
Article:
Vargas Cardona, H.D., Álvarez, M.A. orcid.org/0000-0002-8980-4472 and Orozco, Á.A. (2017) Multi-task learning for subthalamic nucleus identification in deep brain stimulation. International Journal of Machine Learning and Cybernetics. ISSN 1868-8071

https://doi.org/10.1007/s13042-017-0640-5

Noname manuscript No.
(will be inserted by the editor)

Multi-task learning for subthalamic nucleus identification in deep brain stimulation

Hernán Darío Vargas Cardona · Mauricio A. Álvarez · Álvaro A. Orozco

Received: date / Accepted: date

Abstract Deep brain stimulation (DBS) of Subthalamic nucleus (STN) is the most successful treatment for advanced Parkinson's disease. Localization of the STN through Microelectrode recordings (MER) is a key step during the surgery. However, it is a complex task even for a skilled neurosurgeon. Different researchers have developed methodologies for processing and classification of MER signals to locate the STN. Previous works employ the classical paradigm of supervised classification, assuming independence between patients. The aim of this paper is to introduce a patient-dependent learning scenario, where the predictive ability for STN identification at the level of a particular patient, can be used to improve the accuracy for STN identification in other patients. Our inspiration is the multi-task learning framework, that has been receiving increasing interest within the machine learning community in the last few years. To this end, we employ the multi-task Gaussian processes framework that exhibits state of the art performance in multi-task learning problems. In our context, we assume that each patient undergoing DBS is a different task, and we refer to the method as multi-patient learning. We show that the multi-patient learning framework improves the accuracy in the identification of STN in a range from 4.1% to 7.7%, compared to the usual patient-independent setup, for two different datasets. Given that MER are non stationary and noisy signals. Traditional approaches in machine learning fail to recognize accurately the STN during DBS. By contrast in our proposed method, we properly exploit correlations between patients with similar diseases, obtaining an additional information. This information allows to improve the accuracy not only for locating STN for DBS but also for other biomedical signal classification problems.

Keywords Parkinson's Disease · Deep Brain stimulation · MER signals Processing · Multi-Task Gaussian Processes.

Hernán Darío Vargas Cardona
Doctorate Program in Engineering, Universidad Tecnológica de Pereira, Vereda La Julita, Pereira, Colombia
E-mail: hernan.vargas@utp.edu.co

Mauricio A. Álvarez
Faculty of Engineering, Universidad Tecnológica de Pereira, Vereda La Julita, Pereira, Colombia
E-mail: malvarez@utp.edu.co

Álvaro A. Orozco
Faculty of Engineering, Universidad Tecnológica de Pereira, Vereda La Julita, Pereira, Colombia
E-mail: aaog@utp.edu.co

1 Introduction

Parkinson's Disease (PD) is a progressive degenerative condition of the Central Nervous System (CNS). Most common symptoms of PD are the hypokinesia, bradykinesia, lower verbal fluency and tremor (Wright et al 2008; Lees et al 2009; Miocinovic et al 2009). It is known, though, that one of the reasons for its development is related to the deterioration of cells in a structure called substantia nigra pars reticulata (SNr), generating a loss of a neurotransmitter known as dopamine. This makes it impossible for people to control their movements, leading to the primary motor symptoms of PD (NINDS 2004). Patients with PD are usually subjected to drug treatment with Levodopa. In more advanced stages of the disease, it becomes necessary to proceed with a surgical treatment. Deep Brain Stimulation (DBS) of Subthalamic Nucleus (STN) is the most common surgical procedure for PD (Benabid 2003; Maks et al 2009; Krack et al 2003), achieving excellent therapeutical outcomes. The main task in DBS, therefore, is the correct targeting of the STN.

Identification of basal ganglia from analysis and classification of Microelectrode Recording Signals (MER) during DBS, serves as a medical support for the correct localization of a target brain area, and the posterior implantation

of neuroexcitatory microelectrodes. Previous works (Chan et al 2010; Chuang et al 2012) have employed processing approaches based on temporal analysis of spikes. In Chan et al (2010), an unsupervised spike sorting method based on specific wavelet coefficients was implemented. The authors use a spike alignment technique based on multi-peak energy comparison (MPEC), and a dynamic codebook-based template-matching algorithm with a class-merging feature. Another common approach is the time-frequency analysis, which consists in transforming the MER signals to different mathematical representation spaces. Examples include the Short-Time Fourier Transform space (STFT) for power spectrum analysis (Chuang et al 2012; Novak et al 2007), the Wavelet Transform space (WT) (Gemmar et al 2008), and the Hilbert-Huang Transform space (HHT) (Pinzon et al 2009). Within the wavelet space, analysis by adaptive filter banks or adaptive wavelets (AW) is one of the most powerful methods for feature extraction in MER signals (Giraldo et al 2008; Pinzon et al 2010). Another type of features used in analysis of MER are based on nonlinear dynamic analysis (Rodriguez et al 2008), and non-stationary surrogate data methods (Guarin et al 2010).

When the emphasis is on analysis, low complexity classifiers are usually employed. Examples are a Linear Discriminant Classifier (LDC) or a Quadratic Discriminant Classifier (QDC) (Pinzon et al 2010). More sophisticated classifiers have also been used including Support Vector Machines (SVM) with Polynomial Kernel (Guillen et al 2011), and Hidden Markov Models (HMM) (Tahgva 2011; Orozco et al 2006).

To the best of our knowledge, all the methods used so far for basal ganglia identification follow the usual supervised learning paradigm. In a nutshell, each microelectrode recording is transformed to a feature space using some signal-processing representation (i.e. STFT, WT and AW). The feature vector thus obtained, $\mathbf{x}$, has an associated label t, assigned by the specialist. In practice, we usually have access to a set of feature vectors $\mathbf{X}$ obtained from the raw MER signal and the corresponding set of labels $\mathbf{t}$. Based on a subset of $\mathbf{X}$ and $\mathbf{t}$, known as a *training set*, a learning algorithm is put to work, with the hope that the algorithm will exhibit an adequate generalization ability over a different subset of $\mathbf{X}$ and $\mathbf{t}$, known as a *validation set*.

We refer to the setup above as *patient-independent* classification. By this, we mean an scenario for which the development of the classification system does not take into account that there are multiple patients involved in a particular study, that is, for training a particular system, data from different patients is usually used.

Machine learning methods have been employed in different real world problems such as rain flow prediction (Taormina and et. al. 2015; Wu and et. al. 2009), financial time series (Niu and Wang 2014), among others. Inspired by the philosophy behind the multi-task learning framework originated in Machine learning (Caruana 1997; Bakker and Heskes 2003; Bonilla et al 2007; Skolidis and Sanguinetti 2011), we propose in this paper a *patient-dependent* classification system for basal ganglia identification. The idea behind multi-task learning is that by learning simultaneously different but related tasks, it is possible to increase the performance of a learning algorithm (Argyriou et al 2008). The augmented performance is explained due to the transfer of information between tasks. For the patient-dependent classification system, we assume that each patient is related to a task. We also refer to this setup as *multi-patient* learning.

In a patient-dependent system, we want to exploit the fact that for different patients, the symptoms of a particular pathology share similar patterns. Given that MER are non stationary and noisy signals, it is very difficult to extract discriminant features for differentiating brain structures. So, the training of an automatic and accurate system for recognition of STN through MER signals is a challenging task. For this reason, traditional approaches in machine learning fail to recognize accurately the STN during DBS. In contrast, by exploiting not only the information encoded in the features extracted from the signals, but also the implicit correlations among patients suffering a similar disease, we aim at increasing the accuracy for targeting the STN. Thereby, we assume there are some underlying "factors" responsible for generating the disease. However, in most of the cases we completely ignore what those factors really are. In spite of this, those underlying factors reflect themselves on the output signals that can be measured for each patient. After learning these hidden factors using a set of signals from different patients, we might be able to use what we learned to increase the predictive accuracy for a patient not previously seen before.

Several algorithms for multi-task learning have been proposed in the machine learning literature. In this work, we employ the multi-task Gaussian processes framework that exhibits state of the art performance in multi-task problems (Bonilla et al 2007; Alvarez and Lawrence 2008; Chai 2009; Pillonetto et al 2010). To obtain the input vectors $\mathbf{X}$ we use three different techniques: Inter Spike Interval (ISI), Wavelet transform using the base function Daubechies 3 (db3) and adaptive wavelets. We show how the multi-patient learning framework improves accuracy when compared to the usual patient-independent setup, in two different datasets.

Multi-task learning was used for MER signals recognition (Vargas et al 2012). For this paper, we have augmented one of the databases with more patients; we have used two additional methods for feature extraction; we have provided more insights, and detailed explanations for the multi-patient learning framework; and we have included a more rigorous experimental evaluation under several experimental conditions.

2 Materials and Methods

In this section, we introduce the datasets that were used for evaluation. We specify the feature extraction methods employed. We then provide a brief description of the patient-independent and patient-dependent classifiers. Finally, we detail the validation framework.

2.1 Databases

A first database comes from Universidad Tecnológica de Pereira (DB-UTP), Colombia. It contains recordings of surgical procedures for six patients with advanced Parkinson's disease, whose ages were in the range 55 ± 6. The patients signed an informed consent form. Microelectrode recordings were obtained using the ISIS MER system (Inomed Medical GmbH).[1] MER signals were labeled by neurophysiology and neurosurgery specialists from the Institute of Parkinson and Epilepsy of the Eje Cafetero, located in the city of Pereira. In total, there are 600 recordings of one second of duration, sampled at 25 KHz with 16-bit resolution. We consider two classes: 300 recordings belong to the Subthalamic Nucleus, and 300 recordings belong to other brain regions (Thalamus-THAL, Zone Incerta-ZI, Substantia Nigra Reticulata-SNR). Figure 1 shows a sagittal view of a DBS performed in STN and Samples from THAL, ZI, STN and SNr. [2]

A second database comes from Universidad Politécnica de Valencia (DB-UPV). Surgeries were carried out in the General University Hospital of Valencia, Spain, and labeled by specialists in neurophysiology and electrophysiology. The equipment used for data acquisition was the LeadPointTM Medtronic (Medtronics Functional Diagnostics).[3] Each signal is one second long, and sampled at 24 KHz. In total, there are 240 recordings coming from four patients: 120 recordings belong to STN and 120 recordings come from other brain regions.

2.2 Feature extraction methods

We use three feature extraction methods, namely, Inter Spike Interval analysis (ISI) (Fu et al 2005), a classical decomposition through a wavelet transform (WT), and a decomposition with adaptive filter banks (AW) (Deslauriers and Dubuc 1987).

[1] http://www.inomed.com

[2] The interested reader can download this dataset from https://dl.dropboxusercontent.com/u/43310202/DB_UTP.rar

[3] http://www.medtronic.com/

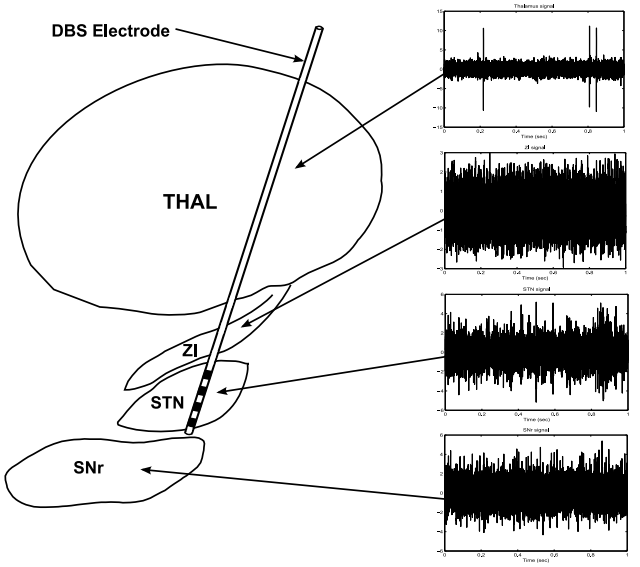

Fig. 1 Cartoon representation of sagittal view for a DBS performed in Subthalamic Nucleus. In this figure, we can see the DBS electrode and samples of MER signals from Thalamus (Thal), Zone Incerta (ZI), Subthalamic Nucleus (STN) and Substantia Nigra pars reticulata (SNr).

2.2.1 Inter Spike Interval (ISI).

The ISI determines the time of occurrence of action potentials for each brain region. This method seeks to organize the electrical activity of the brain according to common patterns between potentials of different areas. In other words, the ISI determines the time of repolarization of brain cells. For this reason, there is a need to isolate each spike in an ISI vector. This procedure is called Spike Sorting (Quiroga et al 2004; Shoham et al 2003). We extract 13 features from the ISI vector: Average Length, Standard Deviation, Maximum length, Minimum length, Mean Instantaneous Frequency (MIF), Standard Deviation of MIF, High Frequency Content Ratio, Low Frequency Content Ratio, dispersion of ISI, dispersion index, Burst Index, asymmetry index and Pause Index. For details, we refer the reader to (Fu et al 2005).

2.2.2 Wavelet Transform (WT).

We apply Wavelet Transform (WT) to the raw MER signals using the mother function Daubechies 3 (db3) with 2 decomposition levels on windows of 80 ms with overlapping of 50%. From the approximation coefficients we calculate the normalized average, the absolute maximum, the kurtosis and energy, obtaining in total eight features (four for each decomposition level).

2.2.3 Adaptive Wavelets (AW)-Dual Scheme.

The adaptive scheme applied to MER signals is shown in Figure 2. X represents the raw MER signal. Signal X_e is obtained by downsampling of X, then X_e is updated with the function U (Update) in order to obtain the approximation coefficients X_L. Signal X_o is obtained by filtering and downsampling X. X_o is then updated with a function P (Prediction) for obtaining the detail coefficients X_H. In this scheme, the prediction function P is adapted while the update function U is fixed. Adaptability is achieved from a decision operator D, which depends on the values obtained from local characteristics of the signal. The decision operator may take two or more values, allowing a choice between two or more filters of dual update. We use a filter with two vanishing moments proposed in (Deslauriers and Dubuc 1987).

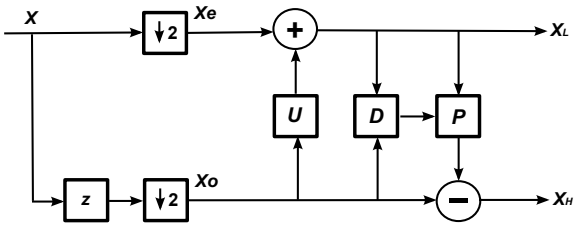

Fig. 2 Dual update adaptive scheme. X_e is obtained by downsampling X. X_e is then updated with the function U (Update) to obtain the approximation coefficients X_L. X_o is obtained by filtering and downsampling X, and then X_o is updated with the function P (Prediction) for obtaining the detail coefficients X_H. D is the decision operator.

The primary update stage is implemented using $X_L[n] = X_e[n]+UX_o[n] = X_e[n]+\frac{1}{2}\left(X_o[n] + X_o[n-1]\right)$, whereas the dual update stage is implemented using $X_H[n] = X_o[n]-P_dX_L[n]$, where P_d may be a filter of second, fourth or sixth order.

We use a dual adaptive scheme to decompose the original signal into two levels. From the approximation coefficients (X_L), we calculate the normalized average, the absolute maximum, the kurtosis and the energy, obtaining 8 features ($\mathbf{x} \in \mathbb{R}^8$) per MER signal. The reader is referred to Giraldo et al (2008) for a detailed description of the above feature extraction method.

2.3 Learning algorithms

We use several standard learning algorithms for classification in the patient-independent context, this is, when no correlation among patients is taken into account. Figure 4 shows a schematic view of patient-independent framework. In the patient-dependent context (multi-patient learning), we use different alternatives of multiple-output Gaussian processes. As we explained before, we consider that patients under the same disease share some underlying factors responsible for generating the disease. Figure 5 shows a schematic view of patient-dependent framework.

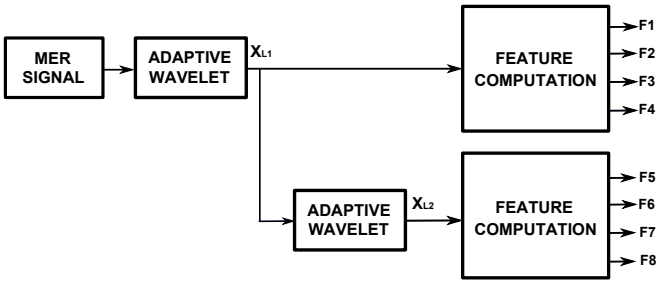

Fig. 3 Feature extraction for MER Signals using Adaptive Wavelets. From approximation coefficients in two decomposition levels ($X_{Lc}, c = 1, 2$), we calculate the normalized average, the absolute maximum, the kurtosis and the energy, obtaining 8 features per sample.

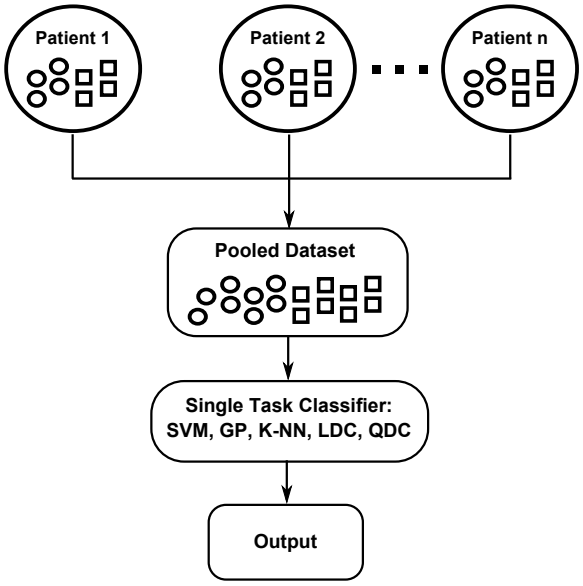

Fig. 4 Patient independent framework. The data is pooled in a single dataset and there is not learning transfer among patients.

2.3.1 Standard classifiers.

We test different parametric and non-parametric classifiers. Within the parametric family, we use the Naive Bayes classifier with a shared covariance matrix among classes, also known as the linear discriminant classifier (LDC) and the Naive Bayes classifier with a different covariance matrix per class, also known as the quadratic discriminant classifier (QDC). Within the non-parametric family, we use the K-nearest neighbors (KNN) algorithm with $K = 1$ and $K = 3$ (KNN1 and KNN3, respectively); a support vector machine with a radial basis kernel (SVM) optimized with quadratic programming. The best parameters for the SVM are found using cross-validation. We also test a Gaussian process re-

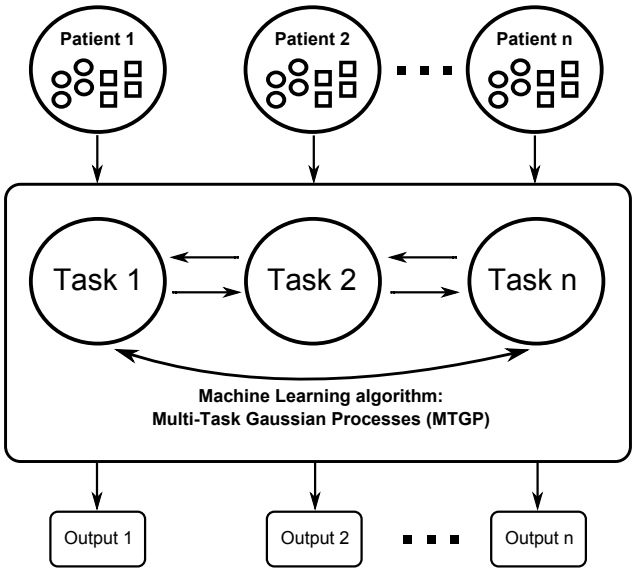

Fig. 5 Patient dependent framework. Each patient is a correlated task. There is learning transfer among patients.

gressor with a RBF (Radial Basis Function) kernel used as a classifier (GPR), and a Gaussian process classifier (GPC) with an ARD (Automatic Relevance Determination) Kernel. For statistical inference in GPR, we use maximum likelihood type-II (Bishop 2006), whereas for statistical inference in GPC we use Laplace approximation (Rasmussen and Williams 2006). The theory behind each of the above classifiers is well known. The interested reader is referred to Bishop (2006).[4] Figure 4 show a scheme of patient independent framework.

2.3.2 Multi-output Gaussian Processes.

Since this a relatively new topic in the machine learning literature, we spend a couple of lines here to describing the different multiple output Gaussian processes methods employed in the experimental section. A detailed description of several alternatives can be found at Alvarez et al (2012).

A general method for multiple output Gaussian processes describes D outputs or tasks $\{f_d(\mathbf{x})\}_{d=1}^{D}$, $\mathbf{x} \in \mathbb{R}^p$, by convolution integrals of latent functions $\{u_q^i(\mathbf{x})\}_{q=1,i=1}^{Q,R_q}$, with smoothing kernels $\{G_{d,q}^i(\mathbf{x}-\mathbf{z})\}_{d=1,q=1,i=1}^{D,Q,R_q}$,

$$f_d(\mathbf{x}) = \sum_{q=1}^{Q}\sum_{i=1}^{R_q}\int G_{d,q}^i(\mathbf{x}-\mathbf{z})u_q^i(\mathbf{z})d\mathbf{z}.$$

Assuming that the latent functions $u_q^i(\mathbf{x})$ are independent Gaussian processes with covariance functions $k_q(\mathbf{x},\mathbf{x}')$,[5] the outputs $f_d(\mathbf{x})$ form a joint Gaussian process with covariance function $k_{d,d'}(\mathbf{x},\mathbf{x}')$ with $d, d' = 1,\ldots,D$, given by

$$\sum_{q=1}^{Q}\sum_{i=1}^{R_q}\int\int G_{d,q}^i(\mathbf{x}-\mathbf{z})G_{d',q}^i(\mathbf{x}'-\mathbf{z}')k_q(\mathbf{z},\mathbf{z}')d\mathbf{z}d\mathbf{z}'. \tag{1}$$

We call this covariance the Convolved Multiple Output Covariance or CMOC.

Assuming that $G_{d,q}^i(\mathbf{x}-\mathbf{z}) = a_{d,q}^i\delta(\mathbf{x}-\mathbf{z})$, being $\delta(\mathbf{x})$ the Dirac delta function, we arrive at a particular case for the covariance function known as the linear model of coregionalization (LMC) (Goovaerts 1997) in the geostatistics literature. The covariance $k_{d,d'}(\mathbf{x},\mathbf{x}')$ reduces then to

$$k_{d,d'}(\mathbf{x},\mathbf{x}') = \sum_{q=1}^{Q}\sum_{i=1}^{R_q}a_{d,q}^i a_{d',q}^i k_q(\mathbf{x},\mathbf{x}') = \sum_{q=1}^{Q}b_{d,d'}^q k_q(\mathbf{x},\mathbf{x}'),$$

where $b_{d,d'}^q = \sum_{i=1}^{R_q}a_{d,q}^i a_{d',q}^i$. The term $b_{d,d'}^q$ accounts for the correlation between the two tasks $f_d(\cdot)$ and $f_{d'}(\cdot)$, for a particular value of q.

A further simplification of the above function, $k_{d,d'}(\mathbf{x},\mathbf{x}')$, can be obtained assuming that $Q = 1$, leading to $k_{d,d'}(\mathbf{x},\mathbf{x}') = b_{d,d'}k(\mathbf{x},\mathbf{x}')$. This model receives the name of the intrinsic coregionalization model (ICM) (Goovaerts 1997).

We assume that the observed tasks $\{y_d(\mathbf{x})\}_{d=1}^{D}$ are given by $y_d(\mathbf{x}) = f_d(\mathbf{x}) + \epsilon_d$, where ϵ_d is a Gaussian white noise following $\epsilon_d \sim \mathcal{N}(0,\sigma_d^2 I)$, being σ_d^2 the variance for the noise. Given a dataset $\mathcal{D} = \{\mathbf{X}_d,\mathbf{y}_d\}_{d=1}^{D}$, where $\mathbf{X}_d = [\mathbf{x}_1^d\cdots\mathbf{x}_{N_d}^d]^\top$, and $\mathbf{y}_d = [y_d(\mathbf{x}_1^d)\cdots y_d(\mathbf{x}_{N_d}^d)]^\top$, the output tasks $\{\mathbf{y}_d\}_{d=1}^{D}$ are jointly Gaussian,

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{0},\mathbf{K}+\boldsymbol{\Sigma}), \tag{2}$$

where $\mathbf{y} = [\mathbf{y}_1^\top\cdots\mathbf{y}_D^\top]^\top$; $\mathbf{K}$ is a block-wise matrix with blocks $\mathbf{K}_{d,d'}$, and $\boldsymbol{\Sigma}$ is a block-diagonal matrix with blocks given by $\sigma_d^2\mathbf{I}$. Elements in $\mathbf{K}_{d,d'}$ are computed using $k_{d,d'}(\mathbf{x}_i^d,\mathbf{x}_j^{d'})$ for $i = 1,\ldots,N_d$, and $j = 1,\ldots,N_{d'}$. The function used for $k_{d,d'}(\cdot,\cdot)$ corresponds to any of the forms CMOC, LMC or ICM.

Predictive distribution for new input data $\{\mathbf{X}_d^*\}_{d=1}^{D}$ is given as

$$p(\mathbf{f}^*|\mathbf{y}) = \mathcal{N}(\mathbf{f}^*|\boldsymbol{\mu}_{\mathbf{f}^*|\mathbf{y}},\mathbf{K}_{\mathbf{f}^*|\mathbf{y}}),$$

where $\mathbf{f}^* = [\mathbf{f}_{1,*}^\top\cdots\mathbf{f}_{D,*}^\top]^\top$, with $\mathbf{f}_{d,*} = [f_d(\mathbf{x}_{*,1}^d)\cdots f_d(x_{*,N_*}^d)]$,[6] and

$$\boldsymbol{\mu}_{\mathbf{f}^*|\mathbf{y}} = \mathbf{K}_*(\mathbf{K}+\boldsymbol{\Sigma})^{-1}\mathbf{y},$$
$$\mathbf{K}_{\mathbf{f}^*|\mathbf{y}} = \mathbf{K}_{*,*} - \mathbf{K}_*(\mathbf{K}+\boldsymbol{\Sigma})^{-1}\mathbf{K}_*^\top,$$

[4] The parametric classifiers, KNN1, KNN3 and the SVM are implemented using the PRTOOLS toolbox obtained from `http://www.prtools.org/`. GPR is implemented using the Gaussian Process Toolbox from `http://staffwww.dcs.shef.ac.uk/people/N.Lawrence/gp/`. GPC is implemented using the Gaussian Process Toolbox from `http://www.gaussianprocess.org/gpml/code/matlab/doc/`.

[5] The latent functions $u_q^i(\mathbf{x})$ share the same covariance $k_q(\mathbf{x},\mathbf{x}')$, irrespectively of the value of i

[6] For simplicity, we assume all the task are evaluated at the same number of tests inputs, N_*. This is mainly to avoid notation clutterness.

where $\mathbf{K}_*$ is a block-wise matrix with blocks $\mathbf{K}_{d*,d}$, and $\mathbf{K}_{*,*}$ is a block-wise matrix with blocks $\mathbf{K}_{d*,d*}$. In turn, matrix $\mathbf{K}_{d*,d}$ has entries $k_{d,d'}(\mathbf{x}^d_{*,i}, \mathbf{x}^{d'}_j)$, for $i = 1, \ldots, N_*$, and $j = 1, \ldots, N_{d'}$. Likewise, matrix $\mathbf{K}_{d*,d*}$ has entries $k_{d,d'}(\mathbf{x}^d_{*,i}, \mathbf{x}^{d'}_{*,j})$, for $i, j = 1, \ldots, N_*$. As before, the function for $k_{d,d'}(\cdot,\cdot)$ corresponds to any of the forms described above for CMOC, LMC or ICM.

For constructing the CMOC in (1), we use Gaussian-like kernels for $G^i_{d,q}(\mathbf{x}-\mathbf{z})$, and $k_q(\mathbf{x},\mathbf{x}')$, leading to a closed form solution for $k_{d,d'}(\mathbf{x},\mathbf{x}')$. For the experimental part, we also set $R_q = 1$. Expression for $G_{d,q}(\mathbf{x}-\mathbf{z})$ follows as

$$G_{d,q}(\mathbf{x}-\mathbf{z}) = \frac{S_{d,q}|\mathbf{P}_d|^{1/2}}{(2\pi)^{p/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{z})^\top \mathbf{P}_d(\mathbf{x}-\mathbf{z})\right],$$

where $S_{d,q}$ is a scale parameter that depends on the input q, and the task d, and $\mathbf{P}_d$ is the precision matrix for task d. For $k_q(\mathbf{z},\mathbf{z}')$, we use

$$k_q(\mathbf{z},\mathbf{z}') = \frac{|\boldsymbol{\Lambda}_q|^{1/2}}{(2\pi)^{p/2}} \exp\left[-\frac{1}{2}(\mathbf{z}-\mathbf{z}')^\top \boldsymbol{\Lambda}_q(\mathbf{z}-\mathbf{z}')\right],$$

where $\boldsymbol{\Lambda}_q$ is the precision matrix for the latent function q.

With the above expressions for $G_{d,q}(\mathbf{x}-\mathbf{z})$, and $k_q(\mathbf{z},\mathbf{z}')$, it can be shown that a standardized form for $k_{d,d'}(\mathbf{x},\mathbf{x}')$, for CMOC, is given as

$$\sum_{q=1}^{Q} S_{d,q}S_{d,q'} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{x}')^\top \mathbf{P}^{-1}_{\text{eqv}}(\mathbf{x}-\mathbf{x}')\right], \qquad (3)$$

where $\mathbf{P}_{\text{eqv}} = \mathbf{P}_d^{-1} + \mathbf{P}_{d'}^{-1} + \boldsymbol{\Lambda}_q^{-1}$.

The details for this construction can be found in Álvarez and Lawrence (2011). [7]

For the LMC covariance, we use a Gaussian kernel for $k_q(\mathbf{x},\mathbf{x}')$,

$$k_q(\mathbf{x},\mathbf{x}') = \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{x}')^\top \boldsymbol{\Gamma}_q(\mathbf{x}-\mathbf{x}')\right],$$

where $\boldsymbol{\Gamma}_q$ is the precision matrix for the latent function q.

For multi-task regression, the usual parameter inference method is based on maximum likelihood. Parameters for CMOC, $\boldsymbol{\theta}_{\text{CMOC}} = \{\{S_{d,q}\}_{d=1,q=1}^{D,Q}, \{\mathbf{P}_d\}_{d=1}^{D}, \{\boldsymbol{\Lambda}_q\}_{q=1}^{Q}\}$, and LMC, $\boldsymbol{\theta}_{\text{LMC}} = \{\{b^q_{d,d'}\}_{d=1,d'=1,q=1}^{D,D,Q}, \{\boldsymbol{\Gamma}_q\}_{q=1}^{Q}\}$, are computed by maximizing the logarithm of $p(\mathbf{y})$ in expression (2), using numerical optimization.

As described above, the CMOC is a generalization of the LMC covariance. While for the CMOC the contribution from the tasks and the inputs is mixed through a convolution operation, this contribution is well separated for the LMC covariance: coefficients $b^q_{d,d'}$ represent the covariance between tasks, for a given q, independently from the input values, while $k_q(\mathbf{x},\mathbf{x}')$ accounts for the covariance between the inputs, independently from the tasks. The practical effect of these different constructions is that the CMOC is generally more flexible than the LMC covariance when it comes to represent outputs with very different behaviors: CMOC would typically need less parameters to describe such variety in the outputs, when compared to the LMC covariance. On the other hand, computing the double integral analitycally for the CMOC is not feasible for all smoothing kernels $G^i_{d,q}(\cdot)$, and covariances $k_q(\cdot,\cdot)$, whereas using the LMC covariance amounts to choosing valid covariance expressions for $k_q(\cdot,\cdot)$.[8] Also, the ICM covariance is the simplest covariance model, and it basically assumes that the difference between the tasks is in their variance, and that their spatial-varying pattern in terms of the inputs is fundamentally the same.

In this paper we use multi-task Gaussian process regression with the CMOC and LMC covariances, for classification purposes. This practice is sometimes known as least-square classification. We refer to the multi-task GP with CMOC as MC and to the multi-task GP with LMC covariance as ML. As mentioned before, for MC, we set $R_q = 1$ for all values of q, and choose Q using cross-validation. For ML, we choose R_q and Q using cross-validation.

We also use the ICM covariance in a multi-task Gaussian process classifier as introduced in Skolidis and Sanguinetti (2011), and refer to this method as MI. [9] In Skolidis and Sanguinetti (2011), the authors use a probit model for relating the observed data $\mathbf{y}$, with the un-observed variables $\mathbf{f}$. Computing the posterior distribution $p(\mathbf{f}|\mathbf{y})$ can not be accomplished in closed form, and the authors use Expectation-Propagation (EP) for computing an approximated posterior.

2.4 Validation

To test the statistical significance of our results, we follow the procedure proposed for model selection in Pizarro et al (2002). We split each dataset in a training set and a validation set. We train the different methods using the training set and then we measure the accuracy and area under the curve (AUC) over the validation set. We repeat this procedure 30 times with a different training set and validation set per repetition. To study if there are differences that are statistically significant among the classifiers, we apply first a

[7] The exact expression for (3) includes an additional scaling factor that was not included, since for a high value of p, that scaling factor makes the kernel goes to zero quickly. A detailed mathematical explanation of this phenomenon is given in Álvarez and Lawrence (2011).

[8] Strictly speaking, we also need the coefficients $b^q_{d,d'}$ to lead to a positive semidefinite function for $k_{d,d'}(\cdot,\cdot)$. This can be enforced by using $b^q_{d,d'} = \sum_{i=1}^{R_q} a^i_{d,q}a^i_{d',q}$, and estimating $a^i_{d,q}$ instead of $b^q_{d,d'}$.

[9] We implement MC and ML using the MULTIGP Toolbox retrieved from `http://staffwww.dcs.shef.ac.uk/people/N.Lawrence/multigp/`. We implement MI using software available at `http://homepages.inf.ed.ac.uk/gsanguin/software.html`.

Lilliefors test for normality over the 30 repetitions of each classifier. If the null hypothesis for normality is rejected, we perform a Kruskal-Wallis test to compare average performances among the classifiers. If the null hypothesis for equal average behaviors is rejected, we perform a multiple comparison test using Tukey-Kramer and Bonferroni to study further, which classifiers are different. All the significance levels are measured at 5%.

2.5 Experimental setup.

We want to evaluate the generalization ability of the patient-independent classifiers (KNN1, KNN3, LDC, QDC, SVM, GPR, and GPC) against the patient-dependent classifiers (MI, MC, and ML) in three different types of experiments. For the first type of experiment, we test the performance of the different classifiers using 50% of the datapoints from each patient for training, and then validate the performance over the other 50% of the datapoints per patient. We refer to this type of experiment as E1. For the second and third types of experiments, we initially perform leave-one out (LOO) over the patients. This is, assume we have D patients for each dataset. We split the patients in two sets: one of the sets has $D-1$ patients, and the other set has only the patient left out. We call the set with $D-1$ patients, the *training set of patients* (TSP), and we call the set with the left out patient, the *validation set of patients* (VSP). In the second type of experiments, we generate the training set for the classifiers with 50% of the datapoints from the TSP, plus 10% of datapoints from the VSP. The performance measures are computer over the 90% of the datapoints from the VSP. We then change the composition of the patients for both sets, the TSP and the VSP, according to the LOO methodology, and compute once again the performance measures. We report the averaged performance measures obtained from having the different patients in the VSP, one at a time. We refer to this type of experiment as E2. The third type of experiments are similar to the second type. The only difference is that the training set for the classifiers is made-up with 50% of the datapoints from the TSP, and none of the datapoints from the VSP. The performance measures are computer over the 100% of the datapoints from the VSP. The reported measures are averaged measures obtained similarly to the ones obtained in E2. We refer to the third type of experiments as E3.

Notice that E3 is an extreme experiment in which we do not want to use any datapoints from the validation patient (from the set VSP) when training the classifiers, as opposed to E2. Actually, we are attempting to identify the STN of a particular patient without using any information from that patient. This setup does not quite fit within the multi-patient learning classifiers described in section 2.3.2, since for the prediction phase we need the estimated values for the parameters $b^q_{d,d'}$, and the estimated values for the parameters associated to the kernels $G^i_{d,q}(\cdot)$, for all values of d, even for the left-out patient. To estimate those parameters, we basically need datapoints related to the left-out patient. To fulfill these requirements, and to avoid the inclusion of any datapoint from the left-out patient (in the training phase), we use as "surrogate" data for that patient, the feature vectors computed as the average of the feature vectors of all the patients in TSP, with their corresponding labels (STN, etc.) for those feature vectors.

3 Results and Discussion

In this section, we report and discuss results of the comparison between the patient-independent classifiers against the patient dependent classifiers. Performance measures are computed over two datasets (DB-UTP and DB-UPV), and under three different experimental setups, namely, experiments E1, E2, and E3, as described in section 2.5.

3.1 Results for DB-UTP

Table 1 shows accuracy and AUC results for E1, E2 and E3 in DB-UTP.

With respect to E1, we notice that when the features are extracted with the wavelet transform (WT) and adaptive wavelet (AW), the methods employing multi-patient learning (MI, MC, and ML) exhibit better performance than methods disregarding correlations between patients (KNN1, KNN3, LDC, QDC, SVM, GPR and GPC). This increased performance is further tested using the hypothesis tests described in section 2.4. For data processed using ISI, accuracy and AUC results are similar in all classifiers. The null hypothesis of equal average behavior between the group of multi-patient learning algorithms and the group of patient-independent algorithms is rejected when we extract features with WT and AW-LS, but it is not rejected when we use ISI. According to the same analysis, the difference in performances between MI, MC and ML is not statistically significant for all processing methods.

With respect to E2. All multi-patient algorithms (MC, MI, ML) show an increased performance when compared to the other learning algorithms. The increased performance is observed for all the feature extraction method employed. The null hypothesis of equal averages between the group of multi-patient classifiers and the group of standard algorithms is rejected. Recall from section 2.5 that in E2, we use 50% of datapoints for each patient in TSP, and 10% of the datapoints for the validation patient in VSP. Performance measures are computed over the remaining 90% datapoints

Table 1 Mean accuracy and area under the ROC for different classifiers applied to DB-UTP featured with adaptive wavelet, wavelet transform and inter spike interval (ISI) for all experiments. KNNX stands for K-nearest neighbors, where X is either 1 or 3. L(Q)DC stands for linear(quadratic) discriminant classifier. SVM stands for support vector machine. GPR stands for Gaussian Process Regressor. GPC stands for Gaussian Process Classifier. MI represents a multi-patient GP classifier with ICM covariance. MC represents a multi-patient GP regressor with CMOC. ML represents a multi-patient GP regressor with LMC covariance.

		ISI		Wavelet Transform		Adaptive Wavelet	
Experiment	Method	Accuracy (%)	AUC	Accuracy (%)	AUC	Accuracy (%)	AUC
E1	KNN1	79,1 ± 1,8	0,844 ± 0,013	90,2 ± 0,7	0,970 ± 0,003	92,1 ± 0,8	0,973 ± 0,004
	KNN3	81,5 ± 2,3	0,826 ± 0,018	90,0 ± 1,1	0,922 ± 0,010	92,0 ± 0,9	0,938 ± 0,010
	LDC	78,7 ± 1,5	0,460 ± 0,041	71,8 ± 1,3	0,628 ± 0,011	76,1 ± 1,6	0,656 ± 0,013
	QDC	$\mathbf{82,6 \pm 2,4}$	0,689 ± 0,030	75,4 ± 0,8	0,703 ± 0,012	84,2 ± 1,0	0,779 ± 0,010
	SVM	80,3 ± 2,4	0,868 ± 0,014	92,0 ± 0,9	0,962 ± 0,006	93,0 ± 1,0	0,973 ± 0,006
	GPR	80,2 ± 1,8	0,911 ± 0,015	92,2 ± 0,5	0,968 ± 0,006	92,9 ± 0,9	0,972 ± 0,004
	GPC	80,2 ± 1,8	$\mathbf{0,911 \pm 0,015}$	90,8 ± 1,0	0,959 ± 0,004	92,3 ± 1,1	0,958 ± 0,008
	MI	79,1 ± 2,0	0,872 ± 0,026	95,0 ± 0,6	$\mathbf{0,985 \pm 0,004}$	$\mathbf{97,0 \pm 0,6}$	$\mathbf{0,992 \pm 0,003}$
	MC	77,3 ± 0,8	0,815 ± 0,011	$\mathbf{95,1 \pm 0,5}$	0,983 ± 0,003	96,4 ± 0,6	0,988 ± 0,004
	ML	79,2 ± 2,3	0,867 ± 0,027	93,4 ± 1,9	0,968 ± 0,012	96,3 ± 0,7	0,986 ± 0,005
E2	KNN1	59,4 ± 1,1	0,557 ± 0,011	83,3 ± 1,2	0,935 ± 0,006	82,5 ± 1,6	0,912 ± 0,010
	KNN3	60,5 ± 1,0	0,517 ± 0,013	80,8 ± 1,2	0,808 ± 0,016	78,7 ± 1,3	0,781 ± 0,024
	LDC	63,2 ± 1,1	0,519 ± 0,021	66,0 ± 0,7	0,562 ± 0,011	63,2 ± 1,2	0,532 ± 0,008
	QDC	63,7 ± 0,9	0,561 ± 0,019	76,0 ± 1,2	0,655 ± 0,014	73,6 ± 1,0	0,566 ± 0,008
	SVM	64,6 ± 1,1	0,689 ± 0,011	84,4 ± 1,2	0,922 ± 0,009	83,1 ± 1,2	0,887 ± 0,018
	GPR	63,3 ± 0,9	0,679 ± 0,010	84,4 ± 1,2	0,924 ± 0,008	83,1 ± 1,6	0,900 ± 0,014
	GPC	56,6 ± 2,8	0,593 ± 0,028	72,8 ± 2,7	0,769 ± 0,028	72,8 ± 2,7	0,769 ± 0,028
	MI	65,9 ± 4,8	0,775 ± 0,025	87,8 ± 1,4	0,930 ± 0,013	$\mathbf{90,8 \pm 1,5}$	0,943 ± 0,012
	MC	$\mathbf{81,6 \pm 1,8}$	$\mathbf{0,890 \pm 0,022}$	$\mathbf{89,8 \pm 0,9}$	$\mathbf{0,956 \pm 0,007}$	90,7 ± 1,1	$\mathbf{0,954 \pm 0,005}$
	ML	58,7 ± 4,1	0,680 ± 0,020	87,5 ± 1,7	0,930 ± 0,010	87,3 ± 1,9	0,910 ± 0,017
E3	KNN1	57,3 ± 6,1	0,540 ± 0,015	72,57 ± 1,97	0,826 ± 0,031	71,38 ± 5,68	0,88 ± 0,021
	KNN3	58,2 ± 2,5	0,520 ± 0,021	70,36 ± 2,25	0,647 ± 0,037	74,72 ± 3,21	0,77 ± 0,030
	LDC	61,3 ± 4,7	0,530 ± 0,019	60,12 ± 2,81	0,558 ± 0,02	59,18 ± 4,46	0,55 ± 0,005
	QDC	62,1 ± 5,9	0,650 ± 0,041	68,77 ± 1,94	0,634 ± 0,046	68,41 ± 5,52	0,57 ± 0,012
	SVM	62,8 ± 4,3	0,670 ± 0,014	71,42 ± 2,87	0,89 ± 0,011	83,18 ± 3,68	0,9 ± 0,018
	GPR	61,9 ± 2,9	0,670 ± 0,014	76,78 ± 1,62	0,825 ± 0,027	73,8 ± 4,77	0,91 ± 0,014
	GPC	55,7 ± 3,2	0,570 ± 0,018	74,12 ± 2,39	0,821 ± 0,033	76,51 ± 3,64	0,78 ± 0,039
	MI	65,3 ± 2,1	0,740 ± 0,028	81,18 ± 2,23	0,871 ± 0,017	$\mathbf{86,66 \pm 5,25}$	0,93 ± 0,007
	MC	$\mathbf{79,6 \pm 1,2}$	$\mathbf{0,860 \pm 0,024}$	$\mathbf{85,99 \pm 1,15}$	$\mathbf{0,919 \pm 0,015}$	85,28 ± 2,96	$\mathbf{0,940 \pm 0,004}$
	ML	57,9 ± 2,9	0,670 ± 0,023	83,22 ± 2,51	0,891 ± 0,019	75,64 ± 8,1	0,75 ± 0,050

for the patient in VSP. We report in Table 1, the averaged performance over the six patients.

With respect to E3, we observe that the multi-patient learning algorithms clearly outperform the standard learning algorithms. The null hypothesis tests further confirm this result. This result is highly relevant, because it shows evidence that the multi-patient methods have an increased ability for generalization, when compared to the patient-independent methods.

It can be seen from Table 1 that the difference of performance of the patient-dependent methods over the patient-independent methods, is larger for experiments E2, and E3 than for experiment E1. In the context of multi-patient learning, this is a sensible result since in E2 and E3, we use a just a few or none datapoints from the patient in the VSP. The increased performance is also more pronounced for ISI and WT, than for AW.

3.2 Results for DB-UPV

Table 2 shows accuracy and AUC results for E1, E2 and E3 in DB-UPV.

We notice for E1 that the mean accuracy and AUC performances for the multi-patient algorithms (MI, MC and ML) are superior to the mean performances of the standard classifiers, except for the ISI feature extraction method. Based on the multiple comparison test, we conclude that MI, MC and ML are not different statistically speaking when compared to SVM. Nevertheless, the post test analysis rejects the null hypothesis of equal average behaviors between MC and SVM and between ML and SVM. The post test analysis also rejects the null hypothesis of equal average behaviors between the multi-patient learning algorithms, and the other standard classifiers (KNN1, KNN3, LDC, QDC, GPR and GPC).

MI with ISI features exhibits the best results for the E2 experiment. MC is the best when using WT. For data processed with AW, the SVM obtains the best identification result. The null hypothesis of equal average behaviors between

Table 2 Mean accuracy and area under ROC for different classifiers applied to DB-UPV featured with adaptive wavelet, wavelet transform and inter spike interval (ISI) for all experiments. The specification for each classifier is given in section 2.3.1 or in the caption of table 1.

		ISI		Wavelet Transform		Adaptive Wavelet	
Experiment	Method	Accuracy (%)	AUC	Accuracy (%)	AUC	Accuracy (%)	AUC
E1	KNN1	75,9 ± 2,1	0,825 ± 0,041	92,7 ± 1,3	0,983 ± 0,007	89,3 ± 1,5	0,957 ± 0,010
	KNN3	80,4 ± 2,1	0,765 ± 0,032	90,2 ± 1,7	0,928 ± 0,020	86,8 ± 1,6	0,875 ± 0,020
	LDC	76,2 ± 2,3	0,775 ± 0,026	87,2 ± 1,3	0,864 ± 0,018	85,8 ± 1,9	0,770 ± 0,025
	QDC	76,7 ± 2,4	0,789 ± 0,019	89,1 ± 1,8	0,903 ± 0,014	86,9 ± 2,5	0,858 ± 0,019
	SVM	$\mathbf{80,7 \pm 1,7}$	$\mathbf{0,842 \pm 0,016}$	93,0 ± 1,7	0,971 ± 0,009	94,4 ± 1,3	0,976 ± 0,008
	GPR	80,1 ± 2,3	0,831 ± 0,018	91,2 ± 1,2	0,970 ± 0,012	89,5 ± 1,4	0,956 ± 0,007
	GPC	80,1 ± 2,3	0,831 ± 0,018	88,1 ± 1,0	0,945 ± 0,009	90,4 ± 1,4	0,904 ± 0,014
	MI	77,3 ± 3,5	0,813 ± 0,023	93,1 ± 1,0	0,975 ± 0,009	95,1 ± 1,2	$\mathbf{0,989 \pm 0,007}$
	MC	76,4 ± 3,3	0,786 ± 0,029	$\mathbf{96,4 \pm 1,0}$	$\mathbf{0,995 \pm 0,003}$	94,6 ± 0,9	0,983 ± 0,007
	ML	78,7 ± 3,6	0,823 ± 0,031	96,1 ± 1,2	0,990 ± 0,006	$\mathbf{95,4 \pm 1,4}$	0,986 ± 0,007
E2	KNN1	74,1 ± 1,8	0,805 ± 0,041	82,5 ± 1,6	0,909 ± 0,018	80,2 ± 2,1	0,900 ± 0,019
	KNN3	74,5 ± 1,9	0,695 ± 0,032	81,7 ± 1,7	0,785 ± 0,018	78,0 ± 1,4	0,717 ± 0,020
	LDC	71,2 ± 1,5	0,693 ± 0,026	82,7 ± 1,3	0,826 ± 0,014	82,0 ± 1,1	0,813 ± 0,015
	QDC	74,7 ± 1,5	0,789 ± 0,019	81,0 ± 2,0	0,847 ± 0,011	79,7 ± 1,5	0,812 ± 0,015
	SVM	76,3 ± 1,2	0,805 ± 0,016	86,6 ± 1,0	0,945 ± 0,005	$\mathbf{86,4 \pm 1,1}$	$\mathbf{0,950 \pm 0,004}$
	GPR	74,5 ± 1,9	0,788 ± 0,019	82,6 ± 2,3	0,905 ± 0,021	81,7 ± 1,8	0,905 ± 0,022
	GPC	74,1 ± 1,8	0,799 ± 0,022	81,5 ± 2,2	0,894 ± 0,020	79,8 ± 2,1	0,905 ± 0,022
	MI	$\mathbf{77,1 \pm 1,8}$	$\mathbf{0,819 \pm 0,023}$	82,0 ± 2,5	0,893 ± 0,024	82,4 ± 2,9	0,910 ± 0,023
	MC	74,3 ± 2,3	0,815 ± 0,029	$\mathbf{87,3 \pm 2,2}$	$\mathbf{0,947 \pm 0,012}$	83,3 ± 2,2	0,928 ± 0,015
	ML	73,3 ± 7,3	0,741 ± 0,031	78,6 ± 3,5	0,851 ± 0,033	82,9 ± 2,2	0,899 ± 0,017
E3	KNN1	71,2 ± 4,1	0,800 ± 0,015	76,3 ± 1,97	0,840 ± 0,031	77,3 ± 4,2	0,860 ± 0,021
	KNN3	70,6 ± 2,4	0,740 ± 0,021	75,4 ± 2,25	0,770 ± 0,037	74,1 ± 3,5	0,630 ± 0,030
	LDC	71,1 ± 4,2	0,770 ± 0,019	74,9 ± 2,81	0,770 ± 0,02	76,4 ± 2,9	0,800 ± 0,005
	QDC	71,8 ± 3,1	0,760 ± 0,041	72,1 ± 1,94	0,810 ± 0,046	79,9 ± 4,3	0,770 ± 0,012
	SVM	72,2 ± 2,3	0,810 ± 0,017	77,6 ± 2,87	0,920 ± 0,011	80,2 ± 3,4	0,886 ± 0,018
	GPR	72,5 ± 1,9	0,800 ± 0,017	73,1 ± 1,62	0,880 ± 0,027	76,4 ± 2,9	0,810 ± 0,014
	GPC	71,8 ± 2,2	0,790 ± 0,014	75,3 ± 2,39	0,870 ± 0,033	74,7 ± 3,1	0,760 ± 0,039
	MI	$\mathbf{76,9 \pm 2,1}$	$\mathbf{0,818 \pm 0,018}$	79,1 ± 2,23	0,890 ± 0,017	81,2 ± 3,9	0,880 ± 0,007
	MC	72,2 ± 2,9	0,790 ± 0,024	$\mathbf{84,1 \pm 1,15}$	$\mathbf{0,930 \pm 0,015}$	$\mathbf{82,5 \pm 2,1}$	$\mathbf{0,890 \pm 0,004}$
	ML	71,8 ± 7,4	0,819 ± 0,023	77,7 ± 2,51	0,810 ± 0,019	82,1 ± 1,9	0,840 ± 0,050

multi-patient classifiers and SVM is not rejected. Therefore, differences between SVM and MI, MC and ML are not statistically significant. Nevertheless, the multi-patient group of classifiers is superior to the other standard methods.

For E3 and ISI features, MI is superior to the other classifiers. When the features are WT and AW, the MC method is better. The null hypothesis test shows that both MI and MC are statistically different when compared to the standard methods.

The general landscape of the results show a similar behavior to the one obtained for DB-UTP. From Table 2 we observe that the difference of performance of the patient-dependent methods over the patient-independent methods, is larger for experiments E2, and E3, and ISI and WT features.

3.3 Discussion

Several approaches have been evaluated for MER signals identification, especially to recognize the STN. In some of these works the contribution is focused in the processing methodology, and other works emphasize in supervised learning techniques, for example: Bayesian classifier, hidden Markov Models, support vector machines (SVM), single Gaussian processes (GP), etc. The mentioned frameworks pool the data from different subjects in a single set, and the classifier is trained with the classical paradigm. The SVM is considered the state of the art for recognition of MER signals, for this reason we compare our proposed approach with this method. In relation to multi-task learning (MTL), there are alternatives such as MTL-KNN, MTL with kernels, decision trees, among others. However those methods are developed for a specific type of data. The Gaussian processes are powerful, flexible and robust to noisy signals and they are not sensible to non-stationary data. Another advantage is that a GP is a kernelized method, for this reason is able to discriminate high dimensional patterns such as processed MER signals. Multi-task Gaussian processes (MTGP) exhibit the state of the art in MTL techniques. Following this notion, other alternative could not achieve the excellent outcomes obtained with MTGP.

Although, there are feasible alternatives for extracting features from MER signals: Spike detection, inter spike interval (ISI), short time Fourier transform (STFT), discrete Wavelet transform (WT), Hilbert-Huang transform, Adaptive Wavelet (AW), statistical descriptors in time, power spec-

trum analysis, non linear dynamic analysis, among others. We adopted ISI, WT and AW because we pretend to analyze the capability of MTGP to identify the STN when we extract the features with a weak method (ISI), an intermediate method (WT), and an advanced method (AW). In this work, the feature extraction approach is secondary. The advantage of using three processing methods is to prove that our proposed recognition methodology outperforms to the classical paradigm of single task learning, under any condition. Given that MTGP improves to LDC, QDC, KNN and SVM for STN identification, no matter the database, processing method or experimental setup. We consider that a different feature extraction method will not affect the results.

To the best of our knowledge, there are not public MER signals databases. The works presented previously in literature always reported results with private datasets. For testing the sensibility of our method to modifications in data, we repeat the experiments in two databases acquired with different sampling frequencies. Our approach showed a better performance when we evaluate both databases, in different testing scenarios (E1, E2, E3). The advantage of our databases is the medical validation of experimented specialists from the Institute of Epilepsy and Parkinson of the eje Cafetero (Colombia) and the Universitat Politecnica de Valencia. Also, DB-UTP and DB-UPV has been tested in similar works previously published. According to this analysis, a different dataset will not affect the results, because the experimental setup in this work consider the worst scenario in E3.

We experimented with several set of parameters for the Support Vector Machine (SVM), Gaussian Process (GP) and Multi-Task GP (MTGP) with LMC and CMOC covariance. K-NN is a non-parametric method. Specifically for the SVM, we used the Radial Basis Function (RBF) kernel. The regularization parameter (γ) for the SVM, is estimated using cross-validation over a grid of values for this parameter. The length-scale hyperparameter of the Gaussian Process (GP) is optimized using the gradient descent algorithm. Also, for the GP and MTGP, we evaluated the number of latent functions (LF) and we found the best results for both methods by setting one LF. The sensitivities of these parameters on the results are high. Specially, the regularization parameter in SVM. For this reason it is necessary a detailed search of the best γ with cross validation. For the GP and MTGP (LMC and CMOC) the number of LF is relevant. However, it is easy to find the best number of LF by a simple comparison of the best results.

As we explained before, E3 is the more extreme scenario for testing the proposed method. While, we do not use datapoints from validation patient, we employ a kind of surrogate data for that patient, defined as the average of the feature vectors of all the patients used in the training phase. This solution worked well for estimating parameters in the left-out patient, because multi-task methods can correlate information among tasks. This additional information obtained from the patients allows to achieve better outcomes compared to traditional approaches for classification, where a correlation is not possible.

4 Conclusions

In this work, we presented a methodology based on multi-task learning for classification of MER signals obtained from Parkinson' disease surgeries. We introduced the term multi-patient learning to refer to this methodology. Results for the multi-patient learning methods outperformed the standard paradigm of supervised classification used in the state of the art for MER signals identification. The increased performance is retained across several feature extraction methods, and in two different datasets.

We also showed that methods using multi-patient learning increase the accuracy over standard learning techniques in setups of low amounts of data for validation patients. Experiment three is particularly illustrative of the way in which multi-patient learning can leverage the performance of the learning system, when none of the datapoints are available for a particular validation patient.

This study is limited by a validation with patients of the same database. When we trained the methods (LDC, QDC, K-NN, SVM, GP, MTGP) with subjects of DB-UTP and we validated with data from DB-UPV (or we trained with DB-UPV and we validated with DB-UTP) the accuracy in STN recognition was reduced considerably. This issue is due to acquisition characteristics, especially the sampling frequency. In this case, DB-UTP was sampled to 25 KHz and DB-UPV was sampled to 24 KHz. The performance of our proposed approach depends of a validation with data from the same database. Therefore, the MTGP is sensible to acquisition protocols of MER signals.

We consider this methodology can be improved with augmented databases. We observed that multi-patient framework can improve the STN identification when we have more subjects in the databases. The correlation among patients with the same pathology is a key information that can not be omitted.

Finally, as future work we would like to evaluate the multi-patient methodology for other problems associated to biosignals identification. For example: epileptic sources (EEG), cardiac anomalies (ECG) and emotion recognition (multimodal signals).

Acknowledgements We would like to thank the MD Hans Carmona Villada and the Institute of Epilepsy and Parkinson of Eje Cafetero, who helped in organizing the database DB-UTP. We also acknowledge Dr. Enrique Guijarro, for providing us with the DB-UPV database. This research has been developed under the project with code 111056934461

financed by Colciencias. H.D Vargas Cardona author is funded by Colciencias under the program: *formación de alto nivel para la ciencia, la tecnología y la innovación - Convocatoria 617 de 2013*.

References

Alvarez M, Lawrence N (2008) Sparse convolved Gaussian process for multi-output regression. In: Advances in Neural Information Processing Systems (NIPS), pp 57–64

Alvarez M, Rosasco L, Lawrence N (2012) Kernels for vector-valued functions: A review. Foundations and Trends in Machine Learning 4:195–266

Álvarez MA, Lawrence ND (2011) Computationally efficient convolved multiple output Gaussian processes. Journal of Machine Learning Research 12:1459–1500

Argyriou A, Evgeniou T, Pontil M (2008) Convex multi-task feature learning. Machine Learning 73(3):243–272

Bakker B, Heskes T (2003) Task clustering and gating for Bayesian multitask learning. The Journal of Machine Learning Research 4:83–99

Benabid A (2003) Deep brain stimulation for Parkinson's disease. Curr Opin Neurobiol 13:696–706

Bishop C (2006) Pattern Recognition and Machine Learning. Springer, USA

Bonilla EV, Agakov FV, Williams CKI (2007) Kernel multi-task learning using task-specific features. In: Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS)

Caruana R (1997) Multitask learning. Machine Learning 28:41–75

Chai KM (2009) Generalization errors and learning curves for regression with multi-task Gaussian processes. In: Advances in Neural Information Processing Systems 22 (NIPS), pp 279–287

Chan H, Wu T, Lee S, Lin M, He S, Chao P, Tsai Y (2010) Unsupervised wavelet-based spike sorting with dynamic codebook searching and replenishment. Neurocomputing 73, no.7-9:1513–1527

Chuang WY, Young KY, Chao CP, Tsai ST, Chen SY (2012) Locating optimal electrodes placement via microelectrode recording in general anesthetic patients during deep brain stimulation. In: IEEE International Conference on Biomedical Engineering and Biotechnology, pp 1048–1051

Deslauriers G, Dubuc S (1987) Interpolation dyadique. In: Fractals, dimensions non entires et applications, p 44–55

Fu Q, Clements M, Mewes K (2005) Neural cell type recognition between globus pallidus externus and globus pallidus internus by gaussian mixture modelling. In: ICASSP, p 53–56

Gemmar P, Gronz O, Henrichs T, Hertel F (2008) Advanced methods for target navigation using microelectrode recordings in stereotactic neurosurgery for deep brain stimulation. In: CBMS ?08: Proceedings of the 21st IEEE International Symposium on Computer-Based Medical Systems, IEEE Computer Society, vol 21, p 99–104

Giraldo E, Castellanos G, Orozco A (2008) Feature extraction for MER signals using adaptive filter banks. In: Electronics, Robotics and Automotive Mechanics Conference, pp 582–585

Goovaerts P (1997) Geostatistics For Natural Resources Evaluation. Oxford University Press, USA

Guarin D, Orozco A, Delgado E, Guijarro E (2010) On detecting determinism and nonlinearity in microelectrode recording signals: Approach based on non-stationary surrogate data methods. In: EMBC 2010, Proceedings of the 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp 4032–4035

Guillen P, Martinez F, Sanchez R, Argaez M, Velasquez L (2011) Characterization of subcortical structures during deep brain stimulation utilizing support vector machines. In: EMBC 2011, Proceedings of the 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp 7949–7952

Krack P, Batir A, Van Blercom N, Chabardes S, Fraix V, Ardouin C, Koudsie A, Limousin P, Benazzouz A, LeBas J, Benabid A, Pollak P (2003) Five-year follow-up of bilateral stimulation of the subthalamic nucleus in advanced Parkinson's disease. N Engl J Med 349(20):1925–1934

Lees A, Hardy J, Revesz T (2009) Parkinson's disease. Lancet 2009 373:2055–2066

Maks C, Butson C, Walter B, Vitek J, McIntyre C (2009) Deep brain stimulation activation volumes and their association with neurophysiological mapping and therapeutic outcomes. J Neurol Neurosurg Psychiatry 80(6):659–666

Miocinovic S, Lempka S, Russo G, Maks C, Butson C, Sakaie K, Vitek J, McIntyre C (2009) Experimental and theoretical characterization of the voltage distribution generated by deep brain stimulation. Exp Neurol 216(1):166–176

NINDS (2004) Parkinson's disease: Challenges, progress, and promise. National Institute of Health pp 6–10

Niu H, Wang J (2014) Financial time series prediction by a random data-time effective rbf neural network. J Soft computing 18:497–508

Novak P, Daniluk S, Elias S, Nazzaro J (2007) Detection of the subthalamic nucleus in microelectrographic recordings in Parkinson disease using the high frequency (500 hz) neuronal background. Neurosurgey 106:175–179

Orozco A, Álvarez M, Guijarro E, Castellanos G (2006) Identification of spike sources using proximity analysis through hidden Markov models. In: EMBC 2006, Proceedings of the 28th annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol 28, pp 5555–5558

Pillonetto G, Dinuzzo F, De Nicolao G (2010) Bayesian online multitask learning of Gaussian processes. IEEE Transactions on Pattern Analysis Machine Intelligence 32(2):193–205

Pinzon R, Garces M, Orozco A, Nazzar J (2009) Automatic identification of various nuclei in the basal ganglia for Parkinson's disease neurosurgery. In: EMBC 2009, Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol 31, p 99–104

Pinzon R, Orozco A, Carmona H, Castellanos G (2010) Feature selection using an ensemble of optimal wavelet packet and learning machine: Application to MER signals. In: Communication Systems Networks and Digital Signal Processing (CSNDSP), vol 7, pp 25–30

Pizarro J, Guerrero E, Galindo P (2002) Multiple comparison procedures applied to model selection. Neurocomputing 48:155–173

Quiroga R, Ndasdy Z, Ben-Shaul Y (2004) Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. Neural Computation 16:1661–1687

Rasmussen C, Williams C (2006) Gaussian Processes for Machine Learning. MIT Press

Rodriguez A, Delgado E, Orozco A, Castellanos G, Guijarro E (2008) Nonlinear dynamics techniques for the detection of the brain areas using MER signals. In: IEEE International Conference on BioMedical Engineering and Informatics, vol 2, pp 198–202

Shoham S, Fellows M, Normann R (2003) Automatic spike sorting using mixtures of multivariate t-distributions. Journal of Neuroscience Methods 127:111–122

Skolidis G, Sanguinetti G (2011) Bayesian multi-task classification with Gaussian process priors. IEEE Transactions on Neural Networks 22:2011–2021

Tahgva A (2011) Hidden semi-Markov models in the computerized decoding of microelectrode recording data for deep brain stimulator placement. World Neurosurgery 75:758–763

Taormina R, et al (2015) Data-driven input variable selection for rainfall-run modeling using binary-coded particle swarm opti-

mzation and extreme learning machines. Journal of Hidrology 529:1617–1632

Vargas H, Orozco A, Alvarez M (2012) Multi-patient learning increases accuracy for Subthalamic nucleus identification in deep brain stimulation. In: EMBC 2012, Proceedings of the 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp 4341–4344

Wright D, Nakamura K, Maeda T, Kutsuzawa K, Miyawaki K, Nagata K (2008) Research and development of a portable device to quantify muscle tone in patients with Parkinsons disease. In: EMBC 2008, Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol 30, pp 2825–2827

Wu C, et al (2009) Methods to improve neural network performance in daily flows prediction. Journal of Hydrology 372:80–93