Biodata analytics for COPD

P. Carvalho, J. Henriques, C. A. Teixeira, R, Couceiro, T. Rocha, L. Mendes, B. Rocha, D. Nunes, I. Chouvarda, N. Maglaveras, R. P. Paiva

Abstract— **WELCOME is an ambitious EU FP7 project which aims to bring about a change in the reactive and integrated care nature of the management of chronic diseases and in particular the Chronic Obstructive Pulmonary Disease and its comorbidities. In order to achieve these goals, the Welcome solution will incorporate several bio-sensors to enable bio-analytics and multi-parametric modeling. In this paper we introduce the biodata analytics module of the Welcome system. This module is composed of several analysis sub-modules dedicated to the analysis of specific biosignals. Inhere we will introduce the modules for high resolution ECG and for chest sounds processing.**

I. INTRODUCTION

COPD (Chronic Obstructive Pulmonary Disease) is a common preventable and treatable disease, characterized by persistent airflow limitation that is usually progressive and associated with an enhanced chronic inflammatory response in the airways and the lung to noxious particles or gases. Exacerbations and comorbidities increase to different scales the overall severity in individual patients.

COPD poses a significant public health burden with a morbidity and mortality rate that is increasing. One WHO report anticipates that by 2030, COPD will become the fourth cause of mortality and seventh cause of morbidity worldwide. It is often associated with several co-morbidities such as cardiovascular disease (such as CHF), metabolic syndrome (including diabetes), osteoporosis, mental health diseases (depression and anxiety) and lung cancer, benefitting from an integrated or co-ordinated care approach.

WELCOME is an ambitious EU FP7 project which aims to bring about a change in the reactive and integrated care nature of the management of COPD and its comorbidities. In order to achieve these goals, the Welcome solution will incorporate several bio-sensors to enable bio-analytics and multi-parametric modelling. More specifically, i) a light, easy and comfortable-to-wear-and-maintain vest including a large number of standalone non-invasive chest sensors (26 sensors) working in concert for measuring and monitoring various parameters of COPD and co-morbidities and ii) off-the-shelf devices with wireless connection capabilities dedicated to the treatment of diabetes co-morbidity (i.e, bio-chemical sensors for blood glucose, lipid profiling, and BNP, potassium and creatinine assessment). The vest will enable the measurement of the following parameters: SpO2, EIT (electrical impedance tomography), chest sounds and high spatial resolution ECG. In this paper we will discuss mainly the Welcome algorithms for chest sound and multi-channel ECG analysis, since the algorithms for EIT are still in an early development stage and the remaining sensors are standard sensors that already provide the required information for the decision support system. In the Welcome approach, chest sounds are applied for the detection and analysis of crackles, wheezes and cough events. Regarding ECG, the analysis of interest is conformed to ECG segmentation and intervals computation, atrial and ventricular arrhythmias episodes detection, atrioventricular block diagnosis and ST deviation assessment.

II. MATERIAL AND METHODS

A. *The ECG analysis module*

Regarding the ECG processing module, the ECG first undergoes a quality control check where contamination by noise and artefacts are identified. The clean ECG segments are then further processed in order to extract the aforementioned diagnosis information.

1) *Pre-processing*

As result, in the ECG analysis performed: i) the baseline wander and noise removal was implemented using a wavelet approach [1]; ii) for the segmentation the method proposed by Sun [2] with some adaptations was implemented. In effect, using morphological analysis, the most important fiducial points have been determined, enabling to characterize QRS complex, P and T waves, as well as the relevant intervals based on those waves identification. In particular the following parameters were computed:

Segmentation: P wave: P onset, P peak and P offset indexes; QRS complex: Q onset, Q peak, R peak, S peak and S offset indexes; T wave: T onset, T peak and T offset indexes.
Intervals: RR, heart rate (bpm), PR interval (s), corrected QT interval (s), Q wave width (s), Q peak height, R peak height, QRS complex duration (s) and corrected JT interval.

2) *AF: Atrial fibrillation*

From the clinical perspective the key characteristics of an AF episode is the absence of P waves before the QRS-T complex, which presents a *sawtooth like* pattern along the cardiac cycle, and the irregularity of the RR intervals. Thus, the proposed strategy makes use of the three principal physiological characteristics of AF, applied by cardiologists in their daily reasoning: *i*) P wave absence/presence, *ii*) heart

This work was supported by EU Project WELCOME (FP7 - 611223).

L. Mendes, P. Carvalho, C. Teixeira, R. P. Paiva, J. Henriques are with the Centre for Informatics and Systems of the University of Coimbra (CISUC), Pólo II, Coimbra, Portugal (e-mail: {lgmendes, carvalho, cteixei, ruipedro, jh} @dei.uc.pt).
N. Maglaveras and I. Chouvarda are with the Lab of Medical Informatics, Aristotle University of Thessaloniki, Greece (e-mail: nicmag@med.auth.gr, ioannach@auth.gr).

978-1-5090-2455-1/16/$31.00 ©2016 IEEE

rate irregularity and *iii*) atrial activity analysis. This knowledge-based approach has the advantage of increasing interpretability of the results to the medical community, while improving detection robustness.

A total of 8 features, $f_i, i=1,..,8$, have been computed to address these characteristics. The features f_1, f_2 and f_4 are time-domain features, whereas f_5, f_7, f_8 are frequency-domain features and f_3 and f_6 features are computed using non-linear measures.

P wave detection: (f_1)·the P wave absence/presence was quantified by measuring the linear correlation of each P wave to a normalized P wave model, created using common P waves extracted from the Physionet QT Database.

Heart rate variability: (f_2, f_3, f_4)·in a first phase the RR interval sequence was modelled as a three-state Markov process, being each interval classified as one of the three states (short, regular or long), and characterized by its state transition probability matrix [3]. Three features were computed based on the probability, on the entropy of the distribution and on the Kullback-Leibler divergence measure between current window and model distribution.

Atrial activity analysis: (f_5, f_6, f_7, f_8)·in a first step the method proposed by [4], was followed to separate the atrial activity components in the ECG signals. Then the spectrum content of the obtained atrial activity was assessed by computing the distance of the spectrum peak to the AF characteristic range, the entropy of the spectrum, the Kullback–Leibler divergence between each window spectrum and generalized bell-shaped membership function, and the ratio between the spectrum area in the AF range and the total area. .

3) *VT: Ventricular tachycardias*

The approach assumes that the fundamental differences in the physiologic origins of normal rhythm and VT, can be discriminated via time ECG shape together with power spectral density analysis. Moreover, the most significant features were selected from an original set of features (found in literature as well as developed during this work) [5][6][7] total of 11 features, $f_i, i=1,..,11$, have been computed to discriminate between normal signals and VT episodes.

Time domain features: ($f_1,...,f_6$)·the first feature estimates the morphology of the signal, computing the amount of time that each beat peaks is above or below a given threshold. The second feature basically assesses the heart rhythm. The next four features assess small and high derivatives in the ECG signal, enabling to detect abnormal signal amplitudes and slopes.

Spectral information: (f_7, f_8, f_9)· the energy contained in different frequencies (three ranges) was used as an approach for characterizing the ECG signal. The PSD was evaluated by windowing segments of signal, computed using the Welch's method.

Non-linear features: (f_{10}, f_{11})· one feature employs a non-linear transform, in particular the multiplication of backward differences, providing an estimation of extreme variations in the ECG. The other estimates the spatial filling index, computed from the ECG phase space reconstruction diagram.

4) *ST deviation*

The ST segment deviation is a measure computed as the difference between the isoelectric point (after the P wave) and the amplitude of the *J* point (segment of the ECG that presents a stable behavior, between the end of the QRS complex and beginning of the T wave). This amplitude, designated as ST segment deviation, is decisive in the assessment of the ischemic condition.

The algorithms implemented to evaluate ST segment deviation follow basically two stages.

1. Baseline removal

Based on R peaks localization, the entire ECG signal is broken into cardiac cycles using the average of the distances between consecutive R peaks. Each cardiac cycle is then submitted to a process of baseline removal using Wolf's method. This method attempts to determine the initial and final heights (H1 and H2) of the interval, using the average of first five samples and the average of last five samples, respectively. Then, the line segment connecting H1 to H2 is subtracted from the ECG, originating a corrected signal in terms of baseline.

2. J point estimation

The method proposed here is based on a time-frequency analysis, showing notable results in what concerns the robustness of the J point identification. It is recognized that time-frequency methods are especially adequate for the detection of small transient characteristic hidden in the ECG, such as the ST segment. Thus, other approach for the estimation of ST deviation was based on a time-frequency analysis, in particular using the Wigner-Ville (WV) transform. The basic idea followed here consists in the division of the time frequency map into characteristic areas and, within each specific area, perform the evaluation of particular characteristics. With respect to ST estimation two time bands and one frequency band were considered. Regarding time band, the areas considered were those on the left (isoelectric line) and on the right (ST segment) of the R peak (assumed to be previously determined). For each time band it is expect to determine regions were there is no signal activity (isoelectric line, interval between the end of P wave and the begin of QRS complex, and ST segment, interval between the end of QRS complex and the begin of T wave). Thus, for those time bands, high frequency band was considered and, in particular, the region where high frequency components presents minimum values. **Error! Reference source not found.** depicts this idea, where an electrocardiogram and its corresponding high time-frequency components are shown. By evaluating the minimum of the sum of the high frequency components in each time band, isoelectric and J points can be obtained. Having determined these points ST deviation is straightforward estimated, as the difference between J and isoelectric values.

5) *AV Blocks*

An atrioventricular block (AV block) is a type of heart block in which the conduction between the atria and ventricles of the heart do not follow a correct path, i.e., when the atrial depolarizations fail to reach the ventricles or when atrial depolarization is conducted with a delay. All methods for AV block detection are supported by the identification of the main waves and intervals, namely the PR interval and QRS complex. To this aim, the segmentation module developed inside this project was used to compute these intervals. Simple clinical rules are applied to identify an AV block and to categorize it into the three clinical relevant categories.

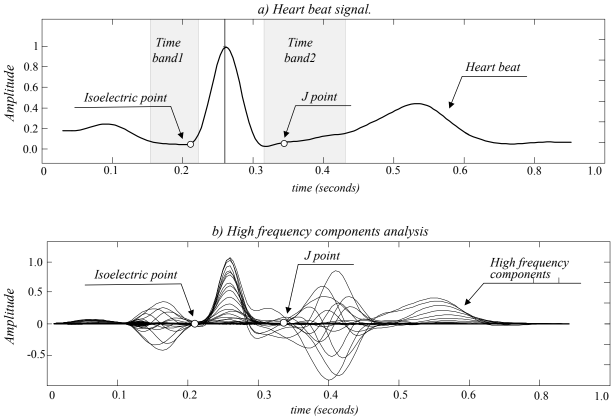

Fig. 1 ST segment deviation. *a*) electrocardiogram, isolelectric and J points *b*) frequency components (Wigner-Ville transform).

B. *Chest sound analysis module*

1) *Cough detection*

Chest sound analysis starts with a cough detector. First, we perform a pre-processing step, were we apply an 8th-order high-pass filter at 80Hz, followed by normalization, after which near-silent segments are discarded (through a process similar to the sound signal quality assessment). For each non-silent segment, seven features are extracted: (1) Mean Inharmonicity, the mean of the pitch inharmonicity, (2,3) Mean and Max Flux, the mean and the maximum of the spectral flux, (4) Max RMS, the maximum of the RMS values, and (5,6,7) Pitch Features. Regarding the classification a two-step classification for cough detection was implemented: (i) the first step is dedicated to speech discrimination. For this reason, we first train a model with the ground-truth annotations for speech and cough, leaving the artifacts out, and we perform a binary classification through multinomial logistic regression. The goal of the second step is to discard other artifacts and only keep the cough segments. In this step we work with all the ground-truth annotations and we don't use features (5) and (6), as they do not add any significant predictive value.

2) *Crackles and wheezes detection*

Two independent classification models, one for wheezes and another for crackles, were developed. Both follow the workflow presented in fig. 2**Error! Reference source not found.**. For each sound channel, we begin by doing the feature extraction and perform a binary classification frame by frame. After, we improve the classification results. To this end, we begin by doing a reduction of the false positives. This reduction is obtained by ensuring that the event is detected at least in two channels. Frames that were classified as noise or cough in the previous processing stages are also discarded. After that we perform the concatenation of the neighborhoods. Finally we count the number of events. We tested the performance of several features to detect wheezes and crackles events. For the detection of wheezes, we tested the performance of 30 features and for crackles we evaluated the performance of 33 features (see Table 1).

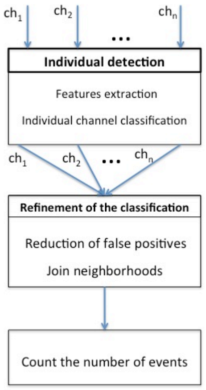
ch$_1$
ch$_2$
...
ch$_n$
Individual detection
Features extraction
Individual channel classification
ch$_1$
ch$_2$
...
ch$_n$
Refinement of the classification
Reduction of false positives
Join neighborhoods
Count the number of events

Fig. 2 Workflow of the crackles/wheezes events detector algorithm.

Table 1: Features tested to detect crackles and wheezes events.

Features	Crackles	Wheezes
Teager energy	x	
Fractal dim. of the WPST–NST	x	
Entropy	x	
WS-SS	x	x
29 Musical features	x	x

The performance of the features to discriminate respiratory sounds with wheezes/crackles events were studied taking into account the Matthews correlation coefficient (MCC) measured after classifying the data using the logistic regression classifier. The MCC is a balanced performance measure, especially suitable when the dataset are unbalanced.

To rank the importance of the features we use the sequential feature selection in the forward direction. Each frame was classified as containing or not containing wheezes/crackles events. For the wheeze features selection the additional data with voice and cough was used. For each classification a stratified 10-fold cross-validation approach with ten Monte Carlo repetitions was used.

The twenty most relevant features for the detection of wheezes and crackles events were used as inputs of the classification models. As with cough, due to the small size of the dataset (that exhibits a great variability of the adventitious sounds between patients) a leave-one-out (volunteer) cross-validation approach was used to test the performance of the detectors. The pre-detection and elimination of artifacts was not done. Since the acquisitions were unsynchronized it was not possible to reduce the false positives based on the simultaneity of the detected events.

III. RESULTS

A. *Data sets*

To assess the performance of the classifiers a subset of the

Physionet MIT-BIH databases was used, namely AF**:** atrial fibrillation database (AF) as well as AF and non-AF episodes collected from 12 patients. Regarding the latter, 1 episode (2 records of 30 mins.) was selected from the “St.-Petersburg Institute of Cardiological Technics 12-lead Arrhythmia Database” and 11 episodes (11 records of 60 mins.) were selected from the 12-lead ECG database collected by our team under the project Cardiorisk; VT: malignant arrhythmia database (VA) and Creighton University ventricular tachyarrhythmia database (CV). For the assessment of the ST deviation the European ST dataset was used.

For the validation of the chest sound analysis module an acquisition protocol was implemented in collaboration with the General Hospital of Thessaloniki ‘G. Papanikolaou’ and at the General Hospital of Imathia (Health Unit of Naoussa), Greece. The protocol includes the collection of chest sounds on 30 volunteers at six recording sites. For each site, lung sounds as well as cough and speech were acquired. The ethical committee of the General Hospital of Thessaloniki ‘G. Papanikolaou’ authorized the data acquisition. The physicians who supervised the acquisition annotated the different events in the timeline. Cough and speech periods are the predominant events in the cough sub-dataset. In total, 343 cough events were annotated. 113 wheezes and 199 crackle events were also annotated.

Table 2: Classification performance (SL/ML-single/multi lead)

	AF (SL)	AF (ML)	VT (VA)	VT (CV)
Sensitivity	79.0	88.5	90.7	91.8
Specificity	91.4	92.9	95.0	96.9

B. Validation: ECG analysis module

1) ECG delineation

The ECG segmentation algorithm validation has been performed using all 105 records from MIT-QT Database. Record lead configurations most similar to MLII have been chosen for testing the algorithm.

2) AF

To validate the proposed AF detection algorithm, 13 records from Physionet and Cardiorisk were used. Respectively 2160 and 2160 windows of 10 seconds, corresponding to AF and non AF episodes, compose the training/validation dataset. The validation was performed using a 6-fold cross validation approach (repeated 20 times) and the average sensitivity and specificity was computed.

3) Ventricular Arrhythmias

A data base of 51 signals was created, involving the three ECG signal classes (normal sinus rhythm, VT and VF). For MVA and CVT data sets, the number of windows was 420 (35 minutes) and 102 (8.5 minutes), respectively.

4) Segmentation and arrhythmias results

The results obtained for the ECG segmentation and the arrhythmias detection by the proposed algorithm are presented in Table 2.

5) ST deviation

A truly validation process could not be done. In fact, the available databases in this area were created to be used for evaluation of algorithms that detect or differentiate between ischemic ST episodes, axis-related non-ischemic ST episodes, etc. This is not the case of the present algorithm, which only considers discrete values of the ST segment deviation without further processing. For this reason, a correlation analysis was carried out. The average results obtained were 0.576 for the proposed method, 0.512 for Taddei’s method and 0.575 for Pang’s algorithm.

C. Validation: Chest sound analysis module

Regarding the chest sound analysis module, table 3 summarizes the achieved results.

Table 3: Classification performance

	Noise	Cough	Wheezes	Crackles
Sensitivity	92.2	93.2	79.0	84.0
Specificity	91.0	87.6		
PPV			90.0	78.1

IV. Conclusions

In this paper we have introduced the biodata analytics module currently being developed inside the Welcome project. It is composed of two main sub-modules: an ECG analysis sub-module and a chest sound analysis sub-module. Currently an EIT analysis sub-module capable of using the EIT data collected by the Welcome vest to assess the ventilation function of the COPD patient is being researched to be incorporated into the Welcome decision support system.

References

[1] Martinez et al (2004); J. Martinez, R. Almeida, S. Olmos, A. Rocha, P. Laguna; A wavelet-based ECG delineator: evaluation on standard databases; IEEE Trans. Biomed. Eng., 51, 4, 570–581.

[2] Sun et al (2005); Y. Sun, K. Chan, S. Krishnan; Characteristic wave detection in ECG signal using morphological transform; BMC Cardiovascular Disorders 5:28 (2005).

[3] Tratnig (2005); R. Tratnig; Reliability of new Fibrillation Detection algorithms for automated External Defibrillators; PhD Dissertation, Technische Universitaet Graz.

[4] J. J. Rieta et al (2004); "Atrial activity extraction for atrial fibrillation analysis using blind source separation," Biomedical Engineering, IEEE Transactions on, vol. 51, pp. 1176-1186, 2004.

[5] Henriques et al (2007); J. Henriques, P. Carvalho, P. Gil, A. Marques, T. Rocha, B. Ribeiro, M. Antunes, R. schmidt, J. Habetha; Ventricular Arrhythmias Assessment; EMBC -2007, 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; Lyon, France, August 23-26, 2007.

[6] Henriques et al (2008); J. Henriques, P. Carvalho, M. Harris, M. Antunes, R. Couceiro, M. Brito, R. Schmidt; Assessment of Arrhythmias for Heart Failure Management; phealth2008 - International Workshop on Wearable Micro and Nanosystems for Personalised Health; Valencia, Spain, May 21-23, 2008.

[7] Rocha et al (2008); T. Rocha, S. Paredes, P. Carvalho, J. Henriques, M. Antunes; Phase space reconstruction approach for ventricular arrhythmias characterization; EMBC-08, 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vancouver, Canada, pp 5470-5473.