
No = 221	0 (False Positive)	221(True Negative)

Table 7. Confusion matrix for Heart Disease dataset

Actual Class	Predicted Class	
	Yes	No
Yes = 137	137(True Positive)	0(False Negative)
No = 160	0 (False Positive)	160(True Negative)

From the results obtained the following equations are used to measure the Accuracy (Eq.(1)), Sensitivity (Eq.(2)), and Specificity (Eq.(3))[29].

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Number of\ Instances} \quad (1)$$

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

$$Specificity = \frac{True\ Negative}{False\ Positive + True\ Negative} \quad (3)$$

The accuracy of the SVM classifier is evaluated using 10-fold cross-validation test. Cross-validation involves breaking a dataset into 10 pieces, and on each piece, testing the performance of a predictor build from the remaining 90% of the data. The classification accuracy was taken as the average of the 10 predictive accuracy values. Sensitivity and Specificity are statistical measures that describe how well the classifier discriminates between a case with positive and with negative class (with and without disease). Sensitivity is the detection of disease rate that needs to be maximized and (1 – Specificity) is the false alarm rate that is to be minimized for accurate diagnosis. The tradeoff between Sensitivity and (1-Specificity), as well as the performance of the classifier, can be visualized and studied using the Receiver Operating Characteristic (ROC) curve. A perfect classifier provides an AUC that equals 1. The empirical results in Table 8 show the results of SVM classification.

Table 8. Classification results of SVM Classifier

Datasets	Accuracy %	Sensitivity %	Specificity %	AUC
Pima Indian Diabetes	98.9247	99.33	98.73	0.9997
Breast Cancer	99	98.62	1	0.999
Cleveland Heart Disease	100	1	1	1

The obtained results prove that selecting the discriminative features for classification has indeed improved the performance of the classifier. The proposed method achieves the highest accuracy for the datasets compared to other methods reported in the literature.

5. CONCLUSION

This research work attempts to emphasize embedding feature selection methods in clinical decision support tools that could empower the medical community to improve the quality of diagnosis through the use of technology. In medical domain reduce in the number of features means reduce in the number of clinical measures to be made and diagnose of the disease with less number of more discriminating features. In this research we propose an approach to identify significant features of the medical datasets that improve the performance of the classifier in accurate classification. The empirical results prove that the optimal feature subset derived for the datasets improve the classification accuracy and two other vital parameters of the medical domain the sensitivity and specificity.

6. REFERENCES

[1] Dilly Ruth, 2002. Data Mining - An Introduction. Available at http://www.pcc.qub.ac.uk/tec/courses/datamining/stu_ notes/dm_book_1.html.

[2] Lavrac N, 1999. Selected techniques for data mining in medicine. Artificial Intelligence in Medicine 16(1), 3-23.

[3] Cios K J and Moore G W, 2002. Uniqueness of medical data mining. Artificial Intelligence in Medicine 26(1-2), 1-24.

[4] Wu X, Holmes G and pfahringer B, 2008. Mining arbitrarily large datasets using heuristic k-nearest neighbor search. In Wobcke W and Zhang M, (Eds) Proc. of Twenty-First Australian Joint conference on Artificial Intelligence, Advances in Artificial Intelligence(AI 2008). LNAI 5360. Auckland, NZ: Springer, 355-361.

[5] Paraskevas Orfanidis and David J. Russomanno, 2008. Preprocessing enhancements to improve data mining algorithms. International Journal of Business Intelligence and Data Mining 3(2), 196-211.

[6] D.A. Bell, H. Wang, 2004. A formalism for relevance and its application in feature subset selection, Machine Learning 41, 175–195.

[13] Polat, K., Gunes, S., & Aslan, A., 2008. A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. Expert Systems with Applications, 34(1), 214–221.

[14] Patil, B. M., et al., 2010. A Hybrid Prediction Model for Type-2 Diabetic Patient. Expert Systems with Applications, doi:10.1016/j.eswa.2010.05.078.

[15] Polat, K., & Günes, S., 2007. A hybrid approach to medical decision support systems: Combining feature selection, fuzzy weighted pre-processing and AIRS. Computer Methods and Programs in Biomedicine, 88(2), 164–174.

[16] Polat, K., Sahan, S., & Günes, S., 2007. Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing. Expert Systems with Applications, 32(2), 625-631.

[17] Polat, K., Tosun, S., & Günes, S. (2006). Diagnosis of heart disease using artificial immune recognition system and fuzzy weighted preprocessing. Pattern Recognition, 39(11), 2186–2193.

[18] Özs_en, S., & Günes S, 2008. Effect of feature-type in selecting distance measure for an artificial immune system as a pattern recognizer. Digital Signal Processing, 18(4), 635–645.

[19] Kahramanli, H., & Allahverdi, N. (2008). Design of a hybrid system for the diabetes and heart diseases. Expert Systems with Applications, 35(1–2), 82–89.

[20] Polat,k, Güne¸s.S , 2007. Breast cancer diagnosis using least square support vector machine. Digital Signal Processing 17, 694–701.

[21] Setiono, R., 2008. Generating concise and accurate classification rules for breast cancer diagnosis, Artific. Intell. Med. 18, 205–219.

[7] Zhang G P, 2000. Neural Networks for Classification: A Survey. IEEE Trans. on Systems Man, and Cybernetics Part C: Applications and Reviews 30(4), 451–462.

[8] Cao, Bin., Shen, Dou., Sun, Jian-Tao., Yang, Qiang., Chen, Zheng. (2007). Feature selection in a kernel space. In International conference on machine learning (ICML) Oregon, USA, June 20–24, pp. 121–128.

[9] Liu H and Motoda H,1998. Feature Selection for Knowledge Discovery and Data Mining. Boston: Kluwer Academic Publishers.

[10] Pena, J. M., Lozano, J. A., Larranaga, P., & Inza, I., 2001. Dimensionality reduction in unsupervised learning of conditional gaussian networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(6), 590–603.

[11] Yu, L., & Liu, H., 2003. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th international conference on machine learning (pp. 856–863).

[12] Asuncion A and Newman D J, 2007. UCI Machine Learning repository. [http://www.ics.uci.edu/~mlearn/MLRepository.html].University of California, Irvine, CA.

[22] Kurgan, Lukasz A., Cios, Krzysztof J., Tadeusiewicz, Ryszard, Ogiela, Marek, & Doodenday, Lucy S. (2001). Knowledge discovery approach to automated cardiac SPECT diagnosis. Artificial Intelligence in Medicine, 149–169.

[23] Chen Y W and Lin C J, 2005. Combining SVMs with Various Feature Selection Strategies. Available at www.csie.ntu.edu.tw/~cjlin/papers/features.pdf.

[24] Guojun, G., Chaoqu, M., & Jianhong, W., 2007. Data clustering theory algorithm and application (1st ed.). ASA-SIAM.

[25] Witten, H. I., & Frank, E., 2005. Data mining: Practical machine learning tools and techniques (2nd ed.). Morgan Kaufmann Publishers.

[26] Cheng-Lung Huang, Hung-Chang Liao b, Mu-Chen Chen c, 2008. Prediction model building and feature selection with support vector machines in breast cancer diagnosis, Expert Systems with Applications, 578-587 doi:10.1016/j.eswa.2006.09.041

[27] Hsu C W and Lin C J, 2002**.** A simple decomposition method for support vector machine. Machine Learning 46(1-3), 219–314.

[28] Yang J and Honavar V, 2001. Feature Subset Selection Using A Genetic Algorithm. In Feature Extraction, Construction and Selection: A Data Mining Perspective. 117-136, 1998, second printing.

[29] Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. Artificial Intelligence in Medicine, 34, 472 113–127