International Journal of Scientific & Engineering Research Volume 3, Issue 9, September-2012
ISSN 2229-5518

A Survey on Predictive Data mining Approaches for Medical Informatics

Dr. S.Appavu Alias Balamurugan,S.Sasikala, Dr.S.Geetha

Abstract- Among various data mining techniques, classification analysis is widely adopted for supporting medical diagnostic decisions. Medical diagnosis is considered as a classification problem: a record represents a given patient's case, predictor features are all patients' data and the class label is the diagnosis. Subsequently, the built classification model is essential and used to predict appropriate classes for novel and uncategorized cases. Medical data often contain irrelevant features and noise. Feature selection is frequently adopted to identify and remove the irrelevant and redundant information as much as possible. The selection of appropriate subset of the available features can yield a compact and easily interpretable representation of the target concept, model the target task adequately, and improve the classification accuracy especially in medical region .The Medical diagnosis is a complex and dynamic system with noisy, non stationary and chaotic data series. The aim of this paper is to explain the potential day by day research contributions of data mining to solve the complex problem of Medical diagnosis prediction. This study paper synthesizes five significant works and explains how data mining is gaining popularity in medical field.

Keywords: Medial diagnosis prediction, feature selection, classification, Health care domain , Dimensionality Reduction, Medical Informatics, Medical Data sets

I. INTRODUCTION

Data mining techniques have become an apparent need in medical applications. In addition to the large pool of techniques that have already been developed in the machine learning and data mining fields, specific applications in medical field have led to a wealth of newly proposed techniques[1]. In an attempt to achieve effective and objective standard of diseases prediction, many researchers have used data mining approach to construct the classifier for medical dataset. The main aim of this review is to make practitioners aware of the benefits, and in some cases even the necessity of data mining techniques. This paper reviews the most important application fields in the Medical (Health care) domain, highlighting the efforts done by the Medical community in developing novel and adapted data mining procedures for enhanced diagnosis accuracy.

A. Feature Selection Techniques in Medical Informatics

Even though data mining framework has many component phases like Pre-processing, Feature Extraction, Feature Selection and Classification, the latter couple of phases play a significant role in improving the medical diagnosis prediction. In many medical applications, enough features may be available to allow training of all base classifiers on mutually exclusive feature subsets. However, the feature subsets used should be carefully selected to ensure good quality of the individual classifiers, as well a high degree of diversity and independence amongst them to encourage constructive disagreements that enhance ensemble performance [2]. Feature selection for this purpose is somewhat different from traditional feature selection used for data reduction in areas characterized by high dimensionality due to the large number of available features, e.g. in remote sensing [3], seismic data processing [4], speech recognition [5], and drug design [6].

In single classifier applications, dimensionality reduction attempts to select a smaller subset of optimum features by excluding irrelevant and redundant features in order to avoid overfitting, improve performance, and speedup both training and prediction for the classifier. This form of feature reduction has been applied to several areas in medicine, including: classification of EEG signals for operating brain—computer interfaces [7], detection of mass lesions in digital mammograms [8], segmenting digital chest radiographs [9], and detection of seizure events in newborn children using EEG data [10].

B. Classification Techniques in Medical Informatics

In data mining, classification is another important task. It maps the data into predefined targets. It is a supervised learning as targets are predefined. The aim of the classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. Then, the classifier is used to predict the group attributes of new cases from the domain based on the values of other attributes.

Classification has been successfully applied to many areas, such as scientific experiments, medical diagnosis, fraud detection, credit approval, target marketing, computer and network system security, etc.

The purpose of classification on medical informatics is to predict the presence or absence of the particular disease based on

- *Dr.S.APPAVU Alias Balamurugan has received Ph. D. degree from Anna university, Chennai ,India.. He is a Professor & Research coordinator at K.l.N.College of Information Technology, Sivagangai ,India..His research interests include data mining and text mining. E-mail: app_s @yahoo.com.*
- *S.SASIKALA is Full time Research Scholar under Information and Communications, Anna University of Technology,Madurai,India E-mail: nithilannsasikala@yahoo.co.in*
- *Dr.S.GEETHA has received Ph. D. degree from Anna university, Thirchy, India. She is an Assiatant Professor at Thiagarajar College of Engineering, Madurai,India.Her research interests include data mining and Stegnography. E-mail: geethabaalan@gmail.com*

the classification model built by training data. Classification in medical diagnosis is relatively more crucial than any other domain because further treatment for the patient heavily depends on the accuracy of the diagnosis.

Numerous classification methods are available in the literature including probabilistic learner, Naive Bayes , a decision tree learner, C4.5 , an instance-based learner, IB1 , Support Vector Machine [14], Artificial Neural Networks , Rough-Sets [15] etc. These algorithms have proved effective in practice and in particular in the clinical domain. In some cases, a classifier paradigm can both be a feature selector as well as a classifier.

The following sections review the significant works on feature selection and classification on medical informatics under the following perspectives – medical data under diagnosis, feature selection and the classifier techniques employed and the major research findings. The commonalities in all the works and the future scope are also discussed at the end of this paper.

II. REFERENCE PAPER DESCRIPTION:

A. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis(Hui-Ling Chen , Bo Yang , Jie Liu, Da-You Liu , 2011)

Medical dataset description:

The term "breast cancer" refers to a malignant tumour that has developed from cells in the breast. In this study, the authors have performed empirical study on the WBCD (Wisconsis Breast Cancer Dataset) taken from UCI machine learning repository (UCI Repository of Machine Learning Databases). The dataset contains 699 instances taken from needle aspirates from patients' breasts, where 16 instances have missing values. Because of the small number of missing data, these cases are discarded from data set and remaining 683 cases are used in the experiment, of which 444 cases belong to benign class and the remaining 239 cases belong to malignant class. Each record in the database has nine attributes.
These nine attributes were found to differ significantly between benign and malignant samples. The nine attributes listed in Table 1 are graded 1–10, with 10 being the most abnormal state. The class attribute was represented as 2 for benign and 4 for malignant cases.

TABLE I

THE DETAIL OF THE NINE ATTRIBUTES OF BREAST CANCER DATA.

C1	Clump Thickness	1-10
C2	Uniformity of cell size	1-10
C3	Uniformity of cell shape	1-10
C4	Marginal Adhesion	1-10
C5	Single Epithelial cell size	1-10
C6	Bar nuclei	1-10
C7	Bland chromatin	1-10
C8	Normal Nucleoli	1-10
C9	Mitoses	1-10

Diagnosis Results:

In this paper, a rough set (RS) based supporting vector machine classifier (RS_SVM) is proposed for breast cancer diagnosis. In the proposed method (RS_SVM), RS reduction algorithm is employed as a feature selection tool to remove the redundant features and further improve the diagnostic accuracy by SVM. The effectiveness of the RS_SVM is examined on Wisconsin Breast Cancer Dataset (WBCD) using classification accuracy, sensitivity, specificity, confusion matrix and receiver operating characteristic (ROC) curves. It was observed that the proposed method achieved the highest classification accuracies (99.41%, 100%, and 100% for 50–50% of training-test partition, 70– 30% of training-test partition, and 80–20% of training-test partition, respectively) for a subset that contained five features (subset #5).

Major Findings:

- This work has explored a new expert system, RS_SVM, for breast cancer diagnosis.
- A combination of five features (i.e., 'Clump Thickness', 'Uniformity of Cell Shape', Marginal Adhesion', 'Bare Nuclei' and 'Mitoses') for classifying breast tumours was identified to be most informative by RS-based reduction algorithm. It implied that these five features were worthwhile to be taken close attention by the physicians when they conducted the diagnosis.

B. Diagnosis of bladder cancers with small sample size via feature selection (T. Warren Liao, 2011)

Medical dataset description:

In this paper, the wrapper approach is used to select the best feature subsets for the bladder cancer with small sample size that was used by Li et al. [12] and Luukka [14]. The data set consists of nine examples with bladder cancer and nine without. Each example is characterized by 13 kinds of gene expression related to bladder cancer, including both oncogenes and tumor suppressor genes. Note that both previous studies used a different class index. In Li et al., a class index between [0.5, 1.5) was randomly assigned to nine patients with bladder cancers while a random class index between [_0.5, 0.5) was given to other nine normal patients. On the other hand, Luudda assigned nine patients with bladder cancers to 'class 1' and nine patients without to 'class 2.'

Diagnosis Results:

They include three mutual correlation based feature selection methods (MC1, MC2, and MC3 for the first, second, and third mutual correlation-based method), seven gene selection criteria (WTS1 and WTS2 for the two versions of Welch t-statistic, FCS1 and FCS2 for the two versions of Fisher correlation score, ICE for independently consistent expression, MDS for mean difference score, and ADS for average difference score), and the Relief algorithm. The MC1 method starts with the feature subset empty and selects one feature at a time from the remaining features to be added based on a selection criterion, which considers feature–feature correlations as well as feature class correlations. The MC2

method starts with the feature subset full and selects one feature at a time from the feature subset to be removed based on another selection criterion, which considers only feature–feature correlations. For each feature subset size, the MC3 method evaluates each possible combination based on a merit criterion and selects the one with the highest merit.

Major Findings:

- The performances of 3-layer feed-forward neural networks and similarity classifiers trained by samples randomly selected from the small data set using only one single feature were tested and found to achieve equal or better performances than the previous studies with one requiring the generation of additional samples and another requiring the use of all 13 features.
- The results also help rank the performance of each individual of the four features identified above and the descending order of goodness is feature 13, 12, 10 and 3.

C.Feature selection from nocturnal oximetry using genetic algorithms to assist in obstructive sleep apnoea diagnosis (Daniel Alvarez, Roberto Hornero, J. Victor Marcos, Felix del Campo, 2011)

Medical dataset description:

The obstructive sleep apnoea (OSA) syndrome is a sleep-related disorder characterised by frequent breathing pauses, which lead to deep oxy-haemoglobin desaturations, blood pressure and heart rate acute changes, increased sympathetic activity and cortical arousals [15]. A wide variety of significant consequences affect people suffering from OSA including hyper somnolence, neuro-cognitive dysfunction, metabolic deregulation or respiratory failure. Moreover, OSA is frequently linked with conditions associated to the main causes of mortality in adults, such as hypertension, stroke or myocardial infarction [15]. It is estimated that approximately 20% of adults have at least mild OSA and 7% of adults have moderate-to-severe OSA. Unlike its high prevalence, 90% of cases in men and 98% of cases in women may go undiagnosed for many years.

Nocturnal pulse oximetry (NPO) has demonstrated to be a powerful tool to help in obstructive sleep apnoea (OSA) detection. The authors have exhaustively analysed a database of blood oxygen saturation (SpO2) recordings (80 OSA-negative and 160 OSA-positive) to obtain further knowledge on the usefulness of NPO. Population set was randomly divided into training and test sets. A feature extraction stage was carried out: 16 features (time and frequency statistics and spectral and nonlinear features) were computed.

Diagnosis Results:

A genetic algorithm (GA) approach was applied in the feature selection stage. The methodology achieved 87.5% accuracy (90.6% sensitivity and 81.3% specificity) in the test set using a logistic regression (LR) classifier with a reduced number of complementary features (3 time domain statistics, 1 frequency domain statistic, 1 conventional spectral feature and 1 nonlinear feature) automatically selected by means of GAs. Their results improved diagnostic performance achieved with conventional oximetric indexes commonly used by physicians. GAs proved to be an effective and robust tool to search for essential oximetric features that could enhance NPO in the context of OSA diagnosis.
The authors have extensively assessed the suitability of GAs for feature selection in the context of OSA diagnosis from NPO. To achieve this goal, their feature selection methodology took into account the number of features, in order to explore each k-dimensional subspace: GAs were applied to obtain the optimal feature subset for a given number of input features, from 2 to $p - 1$, where p is the dimension of the original feature space. An optimal subset was defined as the group of input variables to a LR model that achieved the highest classification performance.

Major Findings:

- An optimal feature subset was found by means of GAs from the proposed oximetric feature set composed of time statistics, frequency statistics, spectral and nonlinear features. M1t, M3t, M4t, MF, PA and CTM were automatically selected, which achieved 90.6% sensitivity, 81.3% specificity and 87.5% accuracy on an independent test set.
- Thus, GAs could be a useful tool to exhaustively investigate an oximetric feature space in order to find feature subsets with high diagnostic power.
- The authors conclude that complementary feature extraction plus exhaustive variable selection methodologies could obtain essential information from SpO2 recordings to improve NPO usefulness in the diagnostic assessment of OSA syndrome.

D.Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases(Juanying Xie , Chunxia Wang ,2011)

Medical dataset description:

The differential diagnosis of erythemato-squamous diseases is a difficult problem in dermatology. The erythemato-squamous diseases all share the clinical features of erythema and scaling, with very little differences. The diseases in this group are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris. Usually a biopsy is necessary for the diagnosis but unfortunately these diseases share many histopathological features as well. Another difficulty for the differential diagnosis is that a disease may show the features of another disease at the beginning stage and may have the characteristic features at the following stages. Patients were first evaluated clinically with 12 features. Afterwards, skin samples were taken for the evaluation of 22 histopathological features. The values of the histopathological features are determined by an analysis of the samples under a microscope. In the dataset constructed for this domain, the family history feature has the value 1 if any of these diseases has been observed in the family, and 0 otherwise. The age

feature simply represents the age of the patient. Every other feature (clinical and histopathological) was given a degree in the range of 0 to 3. 0 indicates that the feature was not present, 3 indicates the largest amount possible, and 1, 2 indicate the relative intermediate values. The datasets consisting 34 features are used in the study and the six classes of erythemato-squamous diseases are shown in Table 2.

Diagnosis Results:

The authors employed a hybrid feature selection method, named IFSFS. In the process of filter, the improved F-score for each feature is calculated and then they are sorted in descending order. A SVM-based diagnosis model with hybrid feature selection method, named IFSFS, for diagnosis of erythematosquamous diseases is used. This hybrid feature selection method combines the advantage of filter method and wrapper method, where the improved F-score is used as an evaluation criteria of filter method and SFS is used as the evaluate system of wrapper method for feature selection to find out the best feature subset, and SVM is used to evaluate the classification accuracy.

Major Findings:

- Experiments have been conducted on four different portions of the erythemato-squamous diseases dataset, and the results show that the proposed method obtained the high classification accuracies (96.65%, 95.83%, 94.44%, and 98.61% for 50–50%, 60–40%, 70–30%, and 80–20% of training-test partition respectively) for the subset that contained 21 features (model #21).
- Considering the results, the SVM-based model with IFSFS developed gives very promising results for diagnosis of erythemato-squamous diseases.

Table 2:
Erythemato-squamous diseases Data set present study:

Erythemato-squamous diseases (number of patients)	Features	
	Clinical	Histopathological
Psoriasis (111)	Feature 1 :Erythema	Feature 12:Melanin incontinence
Seboreic dermatitis (60)	Feature 2: Scaling	Feature 13: Eosrinophils in the infiltrate
Lichen planus (71)	Feature 14: PNL infiltrate	Feature 14: PNL infiltrate
Pityriasis rosea (48)	Feature 4: Itching	Feature 15: Fibrosis of the papillary dermis
Chronic dermatitis (48)	Feature 5: Koebner phenomenon	Feature 16: Exocytosis
Pityriasis rubra pilaris (20)	Feature 6 : Polygonal papules Feature 7: Follicular papules Feature8: Oral mucosal involvement Feature9: Knee and elbow involvement Feature 10: Scalp involvement Feature 11: Family history Feature 34: Age Feature 34: Age Feature 23: Pongiform pustule	Feature 17: Acanthosis Feature 18: Hyperkeratosis Feature 19: Parakeratosis Feature20: Clubbing of the rete ridges Feature 21: Elongation of the rete ridges Feature 22: Thinning of the suprapapillary epidermis Feature 23: Pongiform pustule Feature 24: Munro microabcess Feature 25: Focal hypergranulosis Feature 26: Disappearance of the granular layer Feature 27: Vacuolization and damage of basal layer Feature 28: Spongiosis Feature 29: Saw-tooth appearance of retes Feature 30: Follicular horn plug Feature 31: Perifollicular parakeratosis Feature 32: Inflammatory mononuclear infiltrate Feature 33: Band-like infiltrate

E.Feature selection for a cooperative coevolutionary classifier in liver fibrosis diagnosis (Catalin Stoean , RuxandraStoean , MonicaLupsor , HoriaStefanescu , RaduBadea,2011)

Medical dataset description:

This paper presents an automatic tool capable to learn from a patient's data set with 24 medical indicators characterizing each sample and to subsequently use the acquired knowledge to differentiate between five degrees of liver fibrosis. The indicators represent clinical observations and the liver stiffness provided by the new, non-invasive procedure of Fibro scan.

The chronic hepatitis C data set employed in this paper comes from the 3rd Medical Clinic, University of Medicine and Pharmacy, Cluj-Napoca, Romania, and consists of 722 samples, each described by 24 indicators, with a small number of missing values. The medical attributes that are chosen to provide information that triggers a certain degree of liver fibrosis are outlined in Table 3. The first one is the stiffness indicator from the Fibroscan, while the others represent standard hematological and biochemical exams that are required in a patient with chronic hepatitis C. The five possible degrees of fibrosis have the following meaning and number of representatives:

- F0 (nofibrosis)–29examples;
- F1 (portalfibrosiswithoutsepta)–227examples;
- F2(portalfibrosisandfewsepta)–64examples;
- F3(numerousseptawithoutcirrhosis)–87examples;
- F4 (cirrhosis)–215examples.

TABLE 3

THE DETAIL OF THE NINE ATTRIBUTES OF BREAST CANCER DATA.

Attribute Name	Attribute Number
Stiffness (A1)	A1
Sex (A2)	A2
BMI (body mass index)(A3)	A3
Glycemia (A4)	A4
Triglycerides (A5)	A5
Cholesterol (A6)	A6
HDL Cholesterol(A7)	A7
amino transferase (A8)	A8
Alanin aminotransferase(A9)	A9
Gama glutamyltranspeptidase(A10)	A10
bilirubin(A11)	A11
Alkaline phosphatase(A12)	A12
Prothrombin index(A13)	A13
TQS (QuiqTime)(A14)	A14
INR (prothrombintimeratio)(A15)	A15
Prolonged activatedpartial thromboplastin time(A16)	A16
Haematids (erythrocytes)(A17)	A17
Hemoglobin (A18)	A18
Hematocrit (A19)	A19
Medium erytrocityvolume(A20)	A20
Avg. erytrocitaryhemoglobin(A21)	A21
Avg. concentrationofhemoglobin in aredbloodcell(A22)	A22
Thrombocytes (A23)	A23
Sideraemia (A24)	A24

Diagnosis Results:

The proposed technique combines a hill climbing algorithm that selects subsets of important attributes for an accurate classification and a core represented by a cooperative co-evolutionary classifier that builds rules for establishing the diagnosis for every new patient. The results of the novel method proved to be superior as compared to the ones obtained by other important classification techniques from the literature. Additionally, the proposed methodology extracts a set of the most meaningful attributes from the available ones. One of the commonly used feature extraction mechanisms, principal component analysis (PCA), is thus also employed in the study. This leads to a sizeable reduction of the data dimension- ality from 24 to only 6 assembled attributes. When applying CCEA to the new data, the results are improved by only 1.4%, as opposed to those achieved through the direct application on the original data. It has to be mentioned that there is, however, a major improvement in runtime.

Major Findings:

- Depending on the selected attributes, the best accuracy result obtained as the average over 30 repeated runs of random cross-validation of the CCEA is of 62.11% correctly classified patients.
- Over the 1000 fitness evaluations of the hill climber, the average accuracy is of 55.93%, while the worst test accuracy is of 47.92%.
- The individual that yields the best obtained accuracy only selected nine attributes out of the 24 available and these are the following: stiffness, triglycerides, HDL cholesterol, aspartate amino transferase, gamma-glutamyl transpeptidase, alkaline phosphatase,

prothrombin index, prolonged activated partial thromboplastin time and hematocrit.

III. COMMANILITIES AND DIFFERENCES

In this article, we reviewed the main contributions of data mining research in a set of well-known medical informatics applications. The main issue that emerge as common problems in the medical diagnosis domain is the criticality of the predicted result – more accuracy is needed. To deal with this problem, a wealth of Feature Selection and classification techniques has been designed by researchers in medical informatics, machine learning and data mining. In all these works, the prediction accuracy is taken as the prime evaluation metric and the authors have succeeded in producing superior results. Also among the numerous attributes, the most essential subset of attributes that will enhance the classifier performance has been identified by the authors. As a result of this step, the computational complexity is also reduced to a significant amount.

IV. CONCLUSION

A large and fruitful effort has been performed during the last years in the adaptation and proposal of data mining techniques for more accurate medical diagnosis. The proposal of univariate and multivariate selection algorithms can be considered as one of the most promising future lines of work for this community.

A second line of future research is the development of especially fitted ensemble FS approaches to enhance the robustness of the finally selected feature subsets. We feel that, in order to alleviate the actual small attribute/sample sizes of the majority of medical applications, the further development of such techniques, combined with appropriate evaluation criteria, constitutes an interesting direction for future research.

Other interesting opportunities for future data mining research will be the extension towards upcoming medical domains and the combination of heterogeneous data sources.

REFERENCES

[1] Yvan Saeys, Inaki Inza and Pedro Larran aga "A review of feature selection techniques in bioinformatics", BIOINFORMATICS Vol. 23 no. 19 2007.

[[2] P. Cunningham, J.G. Carney, "Diversity versus quality in classification ensembles based on feature selection", 11th European Conference on Machine Learning (2000), in: Lecture Notes in Artificial Intelligence, R. Lopez de Mantaras and E. Plaza, (Eds.), Springer Verlag, pp. 109-116,2000.

[3] S. Yu, S. De Backer, P. Scheunders, "Genetic feature selection combined with composite fuzzy nearest neighbor classifiers for high-dimensional remote sensing data", IEEE International Conference on Systems, Man, and Cybernetics ,pg 1912-1916,2007.

[4] A.J. Hoffman, C. Hoogenboezem, N.T. van der Merwe, C.J.A. Tollig, "Seismic buffer recognition using mutual information for selecting wavelet based features", IEEE International Symposium on Industrial Electronics,pg 663-667,1998.

[5] D. Garrett, D.A. Peterson, C.W. Anderson, M.H. Thaut, "Comparison of linear, nonlinear, and feature selection methods for EEG signal classification", IEEE Transactions on Neural Systems and Rehabilitation Engineering,pg141–144,2003.

[6] M.A. Kupinski, M.L. Giger, "Feature selection and classifiers for the computerized detection of mass lesions in digital mammography", International Conference on Neural Networks,pg 2460-2463,1997.

[7] M.F. McNitt-Gray, H.K. Huang, J.W. Sayre, "Feature selection in the pattern classification problem of digital chest radiograph segmentation", IEEE Transactions on Medical Imaging, pg 537–547,1997.

[8] P. Zarjam, M. Mesbah, B. Boashash, "An optimal feature set for seizure detection systems for newborn EEG signals", International Symposium on Circuits and Systems, pg 33-36,2003.

[9] A. Blum, P. Langley, "Selection of relevant features and examples in machine learning", Artificial Intelligence, pg 245-271, 1997.

[10] Bo Jin, Y.C. Tang, Yan-Qing Zhang," Support vector machines with genetic fuzzy feature transformation for biomedical data classification", Information Sciences,pg 476–489,2007.

[11] Xiangyang Wanga, Jie Yanga, Richard Jensenb, Xiaojun Liua," Rough set feature selection and rule induction for prediction of malignancy degree in brain glioma" computer methods and programs in biomedicine ,pg 147–156,2006.

[12] Vapnik, V,"The nature of statistical learning theory." New York: Springer,1995.

[13] Cristianini, N., & Shawe-Taylor, J," An introduction to support vector machines:And other kernel-based learning methods",Cambridge, UK: Cambridge University Press",2000.

[14] Luukka, P.,"Similarity classifier in diagnosis of bladder cancer. Computer Methods and Programs in Biomedicine", pg 43–49,2008.

[15] Tang KS, Man KF, Kwong S, He Q," Genetic algorithms and their applications", IEEE Signal Process Mag, pg 22–37,1996.

[16] Hui-Ling Chen , Bo Yang , Jie Liu , Da-You Liu," A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis", Expert Systems with Applications, pg 9014–9022,2011.

[17] T. Warren Liao,"Diagnosis of bladder cancers with small sample size via feature selection",Expert Systems with Applications ,pg 4649–4654,2011.

[18] Daniel Alvarez, Roberto Hornero, J. Victor Marcos, Felix del Campo," Feature selection from nocturnal oximetry using genetic algorithms to assist in obstructive sleep apnoea diagnosis", Medical Engineering & Physics 2011.

[19] Juanying Xie , Chunxia Wang," Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases", Expert Systems with Applications,pg 5809–5815,2011.

[20] Catalin Stoean, RuxandraStoean , MonicaLupsor , HoriaStefanescu , RaduBadea," Feature selection for a cooperative coevolutionary classifier in liver fibrosis diagnosis", Computers in Biology and Medicine ,pg 238–246,2011.