FEATURE SELECTION APPLIED TO THE TIME-FREQUENCY REPRESENTATION OF MUSCLE NEAR-INFRARED SPECTROSCOPY (NIRS) SIGNALS: CHARACTERIZATION OF DIABETIC

Journal of Mechanics in Medicine and Biology
Vol. 12, No. 4 (2012) 1240013 (14 pages)

DOI: 10.1142/S0219519412400131

World Scientific
www.worldscientific.com

FEATURE SELECTION APPLIED TO THE TIME-FREQUENCY REPRESENTATION OF MUSCLE NEAR-INFRARED SPECTROSCOPY (NIRS) SIGNALS: CHARACTERIZATION OF DIABETIC OXYGENATION PATTERNS

SAMANTA ROSATI*, GABRIELLA BALESTRA†
and FILIPPO MOLINARI‡

Biolab, Department of Electronics and Telecommunications
Politecnico di Torino, Turin, 10129, Italy
**samanta.rosati@polito.it*
†*gabriella.balestra@polito.it*
‡*filippo.molinari@polito.it*

Received 15 March 2012
Revised 21 May 2012
Accepted 18 June 2012
Published 3 August 2012

Diabetic patients might present peripheral microcirculation impairment and might benefit from physical training. Thirty-nine diabetic patients underwent the monitoring of the tibialis anterior muscle oxygenation during a series of voluntary ankle flexo-extensions by near-infrared spectroscopy (NIRS). NIRS signals were acquired before and after training protocols. Sixteen control subjects were tested with the same protocol. Time-frequency distributions of the Cohen's class were used to process the NIRS signals relative to the concentration changes of oxygenated and reduced hemoglobin. A total of 24 variables were measured for each subject and the most discriminative were selected by using four feature selection algorithms: QuickReduct, Genetic Rough-Set Attribute Reduction, Ant Rough-Set Attribute Reduction, and traditional ANOVA. Artificial neural networks were used to validate the discriminative power of the selected features. Results showed that different algorithms extracted different sets of variables, but all the combinations were discriminative. The best classification accuracy was about 70%. The oxygenation variables were selected when comparing controls to diabetic patients or diabetic patients before and after training. This preliminary study showed the importance of feature selection techniques in NIRS assessment of diabetic peripheral vascular impairment.

Keywords: Near-infrared spectroscopy; time-frequency distributions; feature selection; diabetes.

1. Introduction

Computer methods have been extensively used to aid the clinical assessment of diabetes-related pathologies, such as neuropathy[1] and sympathetic system impairment,[2] retinopathy and eye fundus damage,[3] central and peripheral reduced perfusion,[4] and gait abnormalities.[5] As diabetes is a complex pathology involving

multiple systems, several researchers proposed the use of classification techniques and decision systems to aid the clinical interpretation of the biomedical data. As an example, Acharya *et al.* proposed a nonlinear dynamics method for the classification of neuropahtic and myopathic myoelectric signals.[6] They showed that nonlinear dynamics analysis allowed discriminating subjects with muscular and neural impairments from controls with 99.3% accuracy. Several researchers focused their analysis on the lower limb muscular performance, because it could be a very early and specific indicator of the progression of peripheral arterial disease and neuropathy.

Besides electromyography, near-infrared spectroscopy (NIRS) was also used to monitor muscle oxygenation in subjects suffering from lower-extremity arterial disease[7] and diabetes.[8] NIRS is a low-cost, non-invasive and portable system that is very adept at monitoring during rest and exercise conditions. The NIRS system consists of near-infrared sources (usually LED or laser diodes) that inject electromagnetic radiation into human tissues. Light absorption allows for the estimation of the concentration of chromophores in the tissues once the effects of scattering have been compensated. The relevant chromophores in human tissues are oxygenated and reduced hemoglobin. Hemoglobin presents different absorption spectra when in reduced or oxygenated form, hence by using infrared light at different wavelengths, it is possible to continuously monitor the concentration changes of the two hemoglobin types in the tissues in a non-invasive manner.[9]

In this study, we used NIRS to monitor the oxygen and carbon dioxide concentration changes in the muscle tibialis anterior of diabetic patients and controls during a simple exercise of ankle flexo-extension, in order to evidence if NIRS parameters could be indicative of the peripheral vascular and metabolic pattern of the subjects. NIRS signals, however, require advanced signal processing procedures, because relevant information is carried by the signals' spectrum. As such signals are non-stationary when recorded during activation (i.e., muscle contraction), we adopted a time-frequency-based analytical procedure. This technique allowed for the extraction of several spectral features (which can exceed 50) that made difficult the interpretation of results. Multivariate analysis and dimensionality reduction approaches could be useful to improve the readability of the results.[10]

Feature selection (FS) allows dimensionality reduction of multivariate data, deleting the unnecessary attributes in order to extract the most significant features for the system description. In fact, a too large number of features does not inevitably increase the classification accuracy: several attributes may be redundant, irrelevant, or even worse, may introduce some kind of noise which decreases the classifier performance.[11]

For most real applications, in which the number of initial variables is relatively medium-high, an exhaustive search of the best feature subset results inapplicable. For this reason, during the past years, several methods for FS based on a heuristic search have been developed.[12–14] Heuristics indentify a wide class of algorithms used in order to solve optimization problems, reducing time and computational costs by

addressing the solution search toward a high-quality space of admissible solutions. Most of them are based on a local search guided by an objective function that provides an estimate of the current solution goodness. The possibility of easily representing complex and non linear objective functions makes the heuristic algorithms a suitable tool for FS.

Several approaches are available for FS and recently, the rough-set theory (RST) has been applied in this field with very satisfactory results. RST was introduced by Pawlak[15] in order to provide an instrument that is able to model imperfect and incomplete knowledge encapsulated in the real data. Its concepts find wide and different areas of application, such as machine learning,[16] knowledge acquisition,[17,18] decision analysis,[19,20] pattern recognition,[21] knowledge discovery from databases and expert systems,[22] and dimensionality reduction.[23] The main advantage of RST-based FS methods is that they do not require any *a-priori* information or model assumptions about data.

In this study, we applied four FS algorithms to a set of features extracted from NIRS signals recorded in diabetic subjects during ankle flexo-extension, with the aim of extracting the relevant NIRS features, which are characteristic of the subjects' peripheral vascular status.

2. Materials and Methods

2.1. *Demographics, data acquisition, and experimental protocol*

This study involved 39 type II diabetic subjects. Each of them performed daily physical activity for one year: 19 subjects carried out adapted physical activity (APA) (age: 66.7 ± 5.7) and 20 patients performed fit walking (FW) (age: 66.0 ± 6.2). Moreover, a group of 16 healthy subjects (Contr) (age: 65.3 ± 3.9) was included in the study as controls.

NIRS signals were recorded on each subject before and after the period of physical exercise by using a commercially available NIRS system (NIRO300, Hammamatsu Photonics, Japan). The recordings of NIRS signals were performed with the emitting probe placed on the left tibialis anterior muscle, approximately in correspondence with the muscle belly. The receiving photodiode was placed 4 cm apart, aligned with the emitters in the direction of the muscle fibers. In this work, three NIRS signals were considered: the changes in the concentration of oxygenated hemoglobin (O2Hb), the changes in the concentration of deoxygenated hemoglobin (HHb), and the tissue oxygenation index (TOI) defined as the ratio of oxygenated to total hemoglobin.

The experimental protocol was structured as follows:

(1) one minute of resting,
(2) ankle flexo-extension for 3 min, and
(3) a final resting period of 3 min in order to observe the recovery phase.

The total duration for the examination was about 7 min.

2.2. *Feature extraction*

In Fig. 1, the left panels show an example of NIRS signals recorded on a healthy subject performing ankle flexo-extension. Panel A1 reports the O2Hb concentration variation during time, Panel B1 the HHb, and Panel C1 the TOI signals. The vertical dashed lines mark the onset and offset of the flexo-extension. The hemoglobin concentration significantly varies during time: flexo-extension corresponds to a decrease in the O2Hb and an increase in the HHb concentrations.

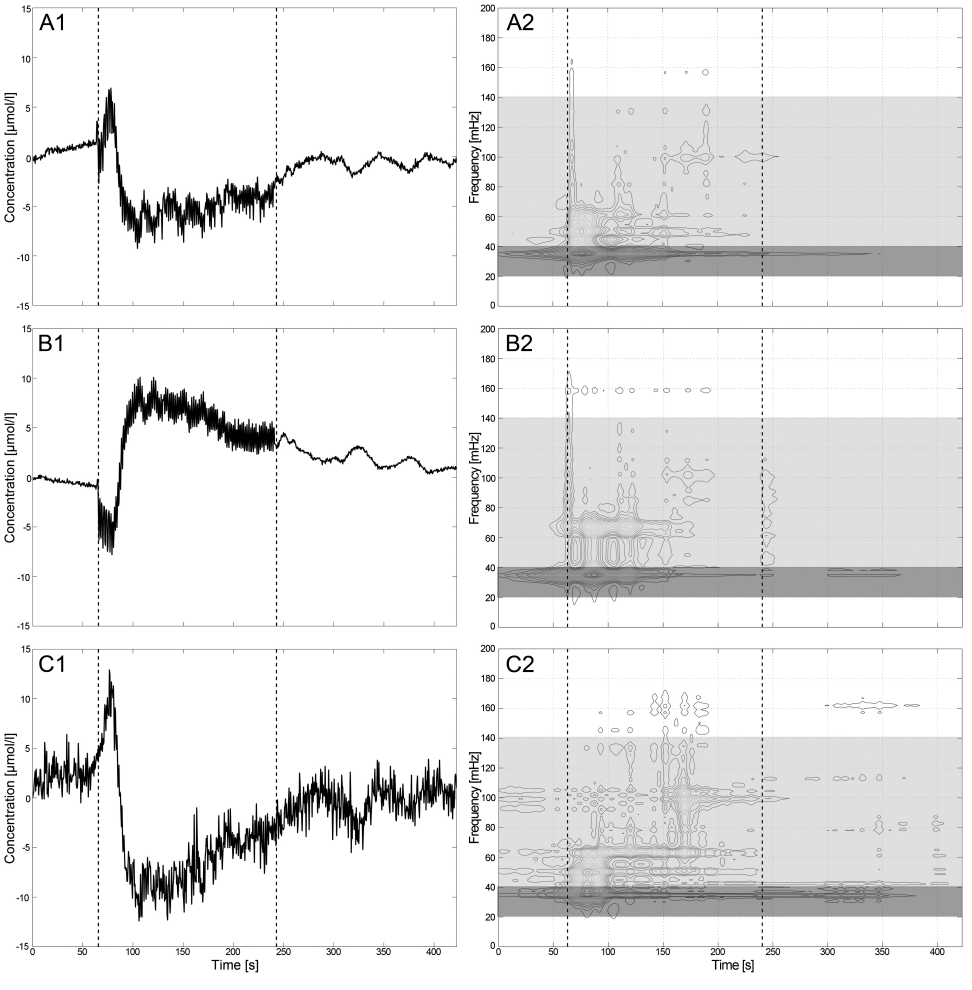

Fig. 1. NIRS signals recorded on a healthy subject performing ankle flexo-extension. Panel A1 reports the O2Hb concentration signal, panel B1 the HHb, and panel C1 the TOI. The right panels show the 15-level contour plot of the Choi-Williams time-frequency distribution of the signals depicted in the left side ($\sigma = 0.05$). The vertical rectangles highlight the LF band (40−140 mHz) and the VLF band (20−40 mHz). The black vertical dashed lines mark the onset and offset of the flexo-extension.

One of the problems of the NIRS recordings is that only relative concentration measures are available. Therefore, the concentration in time of the O2Hb and HHb were not specific of the different groups. We performed a spectral analysis of the NIRS signals because it contains the signature of the sympathetic and para-sympathetic nervous drive.[24]

However, as it emerges from Fig. 1, NIRS signals are clearly non-stationary during the periods of interest and this makes the traditional Fourier-based spectral analysis not applicable. The time-frequency distributions allow for the managing of the non-stationarity, analyzing the signals both in the time and the frequency domain. Specifically, the Choi-Williams (CW) transform,[25] that is a time-frequency distribution belonging to the Cohen's class, was used in this study. The value of σ was set equal to 0.5 for all signals because this value already proved effective in the analysis of several biological signals.

In the right side of Fig. 1, a CW representation of the NIRS signals during the experiment for a healthy subject is reported. Moreover, the time-frequency squared coherence function (SCF), between the concentration signals of O_2Hb and HHb, was evaluated according to the following equation:

$$SCF = \frac{|D_{xy}(t,f)|^2}{D_{xx}(t,f) \cdot D_{yy}(t,f)}, \tag{1}$$

where $D_{xy}(tf)$ is the cross time-frequency CW distribution of O_2Hb and HHb, $D_{xx}(t,f)$ and $D_{yy}(tf)$ are the time-frequency CW representation of O_2Hb and HHb, respectively. As the SCF is a quadratic function, it assumes only values between 0, if the two signals are totally uncorrelated, and 1, if they are totally correlated.

All the signals were preprocessed, converting them to the analytical representation with zero mean. Additionally, a high-pass Chebychev filter was used, with ripple in the stopband and cutoff frequency equal to 25 mHz, in order to remove very slow components.

All time-frequency distributions were analyzed in two specific bands, very low frequencies (VLF) (20−40 mHz) and low frequencies (LF) (40−140 mHz), before, during, and after the flexo-extension. The percentage of signal power in the two bands (referred to the total power of the signal) was calculated for each period. This procedure led to obtain the following 24 variables:

(1) the HHb, O2Hb and TOI power in the VLF band, before, during, and after the flexo-extension (nine variables),
(2) the HHb, O2Hb and TOI power in the LF band, before, during, and after the flexo-extension (nine variables),
(3) the SCF between O2Hb and HHb in the VLF band, before, during, and after the experiment (three variables), and
(4) the SCF between O2Hb and HHb in the LF band, before, during, and after the experiment (three variables).

2.3. *Feature selection algorithms*

In this study four FS procedures were performed and compared. Three of them were based on the RST concepts: QuickReduct Algorithm (QRA), Genetic Rough-Set Attribute Reduction (GenRSAR), and Ant Rough-Set Attribute Reduction (AntRSAR). The last method was the ANalysis Of Variance (ANOVA), based on the computation of the variances within and between groups.

The FS was applied to four datasets in which the rows stand for the subjects and the columns report the 24 features. The datasets were built as follows:

(1) diabetic patients performing APA were taken as subjects and the variables related to the signals acquired before and after the period of physical exercise represented two different classes (APA-PRE vs POST),
(2) diabetic patients performing FW were taken as subjects and the variables related to the signals acquired before and after the period of physical exercise represented two different classes (FW-PRE vs POST),
(3) the features acquired before the period of exercise for control and diabetic subjects were considered as two different classes (PRE-CONTR vs DIAB), and
(4) the features acquired after the period of exercise for control and diabetic subjects were considered as two different classes (POST-CONTR vs DIAB).

MATLAB environment was used to implement all FS procedures. All datasets were tested in order to detect outliers using the Wilks' method.[26] No outlier was found in any dataset. Moreover, as the dependency degree can only be measured on discrete data, a discretization strategy was applied to the four datasets in order to transform continuous values into discrete ones. In this study, three ranges of values were identified for each variable, based on the knowledge about features and the data plots.

2.3.1. *QuickReduct algorithm*

QRA, introduced in Ref. 27, is the basic algorithm employing RST that allows resolving reduct search problems without generating all the possible subsets.

In RST, data are organized as a decision system (or Decision Table−DT) made up of a non-empty set of objects U (the Universe of discourse), each of them characterized by a non-empty set of attributes A. The attributes are divided in a certain number of conditional attributes C, which represent the input features, and a decision attribute D, which is the class the objects belong to.

Given a DT with discretized attribute values, it is possible to find the minimal subset R (reduct) of the original features, using RST, which is the most informative. Different criteria can be used in order to evaluate the relevance of the chosen feature subset;[12] in this study, the dependency degree $\gamma_C(D)$ was employed. It evaluates the dependency of the decision attribute D from the set of the conditional features C and is comprised between 0, if there is no any dependency, and 1, if all values from D are uniquely determined by values of attributes C.

The QRA algorithm starts from an empty subset of features and adds to it the best attributes, until a stopping criterion is satisfied. As the goal of QRA is to find a reduct with the same dependency degree of the whole set of attributes, this parameter is chosen as the stopping criterion. The maximum dependency value results in 1 if the dataset is consistent. Consequently, attributes added to the reduct subset are those producing a larger increase in the dependency degree.

This algorithm, however, is not guaranteed to find a minimal reduct as the feature subset discovered may contain irrelevant or redundant attributes.

2.3.2. *Genetic rough-set attribute reduction*

Genetic Algorithms (GAs), introduced by Holland[28] and belonging to the evolutionary algorithms, are a class of metaheuristics which aims to mimic the natural evolutionary process of species. The main idea is that only the best individuals are able to survive and transmit their genes to the subsequent generations.

In GAs, each individual (chromosome) represents a possible solution to the problem and it is evaluated by means of a fitness function that expresses the solution goodness. Therefore, an adequate choice of the fitness function has to be performed, so that it is directly correlated with the individual suitability. The best individuals are chosen and evolved in order to become parents of a future generation of solutions, until an optimal solution to the problem is found.

Because of the chromosomal structure (usually represented as a binary string), GAs can easily adapt to FS problems. In fact, for these applications, each bit can be associated to a specific feature: a one or zero in a certain position of the string indicates, respectively, if the associated feature is employed or not in the current solution.

There are many applications of GAs using the RST. In Ref. 29, the GenRSAR is introduced — that is an algorithm employing a genetic search strategy in order to determine a feature reduct. It is based on a standard GA structure in which the fitness function takes into account both the size of subset R and its suitability:

$$\text{fitness}(R) = \gamma_R(D) * \frac{|C| - |R|}{|C|}, \tag{2}$$

where $\gamma_R(D)$ is the dependency degree of subset R with respect to the classification $D, |\cdot|$ represents the subset cardinality, and C is the full set of features.

In this study, the initial population was made of 100 randomly generated individuals, the probability of mutation and crossover were set to 0.4 and 0.6 respectively, and the number of generations was equal to 100, as proposed in Ref. 29.

2.3.3. *Ant rough-set attribute reduction*

Ant colony optimization (ACO) is a methodology belonging to the Swarm Intelligent algorithms, inspired to the social behavior of those species which compete for food (ants, bees, etc.). ACO reproduces the same strategy used by ants in order to find the

best path in the direction of the food source. The main idea is that each ant, during its route from the colony nest to the food source, deposits on the ground a chemical hormone called pheromone. This substance will help other ants in selecting in the best trail to arrive at the food. The higher the pheromone quantity on a path, the higher the probability that the ants select that path to go to the food.

From the ACO point of view, real problems can be represented as a graph made up of nodes, representing the solution parts, and edges, the ants' paths. At the beginning of the search process, each ant is placed on a node; a complete solution is reached by adding a new node to the current ant pathway and assessing the next edge to travel.

This methodology, as GAs, is easily adaptable for FS. As ACO requires a problem to be represented as a graph, nodes may be associated to features (one node for each attribute) and edges may denote the choice of the next feature.[29]

In Ref. 29, the AntRSAR is proposed, that is a RST-based ACO in which the dependency degree is used to verify when ants terminate their pathway. This means that the search is stopped when the ant feature subset reaches the same degree of dependency $\gamma_{\mathrm{R}}(D)$ of the whole attribute set. As for the desirability measure, the conditional information entropy $H(D|A)$ produced by an attribute A with respect to the decision feature D[30] is used.

In this study, the number of ants is set equal to the number of features, with each ant starting from a different feature. Furthermore, pheromone levels are set at 0.5 with small random variations added, α and β are set to 1 and 0.1 respectively, and the algorithm terminates after 100 iterations.

2.3.4. *ANOVA analysis*

ANOVA allows the analysis of datasets based on the variances within and between data groups, assuming a linear model for data. In this study, the one-way ANOVA analysis was performed in order to assess if the subject classification has a statistical influence on each feature. We considered the subject classification as the independent variable and the 24 parameters extracted for each subject as the dependent variable, one at a time. Then, the features producing a P value smaller than 5% were selected to constitute a feature subset.

3. Results

All datasets used in this study have a dependency degree equal to 1. The feature subsets selected by the four FS methods are listed in Table 1, in which a "1" indicates the selected variable. The total number of features composing each subset is reported in the last row of Table 1.

All FS methods based on RST return subsets with a dependency degree equal to 1 and with a number of features between 5 and 12. Moreover, AntRSAR returns subsets with the largest number of features. As for ANOVA, it gives feature subsets with a number of variables from 0 to 5 and a dependency degree between 0 and 0.34.

Table 1. Results of the four FS procedures applied to the four datasets. First column contains the 24 variables used as input for the FS strategies. From the second to the last column are the results of QRA, GenRSAR, AntRSAR, and ANOVA applied to APA − PRE vs POST, FW − PRE vs POST, PRE − CONTR vs DIAB, POST − CONTR vs DIAB datasets (1: feature selected). The selected parameters are highlighted in light grey. The last row contains the number of features selected in each subset.

APA PREvsPOST						FW PREvsPOST						PRE CONTRvsDIAB						POST CONTRvsDIAB					
Features	QRA	GenRSAR	AntRSAR	ANOVA	N°of occurrences	Features	QRA	GenRSAR	AntRSAR	ANOVA	N°of occurrences	Features	QRA	GenRSAR	AntRSAR	ANOVA	N°of occurrences	Features	QRA	GenRSAR	AntRSAR	ANOVA	N°of occurrences
$\%P_{VLF}$-post-O_2Hb	1	1	0	1	3	$\%P_{VLF}$-post-O_2Hb	0	1	1	0	2	$\%SCF_{LF}$-post	1	1	1	1	4	$\%P_{LF}$-pre-O_2Hb	1	1	1	0	3
$\%P_{VLF}$-post-TOI	1	1	0	1	3	$\%P_{LF}$-dur-TOI	1	0	1	0	2	$\%P_{VLF}$-pre-O_2Hb	1	1	1	0	3	$\%SCF_{VLF}$-post	1	1	1	0	3
$\%P_{VLF}$-dur-TOI	1	0	1	0	2	$\%SCF_{VLF}$-dur	1	1	0	0	2	$\%P_{VLF}$-dur-O_2Hb	1	0	1	0	2	$\%SCF_{LF}$-post	1	1	1	0	3
$\%P_{LF}$-pre-O_2Hb	0	1	1	0	2	$\%SCF_{VLF}$-post	0	1	1	0	2	$\%P_{VLF}$-pre-TOI	0	1	0	1	2	$\%P_{VLF}$-post-O_2Hb	1	0	1	0	2
$\%P_{LF}$-post-O_2Hb	0	0	1	1	2	$\%SCF_{LF}$-pre	1	0	1	0	2	$\%P_{VLF}$-dur-TOI	1	0	0	1	2	$\%P_{LF}$-pre-HHb	0	1	1	0	2
$\%P_{LF}$-pre-TOI	1	0	1	0	2	$\%SCF_{LF}$-post	1	0	1	0	2	$\%P_{LF}$-pre-TOI	0	0	1	1	2	$\%P_{VLF}$-pre-O_2Hb	1	0	0	0	1
$\%P_{LF}$-post-TOI	0	0	1	1	2	$\%P_{VLF}$-dur-O_2Hb	1	0	0	0	1	$\%SCF_{LF}$-pre	0	1	1	0	2	$\%P_{VLF}$-dur-HHb	0	1	0	0	1
$\%SCF_{VLF}$-post	0	1	1	0	2	$\%P_{VLF}$-pre-HHb	0	0	1	0	1	$\%SCF_{LF}$-dur	1	0	1	0	2	$\%P_{VLF}$-post-HHb	1	0	0	0	1
$\%P_{VLF}$-pre-O_2Hb	1	0	0	0	1	$\%P_{VLF}$-dur-HHb	1	0	0	0	1	$\%P_{VLF}$-post-O_2Hb	0	1	0	0	1	$\%P_{VLF}$-pre-TOI	0	1	0	0	1
$\%P_{VLF}$-dur-O_2Hb	0	0	1	0	1	$\%P_{VLF}$-post-HHb	1	0	0	0	1	$\%P_{VLF}$-pre-HHb	0	0	1	0	1	$\%P_{VLF}$-post-TOI	0	0	0	1	1
$\%P_{VLF}$-dur-HHb	1	0	0	0	1	$\%P_{VLF}$-pre-TOI	1	0	0	0	1	$\%P_{VLF}$-dur-HHb	1	0	0	0	1	$\%P_{LF}$-dur-O_2Hb	0	0	1	0	1
$\%P_{VLF}$-post-HHb	1	0	0	0	1	$\%P_{VLF}$-dur-TOI	0	0	1	0	1	$\%P_{VLF}$-post-TOI	0	0	1	0	1	$\%P_{LF}$-post-O_2Hb	0	0	1	0	1
$\%P_{VLF}$-pre-TOI	1	0	0	0	1	$\%P_{LF}$-pre-O_2Hb	0	0	1	0	1	$\%P_{LF}$-pre-O_2Hb	0	0	1	0	1	$\%SCF_{VLF}$-pre	1	0	0	0	1
$\%P_{LF}$-dur-HHb	0	0	1	0	1	$\%P_{LF}$-dur-O_2Hb	0	1	0	0	1	$\%P_{LF}$-dur-O_2Hb	0	0	1	0	1	$\%SCF_{VLF}$-dur	0	0	1	0	1
$\%P_{LF}$-dur-TOI	0	0	1	0	1	$\%P_{LF}$-post-O_2Hb	0	0	1	0	1	$\%P_{LF}$-post-O_2Hb	0	0	1	0	1	$\%SCF_{LF}$-pre	0	0	1	0	1
$\%SCF_{VLF}$-dur	0	1	0	0	1	$\%P_{LF}$-pre-HHb	0	1	0	0	1	$\%P_{LF}$-pre-HHb	1	0	0	0	1	$\%SCF_{LF}$-dur	0	0	1	0	1
$\%SCF_{LF}$-dur	0	0	1	0	1	$\%P_{LF}$-dur-HHb	0	0	1	0	1	$\%P_{LF}$-post-TOI	1	0	0	0	1	$\%P_{VLF}$-dur-O_2Hb	0	0	0	0	0
$\%SCF_{LF}$-post	0	0	1	0	1	$\%P_{LF}$-pre-TOI	0	0	1	0	1	$\%SCF_{VLF}$-dur	1	0	0	0	1	$\%P_{VLF}$-pre-HHb	0	0	0	0	0
$\%P_{VLF}$-pre-HHb	0	0	0	0	0	$\%SCF_{LF}$-dur	0	0	1	0	1	$\%SCF_{VLF}$-post	0	1	0	0	1	$\%P_{VLF}$-dur-TOI	0	0	0	0	0
$\%P_{LF}$-dur-O_2Hb	0	0	0	0	0	$\%P_{VLF}$-pre-O_2Hb	0	0	0	0	0	$\%P_{VLF}$-post-HHb	0	0	0	0	0	$\%P_{LF}$-dur-HHb	0	0	0	0	0
$\%P_{LF}$-pre-HHb	0	0	0	0	0	$\%P_{VLF}$-post-TOI	0	0	0	0	0	$\%P_{LF}$-dur-HHb	0	0	0	0	0	$\%P_{LF}$-post-HHb	0	0	0	0	0
$\%P_{LF}$-post-HHb	0	0	0	0	0	$\%P_{LF}$-post-HHb	0	0	0	0	0	$\%P_{LF}$-post-HHb	0	0	0	0	0	$\%P_{LF}$-pre-TOI	0	0	0	0	0
$\%SCF_{VLF}$-pre	0	0	0	0	0	$\%P_{LF}$-post-TOI	0	0	0	0	0	$\%P_{LF}$-dur-TOI	0	0	0	0	0	$\%P_{LF}$-dur-TOI	0	0	0	0	0
$\%SCF_{LF}$-pre	0	0	0	0	0	$\%SCF_{VLF}$-pre	0	0	0	0	0	$\%SCF_{VLF}$-pre	0	0	0	0	0	$\%P_{LF}$-post-TOI	0	0	0	0	0
N° of features	8	5	11	4		N° of features	8	5	12	0		N° of features	9	6	11	4		N° of features	7	6	10	1	

The four FS strategies were compared using artificial neural networks (ANNs), based on the percentage of correct classification of the subjects in their classes. The main concept was that a good procedure of FS allows the removal of redundant features so that the reduced set provides at least the same quality of classification of the original set.[31]

Specifically, we tested the whole set of features and all the extracted subsets independently, using a different network for each one. The ANN structure was made of one hidden layer with a number of neurons approximately equal to half the input neurons. As for the neuron activation functions, we used a logarithmic sigmoid function for the hidden layer and a linear function for the output layer. Back-propagation was chosen as the learning algorithm and the mean squared error was

used as performance function. The initial values of interconnection weights were set randomly.

The ANNs were implemented by means of the Neural Network Matlab toolbox using 70% of input data randomly collected as the training set, while the complete dataset was used as the test set. As the number of samples in the datasets was limited, each ANN was run 20 times, with a different training set each time. The results in terms of percentage of correct classification for each subset and for the whole set of features are reported in Figs. 2 and 3 as mean value and best performance, respectively.

The mean percentage of correct classification is above 50% for all subsets, while the maximum values are above 70%. The highest mean values are obtained with PRE - CONTR vs DIAB and POST - CONTR vs DIAB datasets, and this confirmed the hypothesis that the sample number contained in the dataset influences the performances obtainable with ANNs. Moreover, the results achieved using all features are comparable with those obtained with the extracted subsets and are not correlated with the number of selected features.

4. Discussion

In this study, we analyzed the oxygenation pattern of the tibialis anterior muscle of a cohort of diabetic patients undergoing a training protocol (either FW or APA). Our findings, although still preliminary, demonstrated the utility of muscular NIRS in the assessment of the subjects' peripheral vascular pattern. We showed that by selecting the most relevant features of the NIRS signals, it is possible to characterize the

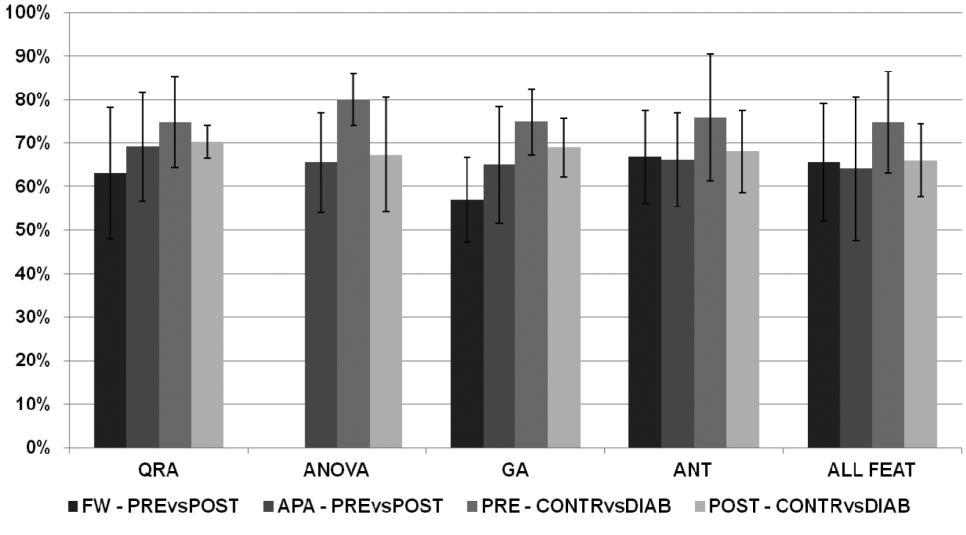

Fig. 2. ANNs results in terms of mean value (on twenty runs) of percentage of correct subjects' classification for each selected feature subset and for the whole set of features, applied to the four datasets.

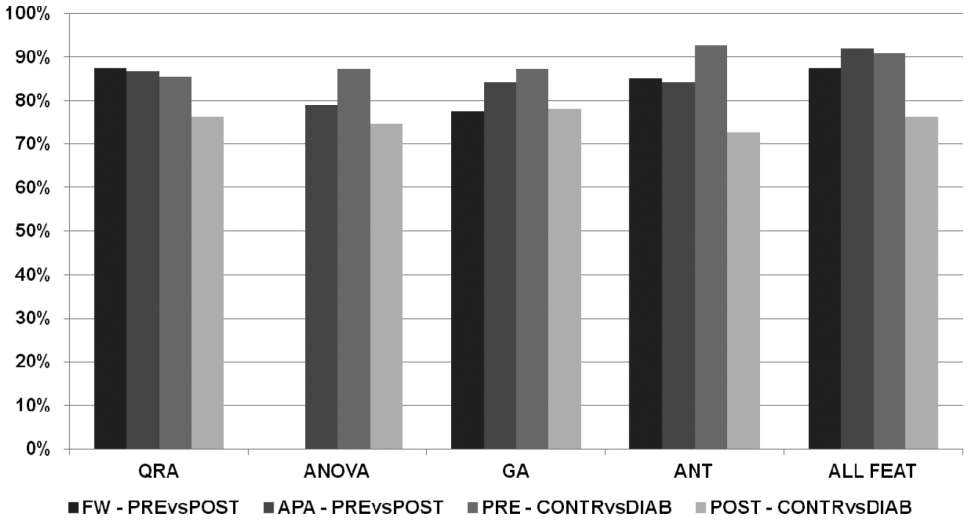

Fig. 3. ANNs results in terms of best value (on 20 runs) of percentage of correct subjects' classification for each selected feature subset and for the whole set of features, applied to the four datasets.

diabetic patients with respect to the activity they performed or against controls, with a classification performance of 70% or higher.

There is a fundamental difference between our study and some other recent researches focused on muscular NIRS.[32,33] Our signal analysis and feature extraction procedure was based on time-frequency distributions, whereas (usually) NIRS signals are analyzed in the time domain. For example, de Blasi *et al.*[8] documented alterations in skeletal muscle blood flow and oxygenation in diabetic patients undergoing haemodialysis with respect to controls. Their analysis was based on the changes in the O2Hb and HHb concentrations recorded during haemodialysis. However, we documented a very high variability in the time course of the NIRS signals (Fig. 1) and did not find any significant time parameter that could discriminate diabetic patients from controls or diabetic patients undergoing different training protocols. We focused our analysis on the spectral changes of the NIRS signals in order to understand if the pathology (diabetes) or the training (FW *vs.* APA) differentiated the signals' spectrum. The VLF band is directly related to long-term metabolic adaptation and regulation, whereas the LF band is linked to the activity of the vagal nerve and the microcirculation vassal reactivity.[24] Since the number of features that can be extracted from the time-frequency analysis is very high, we studied the potentialities of feature selection to extract the relevant information for the characterization of the subjects' microvascular pattern.

Analyzing Table 1 as a whole, we can see that the FS algorithms provided solutions that have a small number of variables in common. Only one feature (% PLF−post−HHb, i.e., the power of the HHb signal in the LF band after the contraction) was never selected. This is probably due to the high concentration of

HHb during the test, which is not significantly different among the subjects. Studies reported that the most significant difference between diabetic and healthy muscles is the O2Hb concentration change as detected by NIRS, rather than the HHb concentration.[8]

If we relate this finding with the classification performances, we can conclude that there are disfferent combinations of the features that are able to distinguish the groups. Moreover, as the classification results obtained using all the features are comparable with the classification results based on the reduced sets, we can deduce that the FS identified discriminative parameters. Considering the four data sets separately, we can observe that there is less difference among the results and that each data set is characterized by a different set of parameters. When comparing diabetic patients pre- and post-APA, all the most discriminative features were direct measures of tissue oxygenation (Table 1, first column) derived either from the O2Hb or TOI signals. Since the percentage power in the bands of the TOI signal was discriminative before (pre), during (dur), and also after (post) the test contraction, we hypothesized that APA effectively improved tissue oxygenation. If we consider the second column of Table 1, relative to FW, we observe a predominance of features computed on the square coherency function between the O2Hb and HHb signals. This indicates an improvement on tissue oxygen-carbon dioxide balancing, thus showing a positive effect of FW on the muscle oxygenation and metabolism. Finally, if we consider the last two columns, which compare controls to diabetic patients before and after training, we again observe a predominance of features computed either on the O2Hb signal or on the TOI signal. In both cases, the SCF in the LF band was also often present as discriminative feature, thus confirming the difference in the microcirculation autoregulatory pattern of the diabetics with respect to controls.

In conclusion, in this study, we applied feature selection strategies to the time-frequency distributions of NIRS signals acquired from diabetic patients and healthy controls before and after a physical training protocol. Different strategies extracted different features, but the classification performance based on the reduced set of features was satisfactory. This preliminary study confirmed the validity of NIRS in the assessment and monitoring of diabetic patients microcirculation impairment and suggested the need for improved classification schemes, in order to gain a better comprehension of the diabetic muscle vascular pattern.

References

1. Acharya RU, *et al.*, Computer-based identification of plantar pressure in type 2 diabetes subjects with and without neuropathy, *J Mech Med Biol* **8**:363−375, 2008.
2. Desai K, Ghista DN, *et al.*, Diabetic autonomic neuropathy detection by heart-rate variability power-spectral analysis, *J Mech Med Biol.*
3. Rossant F, Badellino M, Chavillon A, Bloch I, Paques M, A morphological approach for vessel segmentation in eye fundus images, with quantitative evaluation, *J Med Imaging Health Inf* **1**:42−49, 2011.

4. Last D, Alsop DC, Abduljalil AM, Marquis RP, de Bazelaire C, Hu K, Cavallerano J, Novak V, Global and regional effects of type 2 diabetes on brain tissue volumes and cerebral vasoreactivity, *Diabetes Care* **30**:1193−1199, 2007.
5. Sawacha Z, Gabriella G, Cristoferi G, Guiotto A, Avogaro A, Cobelli C, Diabetic gait and posture abnormalities: A biomechanical investigation through three dimensional gait analysis, *Clin Biomech* **24**:722−728, 2009.
6. Acharya RU, Ng EYK, Swapna G, Michelle YSL, Classification of normal, neuropathic, and myopathic Electromyograph signals using nonlinear dynamics method, *J Med Imaging Health Inf* **1**:375−380, 2011.
7. Comerota AJ, Throm RC, Kelly P, Jaff M, Tissue (muscle) oxygen saturation (StO2): A new measure of symptomatic lower-extremity arterial disease, *J Vasc Surg* **38**:724−729, 2003.
8. De Blasi RA, Luciani R, Punzo G, Arcioni R, Romano R, Boezi M, Menè P, Microcirculatory changes and skeletal muscle oxygenation measured at rest by non-infrared spectroscopy in patients with and without diabetes undergoing haemodialysis, *Crit Care* **13**:S9, 2009.
9. Xue H, Bestonzo M, Acharya RU, Molinari F, Design and implementation of a continuous wave near infrared spectroscopy system for bedside and home monitoring, *J Med Imaging Health Inf* **1**:317−324, 2011.
10. Molinari F, Rosati S, Liboni W, Negri E, Mana O, Allais G, Benedetto C, Time-frequency characterization of cerebral hemodynamics of migraine sufferers as assessed by NIRS signals, *EURASIP J Adv Signal Process*, 2010.
11. Jensen R, Shen Q, *Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches*, Wiley, Hoboken, 2008.
12. Dash M, Liu H, Feature selection for classification, *Intell Data Anal* **1**:131−156, 1997.
13. Somol P, Novovi J, Pudil P, Notes on the evolution of feature selection methodology, *Kybernetika* **43**:713−730, 2007.
14. He Z, Yu W, Stable feature selection for biomarker discovery, *Comput Biol Chem* **34**:215−225, 2010.
15. Pawlak Z, Rough sets, *Int J Comp Inf* **11**:341−356, 1982.
16. Moradi H, Grzymala-Busse JW, Roberts JA, Entropy of english text: Experiments with humans and a machine learning system based on rough sets, *Inf Sci* **104**:31−47, 1998.
17. Feng L, Wang GY, Li XX, Knowledge acquisition in vague objective information systems based on rough sets, *Expert Syst* **27**:129−142, 2010.
18. Matsumoto Y, Watada J, Knowledge acquisition from time series data through rough sets analysis, *Int J Innovations Comput I* **5**:4885−4897, 2009.
19. Greco S, Matarazzo B, Slowinski R, Rough sets theory for multicriteria decision analysis, *Eur J Oper Res* **129**:1−47, 2001.
20. Pawlak Z, Sowinski R, Rough set approach to multi-attribute decision analysis, *Eur J Oper Res* **72**:443−459, 1994.
21. Swiniarski RW, Skowron A, Rough set methods in feature selection and recognition, *Pattern Recogn Lett* **24**:833−849, 2003.
22. Tsumoto S, Automated extraction of medical expert system rules from clinical databases based on rough set theory, *Inf Sci* **112**:67−84, 1998.
23. Thangavel K, Pethalakshmi A, Dimensionality reduction based on rough set theory: A review, *Appl Soft Comput* **9**:1−12, 2009.
24. Obrig H, Neufang M, Wenzel R, Kohl M, Steinbrink J, Einhäupl K, Villringer A, Spontaneous low frequency oscillations of cerebral hemodynamics and metabolism in human adults, *Neuroimage* **12**:623−639, 2000.
25. Choi H, Williams WJ, Improved time-frequency representation of multicomponent signals using exponential kernels, *IEEE Trans Acoust Speech* **37**:862−871, 1989.

26. Wilks SS, Multivariate statistical outliers, *Sankhya* **25**:407−426, 1963.
27. Shen Q, Chouchoulas A, Modular approach to generating fuzzy rules with reduced attributes for the monitoring of complex systems, *Eng Appl Artif Intell* **13**:263−278, 2000.
28. Holand J, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, 1975.
29. Jensen R, Shen Q, Finding rough set reducts with ant colony optimization, *Proc of the 2003 UK Workshop on Computational Intelligence* 15−22, 2003.
30. Jensen R, Shen Q, A rough set-aided system for sorting WWW bookmarks, in Zhong N *et al.* (eds.), *Web Intelligence: Research and Development*, Springer, Berlin, pp. 95−105, 2001.
31. Chen Y, Miao D, Wang R, Wu K, A rough set approach to feature selection based on power set tree, *Knowl-Based Syst* **24**:275−281, 2011.
32. Ferreira LF, Lutjemeier BJ, Townsend DK, Barstow TJ, Effects of pedal frequency on estimated muscle microvascular O2 extraction, *Eur J Appl Physiol* **96**:558−63, 2006.
33. Hirata K, Hara T, Oshima Y, Yoshikawa T, Fujimoto S, Effects of electrical stimulation and voluntary exercise on muscle oxygenation assessed by NIRS, *Osaka City Med J* **52**:67−78, 2006.