34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Categorization of COPD Patient's Health Level through the Use of the CHRONIOUS Wearable Platform

Christos Bellos, Athanassios Papadopoulos, Roberto Rosso, and Dimitrios I. Fotiadis, *Senior Member, IEEE*

Abstract— **The Chronic Obstructive Pulmonary Disease (COPD) is a chronic disease that causes airflow blockage and breathing-related problems. As a Chronic disease it requires specific treatment plan and patient management for a long period of time. Critical factor in the process is the realization of frequent and precise diagnostic tests that describes the health status of the patient. The CHRONIOUS system provides the required easy-to-use wearable platform aiming at the successful management of COPD patients. Several signals and patient's data are stored by the utilization of an ergonomic jacket and through the patient's platform interface. Hybrid techniques based on supervised and unsupervised methodologies were applied for the analysis of the patient's situation. The categorization of health level of the patient to discrete levels is achieved in a continuous base. Useful outcomes in the form of message or advice are extracted appeared on patient's and clinician's devices denoting his health status.**

Index Terms— **Chronic Obstructive Pulmonary Disease, Classification, Random Forests, Support Vector Machine.**

I. INTRODUCTION

CHRONIC Obstructive Pulmonary Disease (COPD) is a chronic disease that appeared as progressive obstruction of the airflow into and out of the lungs and increased shortness of breath. As it happens with most of chronic diseases, they are difficult to be cured and they need much consideration on the planning and the further management of their treatment. An effective supervision of COPD [1] might change its course and progress. The characteristics and symptoms of the disease vary depending on several factors. Some of them are the patient's age, the disease years and the stage of the disease. The contribution of the treatment plans are the recovering of the patient's health status and the improving the quality of their every-day life's. Thus, concerning the clinical needs of the disease, no matter which is exactly the progress, COPD patients need continuous monitoring based on accurate measurements of critical biosignals or medical parameters, which are sufficient in describing each time the health status of the patient.

Due to the technology's great contribution in the management of the chronic disease and the clinicians requests for a more accurate and effective patient's monitoring process, several research projects have been appeared focusing on-site or telemedicine services [2]. The contribution of the recent research and technological innovations is appeared beneficial since much of the management's task is performed utilizing state of the art technologies for the development of a smart wearable system. Body area networks and several types of sensors for the recording of vital signs were utilized [3]. The data that have been collected were preprocessed and analyzed by conventional or more intelligent processes.

The CHRONIOUS system [4] is a generic health status monitoring platform that addresses people suffering from chronic health conditions, which is achieved by the development of an intelligent, ubiquitous and adaptive chronic disease platform which provides services to clinicians and their patients. This platform has been developed integrating several state of the art sensors which satisfies both patients and clinicians' requirements [5]. Beyond platform's hardware, efficient algorithmic processes are responsible for the analysis and the classification of the status of the patient providing information to the patient and the clinician contributing in the modification of the predesigned treatment. In this work, the intelligent part of the CHRONIOUS system is presented. Technical details about the data collection, preprocessing, analysis and their classification is shown. The evaluation of the different classification techniques as well as their effectiveness in patient's categorization is validated. In such a way, the intelligent system of the CHRONIOUS platform provides an innovative and highly effective categorization tool for the management of the health status of the COPD patients.

II. MATERIALS AND METHODS

A. Data Collection

Data from 21 patients suffering from Chronic Obstructive Pulmonary Disease [6] (COPD) have been recorded. The data collection took place in a pilot hospital (Careggi Hospital, Italy) and the wearable platform was facilitated for one day of recording. The inclusion criteria for patient enrollment in the data collection phase are indicated by

Manuscript received March 15, 2012. This work is part funded by the European Commission through the IST Project FP7-ICT-2007–1– 216461 CHRONIOUS.

C. Bellos, A. Papadopoulos, D.I. Fotiadis are with the Foundation for Research and Technology - Hellas, Institute of Molecular Biology and Biotechnology GR 45110, Ioannina Greece (e-mail: cbellos@cc.uoi.gr, thpapado@cc.uoi.gr, fotiadis@cs.uoi.gr). R. Rosso is with TESAN Telematic & Biomedical Services S.p.A., Vicenza, Italy (e-mail: rosso@tesan.it).

clinicians: 45≥age≤80 years, 18≥Body Mass Index (BMI)≤35, 120cm≤Cinrcunpherence thorax ≤178 cm, no deformity of the chest wall, ex or current smokers, stage III-IV in the GOLD categorization [7]. Clinicians also indicated the exclusion criteria of the data collection: psychiatric disorders, other severe comorbidities with compromised life expectancy in the study period, illiteracy, refusal in participation.

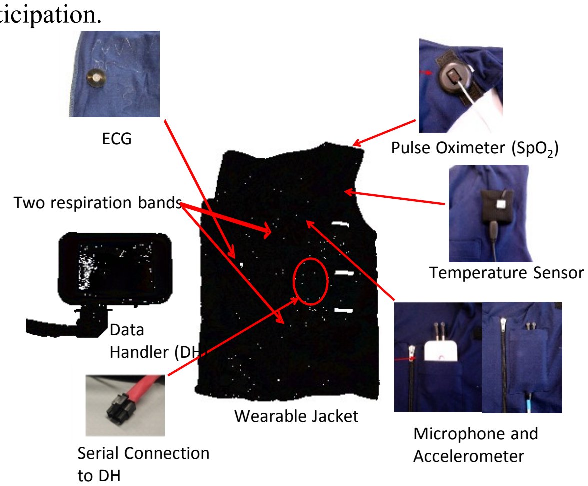

Fig. 1. The CHRONIOUS wearable system and the positions of the attached sensors to the jacket.

The dataset consists of information acquired by three main sources: 1) Sensors: signals acquired by sensors that are attached to the wearable platform (Fig. 1). The wearable platform contains also the data handler (DH) device, which collects all the signals coming from the body sensor network and transmits them to a portable Smart Device (SD), 2) Database: static information (e.g. demographics data) and 3) external devices (i.e. spyrometer, body weight, systolic and diastolic blood pressure and blood glucose). Clinicians provided also an extra feature that specifies the abnormality assessment of the situation, constituting the class index for the specific instance. There are annotated four different levels of severity (no severity, mild, moderate and severe).

B. The Method

The sensor's acquired data are transferred to the SD via Bluetooth. Then signal denoising methods are applied and the processed signals enter the Feature Extraction component where useful features are extracted. Then, the heterogeneous data that are acquired from various sources (i.e. Wearable sensors, Database, External Devices) are fused and a multi-dimensional vector is formed (NxM matrix, where M is the number of attributes plus the class of each instance and N is the number of the instances). Then redundant information is removed by utilizing sensitivity analysis and finally the Decision Support System (DSS) is triggered in order to assess the patient's current health status.

Preprocessing and Signal de-noising

ECG and respiration signals need to be pre-processed and de-noised in order to properly extract features [8]. The preprocessing of ECG is performed in three steps: initially, the base line wander noise is removed to eliminate the very low frequencies, then high frequency noise is removed and finally QRS detection is applied [9]. The respiration rate is calculated using the reference respiration signal, where a dominant frequency detection algorithm is applied based on short-time Fourier transform (STFT). In this study, the Hamming window is applied. A 60 seconds window size has been selected due to the fact that frequency components of the respiration signal are very low (< 2Hz).

Feature Extraction, Selection and Fusion

Features are extracted from various sensors' acquired signals and enter the Heterogeneous Data Fusion component which fuses all available data acquired from different sources and in various formats (i.e. XML data, binary, database stored information). Furthermore, two different feature selection (FS) algorithms [10] are applied to identify the most important clinical-pathological information for individualized health status identification. These FS algorithms are the GainRatioAttributeEval (FS1) and the Correlation-based Feature Subset Selection (FS2).

Decision Support System

The aim of the DSS is to classify the patient's current health status according to four different levels of severity (no severity, mild, moderate, severe). The constructed dataset is classified by two parallel classifiers, which form a system that combines an expert system and a supervised classifier.

The supervised classifier that is developed is a Support Vector Machine (SVM). The specific categorization difficulty that the SVM faces is characterized as multi-class due to the four levels of severity. The SVMs were mainly designed for binary classification [13] but several methods [14] have been proposed to effectively extend a developed one-against-all SVM to a multiclass classification problem. Several methods propose the construction of the classifier by combining several one against all classifiers while others propose to consider all classes at once. In this analysis, the one-against-all SVM classification approach has been implemented. We have utilized three kernels [15] of the developed SVMs; Polynomial, Sigmoid and radial basis function (RBF) while their respective parameters (C, degree and gamma) are optimized with the differential evolution [16] (DE) method.

The second classifier of the parallel system is an expert system, which is implemented using Random Forests [11]. After constructing the forest, several rules are extracted from the constructed pathways and trees and evaluated to facilitate the monitoring of the chronic disease and assess the severity of the patient's health status. This RF algorithm consists of many decision trees where a new subset of samples is selected from the dataset for the construction of each tree. Initially, a subset m of the features is picked up as the splitter of the tree node [12]. Then, the values of the current feature are utilized to sort the available data and finally the Gini index is calculated. The samples that are not participating in the construction of the tree are the out-of-bag (OOB) samples of the classification problem. OOB error is called the error of the RF classification problem that is calculated by using these samples and the average of the OOB errors of all trees. Finally, the majority voting procedure is used to select the predicted class.

Rule Extraction

The extracted tree from the implemented classical RF, described in the previous paragraph, is converted to rules. The classification accuracy is defined as the ratio of the number of corrected classified samples to the total number of samples and the trees with accuracy larger than the median accuracy of all trees are selected. The number of the training samples is used to define the weight of the leaf node, while the branches of the selected trees are selected and converted to rules. Finally, the rules are created by reading the pathway of the tree from top to the specific leaf node.

Although this study is preliminary using very limited number of patients, the created set of rules after an effective training can be used to assess the severity of a COPD patient's health status. Each node of the tree contains a selected feature, the leaves represent classifications (i.e. Level 1 corresponds to class 1 and Level 2 corresponds to class 2) and the branches represent conjunctions of features that lead to those classifications. For example one of the generated rules is the following:

If Respiration Amplitude>=0.00318 and Mean RS Distance<=0.043 and SpO_2>=87.2167 and Respiration Rate>=26.1364 then the instance is classified as class 2 (Level 2).

III. Results

The system utilizes a Random Forest (RF) classifier in order to enhance the accuracy and extract useful rules that can easily be interpreted. The results are presented analytically in Table I for three different numbers of trees, which provide the best accuracy, as depicted in Fig. 2.

TABLE I
RESULTS OF THE RANDOM FOREST APPLIED CLASSIFIER UTILIZING VARIOUS PARAMETERS

	Number of Trees: 15		Number of Trees: 18		Number of Trees: 21	
	accTe	oobAc	accTe	oobAc	accTe	oobAc
Gini - no FS	68.3%	61.5%	67.3%	60.5%	65.3%	61.8%
Gini – FS1	86.3%	74.6%	84.7%	77.4%	**86.3%**	**78.6%**
Gini – FS2	86.3%	74.4%	82.3%	76.7%	84.3%	76.9%
ReliefF - no FS	69%	56%	70.7%	54.7%	71%	58.1
ReliefF– FS1	80.7%	76.5%	84.7%	77.8%	**84.7%**	**81.4%**
ReliefF - FS2	72.7%	75%	82.3%	76.7%	76.3%	74.8%

Various parameters (i.e. number of trees, evaluation measure, feature selection algorithm applied) have been utilized in order to compare the results and optimize the performance of the Random Forest. Also, there have been used two evaluation measures, the Gini index and the ReliefF estimator. The columns of the Table I with label "accTe" represent the accuracy in the Testing dataset, while the columns with label "oobAc" represent the accuracy in the out-of-bags samples.

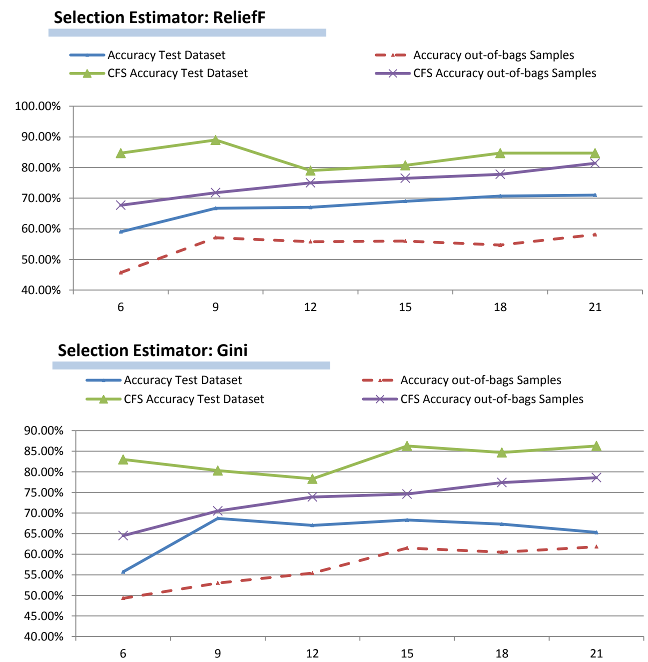

Fig. 1. Chart for the comparison of the accuracy deviation by utilizing different number of trees

As it is described in the previous section, a one-against all SVM classifier is also developed in parallel to the RF. There were developed four different SVMs, each one for each level of assessment and the final result was chosen after a majority voting procedure.

TABLE II
RESULTS OF THE IMPLEMENTED SVMS

Kernel	Feature Selection	1st Level	2nd Level	3rd Level	4th Level
RBF	No FS	88%	88%	91%	74%
	CFS	99%	87%	86%	82%
	Gain	98%	87%	91%	82%
Sigmoid	No FS	88%	89%	87%	54%
	CFS	84%	91%	78%	65%
	Gain	82%	91%	83%	64%
Polynomial	No FS	69%	80%	57%	61%
	CFS	78%	84%	69%	80%
	Gain	68%	79%	46%	75%

The parameters of each applied kernel are optimized and the results are presented in Table II (utilizing stratified folds cross-validation for splitting the dataset to training and testing).

TABLE III
SAMPLE SET OF RULES EXTRACTED FROM RANDOM FORESTS

Rules	Decision	Clinical Knowledge
IF Heart Rate<70.57 AND SpO2>93.69 AND Respiration Rate<24.3 THEN	Level 1	Soft clinical indicator
IF Exhalation Duration<1.2 AND Respiration Rate<40 AND Heart Rate>70.7 AND LF>14.5 AND SpO2>87.8 THEN	Level 2	Moderate clinical indicator
IF Humidity<34 AND Respiration Rate>26 AND SpO2<87.8 THEN	Level 3	Hard clinical indicator
IF SpO2<90.35 AND Humidity<33 AND Ratio LF/HF>0.38 AND Respiration Rate>26 AND Heart Rate>68.8 AND Exhalation Duration<1.18 THEN	Level 3	Hard clinical indicator
IF SpO2<85 AND Respiration Rate<26 AND Heart Rate>70.5 THEN	Level 4	Severe clinical indicator

Finally, a sample set of rules extracted from Random Forests is shown in Table III. It is important to underline that they provide a combination of the medical knowledge concerning the disease which allows to physicians to diagnose the disease and its severity with high specificity.

IV. DISCUSSION

The CHRONIOUS Intelligent System provides an estimation of the severity of the COPD patient's condition. The aim of the developed classifiers is to limit the decision error increasing, in the same time, the system's accuracy.

There are some positive results obtained from the implementation of the Random Forest. Utilizing the Gini index as the evaluation measure and the FS2 (i.e. the Correlation-based Feature Subset Selection algorithm) the accuracy of the system is improved compared with the initial approach. In contrast with the SVM results we could point out that the Random Forest classifier is less accurate than the SVM. Almost all the implemented SVMs for the different levels give lower classification error, except the Level 1 where overtraining issues were met. Moreover, as indicated in Table I, we concluded to the implementation of a Random Forest with 9 trees since it provides the best ratio accuracy to performance time.

Furthermore, the results obtained from the SVM classification system are encouraging. According to the accuracy of the implemented SVMs classifiers, displayed in Table I, we conclude that the performance of the RBF kernel provides the highest classification result. On the other hand, the accuracy of 99% and 98% that arises in the implemented Level 1 SVM with RBF kernel and CFS and Gain Feature Selection algorithms applied respectively are not consistent due to various overtraining problems since the number of samples in this Level is quite poor to obtain a successful training of the respective SVM system.

Finally, it should be mentioned that the clinical knowledge, pointed out in the third column of Table III, was provided by the clinical partners of the CHRONIOUS project. Based on this medical knowledge, the rules that are generated from implemented Random Forests are fully justified.

V. CONCLUSIONS

The Intelligent System offers an innovative and a completely portable system aiming at the effective management of the health status of the patients suffering from COPD both indoors and outdoors.

As for the future work, the employment of larger datasets will improve the performance of the CHRONIOUS system and enhance the validation procedure. There is planned a second data collection phase where more COPD patients will be recruited. Moreover, another clinical protocol is designed aiming to recruit Chronic Kidney Disease (CKD) patients in order to enhance the interoperability of the Intelligent System and its adaptation to other chronic diseases requirements.

REFERENCES

[1] K. Atsou, C. Chouaid, G Hejblum, Variability of chronic obstructive pulmonary disease key epidemiological data in Europe: systematic review. BMC Med. 2011.

[2] S. McLean, U. Nurmatov, J. Liu, C. Pagliari, J. Car, A. Sheikh: Telehealthcare for chronic obstructive pulmonary disease. Cochrane Database Syst Rev. 2011.

[3] M. Chen, S. Gonzalez, A. Vasilakos, H. Cao, V. Leung: Body Area Networks: A Survey, Mobile Netw Appl. 2011.

[4] A. Papadopoulos, D.I. Fotiadis, M. Lawo, F. Ciancitto, C. Podolak, R. L. Dellaca, G. Munaro, R. Rosso: CHRONIOUS: A Wearable System for the Management of Chronic Disease. 9th International Conference on Information Technology & Applications in Biomedicine, Larnaca, Cyprus, 2009.

[5] C. Bellos, A. Papadopoulos, D.I. Fotiadis and R. Rosso: An Intelligent System for Classification of Patients Suffering from Chronic Diseases. 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Buenos Aires, Argentina, August 31 - September 4, 2010.

[6] G. Viegi, F. Pistelli , D.L. Sherrill, S. Maio, S. Baldacci and L. Carrozzi: Definition, epidemiology and natural history of COPD. Eur Respir J; 30: pp. 993–1013, 2007.

[7] GOLD 2008 Guidelines (Global Initiative for Chronic Obstructive Lung Disease).

[8] C. Bellos, A. Papadopoulos, R. Rosso και D.I. Fotiadis: Extraction and Analysis of features acquired by wearable sensors network. 10th International Conference on Information Technology and Applications in Biomedicine, Corfu, Greece, 2-5 November 2010.

[9] I. Fatimah, N.A. Osman, J. Usman and N.A. Kadri: Performance Evaluation of Coifman Wavelet for ECG Signal Denoising. Proceedings of IFMBE (pp. 419-422). IFMBE, 2007.

[10] D. Taniar: Data Mining and Knowledge Discovery Technologies. Part of the IGI Global series named Advances in Data Warehousing and Mining (ADWM) vol. 2, 2007.

[11] L. Brieman: Random forests. Machine Learning Journal, 45, pp. 5-32, 2001.

[12] E.E. Tripoliti, D.I. Fotiadis, M. Argyropoulou, G. Manis: A six stage approach for the diagnosis of Alzheimer's disease based on fMRI data. Journal of Biomedical Informatics, vol. 43, pp.307-320, 2010.

[13] L. Rokach, O. Maimon: Data Mining with Decision Trees, Series in machine perception and artificial intelligence. Vol.69, pp.71-71, 2008.

[14] C. Hsu and C. Lin.: A Comparison of Methods for Multiclass Support Vector Machines. IEEE Transactions on Neural Networks, 2002.

[15] M. Kumar and M. Gopal: Reduced one-against-all method for multiclass SVM classification. Expert Systems with Applications, 2011.

[16] N.E. Ayat, M. Cheriet and C.Y. Suen: Automatic model selection for the optimization of SVM kernels. Pattern Recognition 38 1733 – 1745, 2005.