UNIVERSIDAD NACIONAL DE COLOMBIA

Time-series representation framework based on multi-instance similarity measures

Marco de representación de series de tiempo basado en medidas de similitud de múltiples instancias

Julián Camilo Caicedo Acosta

Universidad Nacional de Colombia
Faculty of Engineering and Architecture
Departement of Electrics, Electronics and Computation Engineering
Manizales, Colombia
2019

Time-series representation framework based on multi-instance similarity measures

Julián Camilo Caicedo Acosta

Thesis submitted as a partial requirement to receive the grade of:
Magister en Ingeniería - Automatización Industrial

Advisor:
Ph.D. Germán Castellanos Domínguez
Co-Advisor:
Ph.D. David Cárdenas Peña

Academic Research Group:
Signal Processing and Recognition

Universidad Nacional de Colombia
Faculty of Engineering and Architecture
Departement of Electrics, Electronics and Computation Engineering
Manizales, Colombia
2019

"Your work is going to fill a large part of your life, and the only way to be truly satisfied is to do what you believe is great work. And the only way to do great work is to love what you do. If you haven't found it yet, keep looking. Don't settle. As with all matters of the heart, you'll know when you find it."

-Steve Jobs-

Acknowledgment

To my mother for being the pillar of my life. To my grandparents, my aunt and my cousin for unconditional support. To the woman who fills my days with light, my angel, Natalia. I love them. To Professor Germán Castellanos for the dedication and all the knowledge provided. To David Cárdenas, for being a guide and inspiration. To Collazos, Tobar, Guillo, Felipe, Anderson, Luisa, Frank, and other companions of the signal processing group. To my friends, colleagues, and acquaintances who contributed a grain of sand in this process. To all of them, Thank you.

This research would not have been possible without the financial assistance of the research project “Desarrollo de un sistema integrado de monitoreo de actividad cerebral a partir de registros electroencefalográficos en pacientes bajo anestesia general para ambientes quirúrgicos”, identified with code 111974454838 granted by COLCIENCIAS, for the beginning of this investigation. And to the scholarship of the ”Bécate Nariño”program awarded by the CEIBA Foundation for the last year of this research.

Abstract

Time series analysis plays an essential role in today's society due to the ease of access to information. This analysis is present in the majority of applications that involve sensors, but in recent years thanks to technological advancement, this approach has been directed towards the treatment of complex signals that lack periodicity and even that present non-stationary dynamics such as signals of brain activity or magnetic and satellite resonance images. The main challenges at the time of time series analysis are focused on the representation of the same, for which methodologies based on similarity measures have been proposed. However, these approaches are oriented to the measurement of local patterns point-to-point in the signals using metrics based on the form. Besides, the selection of relevant information from the representations is of high importance, in order to eliminate noise and train classifiers with discriminant information for the analysis tasks, however, this selection is usually made at the level of characteristics, leaving aside the Global signal information. In the same way, lately, there have been applications in which it is necessary to analyze time series from different sources of information or multimodal, for which there are methods that generate acceptable performance but lack interpretability. In this regard, we propose a framework based on representations of similarity and multiple-instance learning that allows selecting relevant information for classification tasks in order to improve the performance and interpretability of the models.

Keywords: Time-Series analysis, Similarity, Multiple instance learning, EEG, MRI, Satellite images

Resumen

El análisis de series de tiempo juega un papel importante en la sociedad actual debido a la facilidad de acceso a la información. Este análisis está presente en la mayoría de aplicaciones que involucran sensores, pero en los últimos años gracias al avance tecnológico, este enfoque se ha encaminado hacia el tratamiento de señales complejas que carecen de periodicidad e incluso que presentan dinámicas no estacionarias como lo son las señales de actividad cerebral o las imágenes de resonancias magnéticas y satelitales. Los principales retos a la hora de realizar en análisis de series de tiempo se centran en la representación de las mismas, para lo cual se han propuesto metodologías basadas en medidas de similitud, sin embargo, estos enfoques están orientados a la medición de patrones locales punto a punto en las señales utilizando métricas basadas en la forma. Además, es de alta importancia la selección de información relevante de las representaciones, con el fin de eliminar el ruido y entrenar clasificadores con información discriminante para las tareas de análisis, sin embargo esta selección se suele hacer a nivel de características, dejando de lado la información de global de la señal. De la misma manera, últimamente han surgido aplicaciones en las cuales es necesario el análisis de series de tiempo provenientes de diferentes fuentes de información o multimodales, para lo cual existen métodos que generan un rendimiento aceptable pero carecen de interpretabilidad. En este sentido, en nosotros proponemos un marco de trabajo basado en representaciones de similitud y aprendizaje de múltiples instancias que permita seleccionar información relevante para tareas de clasificación con el fin de mejorar el rendimiento y la interpretabilidad de los modelos.

Palabras clave: Análisis de series de tiempo, Similitud, Aprendizaje de múltiples instancias, EEG, MRI, Imágenes satelitales.

List of Figures

List of Tables

Contents

1 Intronduction

1.1. Motivation

Today, machine learning and pattern recognition applications play a significant role in our society due to a large amount of information currently handled and the variety of technological advances to acquire it. The scientific advances regarding this theme are broad and mainly focus on marketing, entertainment systems, computer vision, smart agriculture, and brain-computer interfaces systems (BCI), among others.Besides, the vast majority of the mentioned approaches require algorithms that allow the analysis of time series. This type of analysis is fundamental in the area of medicine since it allows to analyze physiological biomarkers and discover dynamics in the behavior of patients in order to support the assisted diagnosis of pathologies, such as in the analysis of brain activity. Another critical approach is that of the brain-machine interfaces, in which the treatment of time series of brain activity allows to improve the rehabilitation of patients with disabilities and even the adequacy of robotic prostheses. Likewise, the analysis of time series has been directed towards the treatment of images, due to the accessibility to sensors such as magnetic resonance sensors and satellites, which provide sequences of images separated by a specific time interval.

1.2. Problem statement and literature review

According to the above, time series analysis techniques focus on finding patterns that represent their behavior. In the literature, methods have been proposed that allow modeling the behavior of time series in order to predict different types of variables, such as [Chakraborty et al., 1992], which uses neural networks to forecast the behavior of multivariate time series from flour prices signals. In the same direction, other authors have investigated the behavior of time series with purposes related to marketing and economics, based on regression models [Hanssens, 1980, Stock and Watson, 1988]. On the other hand, other works focus their methodologies on the clustering and classification of time series, using measures such as Kullback-Leibler discrimination information and Chernoff information measure, thus addressing nonlinearity in the signals [Kakizawa et al., 1998, Schreiber and Schmitz, 1997]. More recently, time series analysis has been widely used in various fields of research, such as biology for clustering and functional identification of multiple genes [Fujita et al., 2012, Pyatnitskiy et al., 2014], the discovery of climate indices [Ji et al., 2013], and discovering energy consumption pattern [Iglesias and Kastner, 2013]. In addition, applications focused on the urban environment have also been presented such as the analysis of potential violations, early earthquake alerts, and the analysis of population behavior [Shumway, 2003, Liu et al., 2014, Sadahiro and Kobayashi, 2014]. Other approaches proposed in the state of the art have to do with the finances that are presented in order to understand and improve the performance of the estimates in tasks such as: finding seasonality pattern in the signals [Kumar et al., 2002], personal income pattern recognition [Bagnall et al., 2003], creating efficient portfolio [Guam and Jiang, 2007], and discovery patterns from stock time-series [Aghabozorgi and Teh, 2014]. Finally, one of the greatest applications of time series analysis focuses on the medicine and psychology areas, allowing the exploration, identification and discrimination of pathologies, mainly through the detection of brain activity [Wismüller et al., 2002, Gullo et al., 2012, Miao et al., 2017]. Besides, this signals also allow the analysis of human behaviour in psychological domain according to the emotional behaviour in social networks [Kurbalija et al., 2012, Jiang et al., 2014].

One of the main components of time series analysis is its representation to address classification and clustering problems. In this sense, methods have been presented to characterize this type of signals, such as: Discrete Fourier Transform (DFT) [Faloutsos et al., 1994], Discrete Wavelet Transform (DWT) [Kawagoe and Ueda, 2002], Singular Value Decomposition (SVD) [Korn et al., 1997], Piecewise LinearApproximation [Keogh and Pazzani, 1998], Chebyshev Polynomials [Cai and Ng, 2004], Symbolic Approximation [Keogh et al., 2004], and more recently, Choi-Williams Distribution (CWD) [Alazrai et al., 2018] and Common Spatial Patterns (CSP) [Blankertz et al., 2008], the last two used in this work. The above representation methods need a measure that allows comparing the time series characterized, thus, the search for adequate time-series similarity is given close attention, centering mainly on finding shape-based relationships. To this end, the point-to-point comparison is the most implemented algorithm (Euclidean distance (ED) [Faloutsos et al., 1994], Cross-correlation [Golay et al., 1998], Dynamic Time Warping [Berndt and Clifford, 1994], Edit Distance with RealPenalty (ERP) [Chen and Ng, 2004], Minimal Variance Matching(MVM) [Latecki et al., 2005], among others), being efficient for short-time series or with periodic waveforms to extract local temporal and/or frequency information. However, they may have limited ability to capture structural similarity of longtime series which have repetitive waveforms, for instance, electrocardiography(ECG) and electroencephalography(EEG) signals. A more appropriate alternative to determine similarity between long sequences is to measure their similarity based on higher-level structures. Several structure or model-based similarities have been proposed that extract global features such as trend, autocorrelation, skewness, and model parameters from data [Nanopoulos et al., 2001, Wang et al., 2006]. However, it is not trivial how to determine relevant features, and compute distances given these features [Keogh, 2004].

In this regard, in order to capture the high-level structural information of time series, [Lin et al., 2012] proposed a bag-of-patterns(BoP) representation. The temporal order of local segments, i.e., local patterns, in a time series is ignored and all the local segments in the time series are histogrammed to construct a bag-of-pattern representation The bag-of-patterns representation is effective to capture the structural similarity of time series. However, one drawback of the bag-of-patterns representation is that its dimension may be very high, which limits its application for large datasets [Wang et al., 2013a].In order to solve this problem, algorithms for the selection of instances such as [Li et al., 2009] and [Fu et al., 2010] have been proposed. However, these algorithms require a selection criterion that does not take into account the classes and that often choose non-discriminant instances, further, these may present failures in multimodal time series analysis. For multimodal representations, approaches such as [Mühling et al., 2012] and [Zhu et al., 2006] have been proposed, which despite having acceptable performance, lack interpretability.

1.3. Objectives

According to the previous approach, we present the following objectives:

1.3.1. General objetive

Develop a framework based on a similarity representation of time series that allows selecting relevant information for classification tasks in order to improve the performance and interpretability of the models.

1.3.2. Specific objetives

- Develop a feature representation approach for combining multiple data sources through kernel methods, that improves the performance on multimodal problems while preserving the model interpretability.

- Develop a feature selection methodology based on L1 regularization devoted to time series similarity representations, aiming to select discriminant information that improves classification task performance.

- Develop an instance-level bag optimization methodology based on expansion and instance selection that allows enhancing the time series representation providing interpretability and supporting discrimination tasks.

2 Mathematical preliminaries

In this chapter, a brief introduction of mathematical methods used for the development of this work are provided. The content of this chapter is based on book[Theodoridis and Koutroumbas, 2009]

2.1. LASSO sparse regression model

Given a linear regression with standardized predictors $\boldsymbol{X} : \{\boldsymbol{x}_{ij} \in \mathbb{R}^D\}, i{=}1, \ldots, M$ and centred respose values l_i, lasso solves the follow regression problem, finding vector u to minimize [Tibshirani, 1996]:

$$\sum_{i=1}^{M} \left(l_i - \sum_{j} \boldsymbol{x}_{ij} u_j \right)^2 + \lambda \sum_{j=1}^{D} |u_j|, \tag{2-1}$$

this is equivalent to:

$$\mathbf{u} = \arg\min_{\boldsymbol{u}} \frac{1}{2} \left\| \boldsymbol{X}\boldsymbol{u} - \boldsymbol{l} \right\|_2^2 + \lambda \left\| \boldsymbol{u} \right\|_1 \tag{2-2}$$

where $\|.\|_1$ denotes the l_1-norm, $\boldsymbol{l} \in \mathbb{R}^M$ is a vector containing class labels $\{+1, -1\}$. To optimize $\boldsymbol{u}$ in equation (2-2), the coordinate descent algorithm is adopted as [Friedman et al., 2010]:

$$\widetilde{u}_j \leftarrow S\left(\sum_{i=1}^{M} x_{ij} \left(l_i - \widetilde{l}_i^{(j)} \right), \lambda \right) \tag{2-3}$$

where $\widetilde{l}_i^{(j)} {=} \sum_{d \neq j} \left(x_{i,d} \widetilde{u}_d \right)$ is the fitted value excluding the contribution from $x_{i,j}$, and $S(a, \lambda)$ is a shrinkage-thresholding operator defined as below:

$$\mathrm{sign}(a)(|a| - \lambda)_+ = \begin{cases} a - \lambda & if \quad a > \lambda \\ 0 & if \quad |a| \leqslant \lambda \\ a + \lambda & if \quad a < -\lambda \end{cases} \tag{2-4}$$

So, we calculate the optimized sparse vector $\widetilde{\mathbf{u}}$, satisfying equation (2-2) by repeating the update until convergence. The column vectors in $\mathbf{X}$ are those zero-entries in $\widetilde{\mathbf{u}}$, which are excluded from an optimized feature set $\widetilde{\mathbf{X}}$ under assumption that it has lower dimensionality than $\mathbf{X}$. The real-valued λ determines the sparsity degree of $\widetilde{u}$, and hence it rules the selection of feature sets.

2.2. Support Vector Machine

Let a $\boldsymbol{x}_i, i{=}1, \ldots, M$, be the feature vectors of the training set $\boldsymbol{X}$. These belong in two classes $\{+1, -1\}$, the goal is design a hyperplane:

$$\boldsymbol{w}^\top \boldsymbol{x} + w_0{=}{\pm}1, \tag{2-5}$$

and for the case of nonseparabe classes, we introduce the slack variables form, such as:

$$l_i \left[\boldsymbol{w}^\top \boldsymbol{x} + w_0\right] \geq 1 - \xi_i \tag{2-6}$$

Being ξ_i the slack variables. The goal now is to make the margin as large as possible but at the same time to keep the number of points wiht $\xi_i > 0$ (vectors that are misclassified) as small as possible. This is equivalent to minimize the follow cost function:

$$\min_{w,w_0,\xi} \left(J(w, w_0, \xi){=}\frac{\|w\|^2}{2} + C\sum_{i=1}^{M} \xi_i \right), \tag{2-7}$$

$$\begin{aligned} s.t. \qquad & y_i \left(\boldsymbol{w}^T \boldsymbol{x}_i + \boldsymbol{w_0}\right) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \qquad i{=}1, \cdots, N \end{aligned}$$

After applying the corresponding Lagrangian $\boldsymbol{\alpha}$ the dual form of the SVM is as follows:

$$\max_{\alpha} \left(-\frac{1}{2}\sum_{ij}^{M} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i \boldsymbol{x}_j + \sum_{i}^{M} \alpha_i \right),$$

$$\begin{aligned} s.t. \qquad & \sum_{i}^{M} \alpha_i y_i{=}0, \\ 0 \leq \alpha_i \leq C. \qquad & i{=}1, \cdots, M \end{aligned} \tag{2-8}$$

Where, $\boldsymbol{\alpha}$ is the vector of weights of support vectors. Solving the problem of cuadratic programming in eq.2-8, the samples are classified [Theodoridis and Koutroumbas, 2009].

3 Multiple Instance Representation

3.1. Multiple Instance Learning (MIL)

MIL is a supervised learning paradigm formally introduced by [Dietterich et al., 1997] for drug activity prediction tasks, which involve ambiguous training samples called *Bags*. Each bag is a set of feature vectors named instances, and training labels are only attached to bags instead of instance labels, i.e, $\boldsymbol{B}_i \in \mathbb{R}^{D \times N_i} : \{\boldsymbol{x}_{ij} \in \mathbb{R}^D, j = 1 \ldots N_i\}$, with label l_i. Where $\boldsymbol{x}_{ij}$ represents j-th instance in i-th bag, D is the number of features and N_i is the total number of instances in the bag $\boldsymbol{B}_i$. The primary assumption for MIL is that a positive bag contains at least one positive instance and a negative bag only contains negative instances, however other studies propose the assumption that each instance contributes in some way to the classification of the bags [Zhou et al., 2012].

3.2. MIL-based Similarity Representation

From the resulting feature selection stage, we embed trials in a vector space aiming to apply conventional classification machines [Chen et al., 2006] through an instance-level feature mapping based on Diverse Density (DD) framework [Maron and Lozano-Pérez, 1998]. DD assumes that exist a single target concept which can be used to label the individual instances correctly. Then, the diverse density of a given concept $\boldsymbol{v}$ is defined as the probability that the concept is the target given the training bags:

$$DD(\boldsymbol{v}) = P(\boldsymbol{v}|\boldsymbol{B}_1, \cdots, \boldsymbol{B}_i, \cdots, \boldsymbol{B}_M), \tag{3-1}$$

DD also can be interpreted from a feature selection point of view. Given a set of concepts $\boldsymbol{V}$, each concept $\boldsymbol{v} \in \boldsymbol{V}$ defines an attribute or feature denoted as $h_{\boldsymbol{v}}$, i.e., $h_{\boldsymbol{v}}(\boldsymbol{B}_i) = P(\boldsymbol{v}|\boldsymbol{B}_i)$, then, if $\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_{N_v}\}$ where N_v is the total number of concepts, the feature vector for i-th bag is defined as follows:

$$[h_1(\boldsymbol{B}_i), h_2(\boldsymbol{B}_i), \cdots, h_{N_v}(\boldsymbol{B}_i)]^\top = [P(\boldsymbol{v}_1|\boldsymbol{B}_i), P(\boldsymbol{v}_2|\boldsymbol{B}_i), \cdots, P(\boldsymbol{v}_{N_v}|\boldsymbol{B}_i)]^\top \tag{3-2}$$

As result, a bag is a point in the new feature space and $[P(\boldsymbol{v}_{N_v}|\boldsymbol{B}_1), \cdots, P(\boldsymbol{v}_{N_v}|\boldsymbol{B}_M)]$ represents the feature h_{N_v} for all bags. Later, maximize the DD function is equivalent to selecting a feature h_{N_v} that maximize $f(h_{N_v}) = \prod_{i=1}^{M} h_{N_v}(\boldsymbol{B}_i)$=$\prod_{i=1}^{M} P(\boldsymbol{v}_{N_v}|\boldsymbol{B}_i)$ using a similarity measure based on the most-likely-cause estimator proposed in [Chen et al., 2006], as follows:

$$P(\boldsymbol{v}_{N_v}|\boldsymbol{B}_i) \propto s(\boldsymbol{v}_{N_v}, \boldsymbol{B}_i)=\max_{j} \exp\left(-\frac{\|\boldsymbol{v}_{N_v} - \boldsymbol{x}_{ij}\|^2}{\sigma_s^2}\right), \tag{3-3}$$

Further, this framework assumes there may exist more than one target concept to represent a bag, i.e., each instance in the training bags can approximate a target concept. Then, we compare each new itraining instance $\boldsymbol{x} \in \mathbb{R}^D$ against all the bags to form the similarity matrix $\boldsymbol{S} \in \mathbb{R}^{N_c \times M}$ as below:

$$\boldsymbol{S} = \begin{bmatrix} s(\boldsymbol{x}_1, \boldsymbol{B}_1) & \cdots & s(\boldsymbol{x}_1, \boldsymbol{B}_M) \\ s(\boldsymbol{x}_2, \boldsymbol{B}_1) & \cdots & s(\boldsymbol{x}_3, \boldsymbol{B}_M) \\ \vdots & \ddots & \vdots \\ s(\boldsymbol{x}_{N_c}, \boldsymbol{B}_1) & \cdots & s(\boldsymbol{x}_{N_c}, \boldsymbol{B}_M) \end{bmatrix}, \tag{3-4}$$

being, $s(\boldsymbol{x}_{j_c}, \boldsymbol{B}_i), j_c \in [1, N_c]$ the similarity function which holds the probability of an instance belonging to i-th bag, N_c represent the total number of training instances concatenated for all training bags, M is the samples number and σ is the bandwidth of Gaussian kernel.

4 Time Series Feature Representation for Combining Instance of Multiple Sources

Initially, we focus on problems that show time-series with ambiguous sequentiality some of them require a multimodal approach, such as the time series of satellite images that present records from several sensors, as well as data records of different nature as medical imáges, even that present missing data issues. These records are mainly separated by a significant amount of time [Verbesselt et al., 2010], and even by variable time intervals or that present abnormality in the sequences [Xiang et al., 2014]. In this sense, we propose a multimodal time-series MIL representation based on multiple kernel learning (MKL) named MILMKL. The proposed methodology is developed as follows:

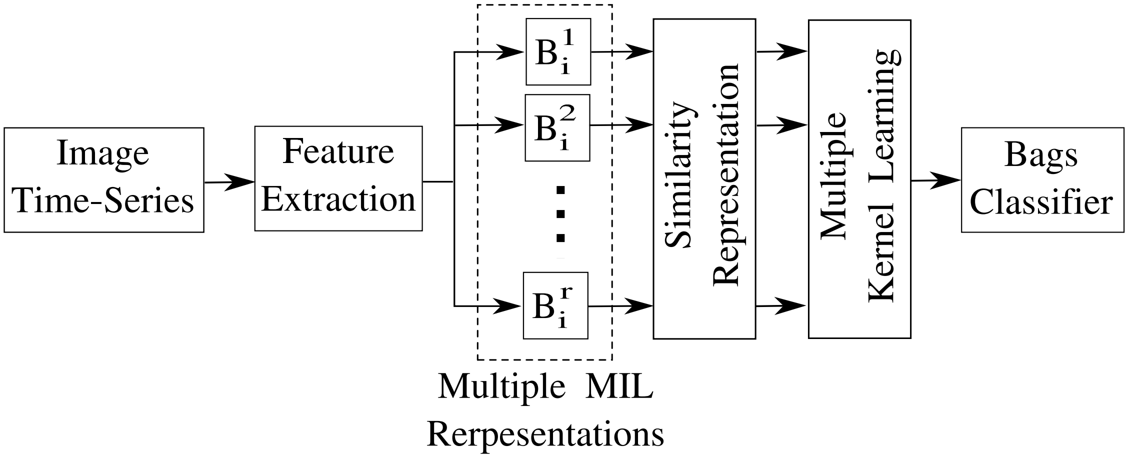


Figura **4-1**: General scheme of proposed methodology for combining instances of multiple source

4.1. Materials and Methods

4.1.1. Dataset and Pre-processing

For this methodology, we use two databases of image time series. The first corresponds to time series of satellite images and the second to time series of medical images:

- *Mato Grosso MODIS image time series [1]:* This data set uses the MOD13Q1 product National Aeronautics andSpace Administration from 2001 to 2016, provided every 16 days at250-meter spatial resolution in the sinusoidal projection [Picoli et al., 2018], and include maps of land cover classification for the state of Mato Grosso, Brasil. The MODIS images were inserted into the SciDB database in order to create a three-dimensional array of satellite data. In this sense, the ground samples consisting of 2,115 time series with known labels obtained from the web time series service (WTSS), available in the SITS [2] R package. Further, nine land cover classes are defined: (1) forest, (2) cerrado, (3) pasture, (4) soybean-fallow (single crop), (5) fallow-cotton (single crop), (6) soybean-cotton (double crop), (7) soybean-corn (double crop), (8) soybean-millet (double crop), and (9) soybean-sunflower (double crop).

- *Alzheimer's Disease, employing data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [3]:* This data set gathers several imaging (serial MRI), other biosignals, and neuropsychological assessments to characterize MCI and AD patients. The main goal of this initiative is to evaluate the combined prognostic value of several AD biomarkers, clinical and neuropsychological assessments. The ADNI database includes more than 900 subjects aging between 50 years to 90+ years with an annual follow-up of 3 years. All ADNI subjects were scanned at the screening with a battery of medical tests. The patients attended the clinicians every six months and were diagnosed as healthy, MCI, MCI converted to AD, MCI converted to healthy, AD, AD converted to MCI, healthy converted to MCI by clinical specialists on neurology and neuropsychology based on the testing battery. Despite the fact that all patients had been followed up to ten years after the first screening visit, they differed in the scheduled time and number of visits. In this study, we select a set of 570 adult patients, belonging to one of the following classes:*i)* Stable MCI state (*sMCI*), patients who had been diagnosed with MCI for at least 36 months and *ii)* converted MCI state (*cMCI*), patients converted to the AD state within 36 months after the baseline diagnosis of MCI. For processing the MRIs recorded at each patient visit, we use the widely known `FreeSurfer` [4] pipeline because of its test-retest reliability across scanner manufacturers and field strengths.

[1] https://doi.pangaea.de/10.1594/PANGAEA.881291

[2] https://github.com/e-sensing/sits

[3] www.adni-info.org

[4] freesurfer.nmr.mgh.harvard.edu

4.1.2. Feature Extraction and Multiple MIL Representations

In order to address multimodal time-series problems, we propose an approach based on multiple MIL representations, so that, r different bags, each with a type of instance, whether of different nature or from different sources. Then, we define a image time serie based on bags representation in Section 3.1 as below:

$$\boldsymbol{Z}_i : \{\boldsymbol{B}_i^r,\ \forall r \in [1, \Phi]\},\ \ i \in [1, M]\,, \tag{4-1}$$

Being Z_i a multimodal image time series and $\boldsymbol{B}_i^r \in \mathbb{R}^{D \times N_i^r} : \{\boldsymbol{x}_{ij}^r \in \mathbb{R}^D, j = 1 \dots N_i^r\}$ represents i-th bag formed by feature vectors over time (instances) of r type. Next, we calculate a set of similarity matrix $\boldsymbol{S} = \left[s(\boldsymbol{x}_{ij}^r, \boldsymbol{B}_i^r), \forall i, j\right]$ (one for each type of bag r) as in Section 3.2.

Further, in the feature extraction stage, for the land cover classification the vegetation we use the index bands NDVI and EVI along with the nir and mir bands over time for each time series as features. On the other hand, for ADNI dataset we use features from the `FreeSurfer` processing pipeline extracts a set of morphological measurements, characterizing cortical and subcortical brain structures. The first stage adjusts the bias field of the structural images, normalizing their intensity. The second stage performs the skull stripping, segmentation of the gray and white matter, and brain parcellation. The third stage tessellates the resulting structures, computing the thickness, area, and volume as the set of morphologic features for describing each brain structure. Volume measurements are further normalized concerning the Total Intracranial Volume [Buckner et al., 2004]. The MRI processing result is a feature matrix that holds 2425 instances with D=312 features summarized in table **4-1**.

Type	*Number of features*	*Units*
Cortical Volumes (CV)	69	mm^3
Subortical Volumes (SV)	39	mm^3
Surface Area (SA)	68	mm^2
Thickness Average (TA)	68	mm
Thickness Std. (TS)	68	mm
Total number of features (P)	312	
Total number of structures (D)	61	

Table **4-1**: List of morphological features extracted by `FreeSurfer`

4.1.3. Multiple kernel learning for bag classification

Given the instance-based representation $\boldsymbol{s}$ (see Section 3.2) of a bag with label l_i, we incorporate the generalized inner product to measure the similarity between a couple of bags. So, the following kernel function implements the similarity:

$$\kappa(\boldsymbol{B}_i, \boldsymbol{B}_{i'}) = \langle \varphi(\boldsymbol{s}_i), \varphi(\boldsymbol{s}_{i'}) \rangle; \quad \forall i, i' \in [1, M], \tag{4-2}$$

where notation $\langle \cdot, \cdot \rangle$ stands for the inner product and $\varphi(\cdot){:}\mathbb{R}^{N_c} \to \mathcal{H}$ maps from the instance-based feature space $\mathbb{R}^{N_c}$ into a Reproduced Kernel Hilbert Space (RKHS) $\mathcal{H}$, so that $|\mathcal{H}| \gg N_c$.

Although the kernel function allows implementing operations on RKHS by the above introduced pair-wise similarity of bags, the instances provided with multiple sets of features (usually, having different nature) result in different mappings of instance-based features, generating a set of kernel functions $\{\kappa_r(\cdot, \cdot){:}r{=}1, \ldots, \Phi\}$ that need to be combined into a single one to attain the similarity of a pair of bags. To solve this issue, Multiple Kernel Learning (MKL) approaches attempt to condense several reproduced spaces into a single representation through a convex sum of kernels, relying on two properties: i) The positive-definite kernel weights enable to extract the relative importance of the combined kernels by looking at them, ii) Imposed on the weights, the nonnegative constraint corresponds to scaling the feature spaces and using the concatenation of them as the combined feature representation [Gönen and Alpaydın, 2011]. Thus, we use the following combined feature representation:

$$\kappa_{\boldsymbol{\mu}}(\boldsymbol{B}_i, \boldsymbol{B}_{i'}) = \sum_{r=1}^{\Phi} \mu_r \kappa_r \left(\boldsymbol{B}_i^r, \boldsymbol{B}_{i'}^r\right), \tag{4-3}$$

where $\boldsymbol{B}_i^r$ stands for i-th bag with instances in the r-th input feature space, that is, $\boldsymbol{x}_{ij}^r \in \mathbb{R}^{D_r}$. The kernel function $\kappa_r{:}\mathbb{R}^{N_c} \times \mathbb{R}^{N_c} \to \mathbb{R}$ maps a bag from the r-th instance-based feature representation, $\boldsymbol{s}^r$, to the r-th RKHS. Without loss of generalization, we denote all instance-based feature mappings as N_c-dimensional since the same number of concepts is used to build them. Vector $\boldsymbol{\mu} \in \mathbb{R}^{\Phi}$, corresponding to the mixture weights, is constrained by $\mu_r \geq 0$ and $\sum_{r=1}^{\Phi} \mu_r{=}1$, aiming to reproduce a convex function of the combining kernels κ_r. We optimize $\boldsymbol{\mu}$ by maximizing the Centered Kernel Alignment (CKA) that estimates the correlation between an ideal target kernel $\boldsymbol{L}$, built from the provided labels, and the convex-combined kernel $\boldsymbol{K}_{\boldsymbol{\mu}}$ [Cortes et al., 2012]:

$$\rho(\boldsymbol{\mu}) = \frac{\left\langle \bar{\boldsymbol{K}}_{\boldsymbol{\mu}}, \bar{\boldsymbol{L}} \right\rangle_F}{\|\bar{\boldsymbol{K}}_{\boldsymbol{\mu}}\|_F \|\bar{\boldsymbol{L}}\|_F}, \tag{4-4}$$

where $\langle \cdot, \cdot \rangle_F$ and $\| \cdot \|_F$ are the Frobenius inner product and matrix-based norm, respectively. Notation $\bar{\boldsymbol{K}}$ stands for the centered version of $\boldsymbol{K} \in \mathbb{R}^{M \times M}$ computed as $\bar{\boldsymbol{I}}\boldsymbol{K}\bar{\boldsymbol{I}}$ with $\bar{\boldsymbol{I}}{=}\boldsymbol{I}_M - \boldsymbol{1}_M\boldsymbol{1}_M^\top/M$, $\boldsymbol{I}_M$ as the M-sized identity matrix, and $\boldsymbol{1}_M$ as a column vector of M ones. Given that CKA maximization criterion jointly determines the mixture weights of a convex combination of kernels, maximizing equation (4-4) in terms of $\boldsymbol{\mu}$ is equivalent to solving the following constrained quadratic optimization problem:

$$\begin{aligned} &\min_{\boldsymbol{\mu}\in\mathbb{R}^{\Phi}} \{\boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu} - 2\boldsymbol{b}^\top\boldsymbol{\mu}\} \\ &\text{s. t.}: \mathbf{1}_{\Phi}^\top\boldsymbol{\mu} = 1 \\ &\qquad \boldsymbol{\mu} \geq \mathbf{0}_{\Phi} \end{aligned} \tag{4-5}$$

where matrix $\boldsymbol{A}{=}\big[\langle\bar{\boldsymbol{K}}_r, \bar{\boldsymbol{K}}_{r'}\rangle_F : \forall r, r'\big] \in \mathbb{R}^{\Phi\times\Phi}$ holds the Frobenius product among the combining kernels, and vector $\boldsymbol{b}{=}\big[\langle\bar{\boldsymbol{K}}_r, \bar{\boldsymbol{L}}\rangle_F : \forall r\big] \in \mathbb{R}^{\Phi}$ is the Frobenius product between the combining kernels and target kernel.

It has been proven that restricting the mixture weights to be nonnegative agrees with scaling their corresponding reproduced features and concatenate them to build a new RKHS. Hence, the convex combination of kernels allows interpreting the discriminative relevance of each set of features in terms of its assigned weight [Gönen and Alpaydın, 2013].

4.1.4. Classification and Performance Assessment

In order to validate the proposed methodology, we present two paths, one for each database to study. For the Mato Grosso database, we build bags in two different ways. The first, define instants of time in all featured bands for the same pixel as instances, in this way we obtain a kernel for each measurement of the sensors. So for the second approach, we define the instances as the features taken from each sensor over time. That is, kernels are formed for each band of the sensors present in the image. We also use a cross-fold validation of 5 folds and an SVM classifier with Gaussian kernel in order to compare ourselves with other state-of-the-art methods. Further, we present the performance measures mostly used in Land Cover classification tasks called User (UA) and Producer (PA) accuracy together with the general accuracy [Foody, 2002].

For ADNI dataset we define the bags and instances as the subjects and their MRI volumes, respectively. Then, to write the diagnosis problem regarding the multiple-kernel learning, we further designate the r-th feature space as the morphological features, describing the r-th brain structure extracted by `FreeSurfer`. Consequently, the resulting mixture vector weighs each structure according to its discrimination capability. For the sake of comparison, we consider two baseline approaches based on the listwise deletion (LD) to deal with missing data by removing the incomplete visits and patient attendance as suggested in [Peugh and Enders, 2004]. The first baseline approach, termed *augmented LD*, increases the feature space by concatenating the D features extracted from each selected instance so that the bag becomes a single instance. Another approach, termed *MIL-LD*, performs the multi-instance learning after the listwise deletion. Managing the bag representations, two standard classification machines, Support Vector Machine (SVM) and Kernel k-Nearest Neighbors (Kk-NN), are used for solving the binary diagnosis task. In either case, we employ the same Gaussian kernel due to its universal approximating property, estimating its bandwidth parameter as addressed in [Álvarez-Meza et al., 2014]. Also, the number of neighbors of Kk-NN and SVM parameter (that rule the trade-off between the slack variable penalty and the margin, respectively) are off-line tuned by an exhaustive search algorithm, fitting the best-averaged validation performance.

The considered bag representations and classifiers are evaluated regarding four well-known performance measurements, namely, accuracy (Acc), sensitivity (S_n), specificity (S_p), and F1-score (F_1). Note that the performed results are reported as the measures averaged over five validation folds and all testing patients.

4.2. Results and Discussion

Land use classification is fundamental, mainly in smart agriculture. Our proposed method has broad applicability for the classification of agricultural areas as evidenced in Table **4-2**, which shows high PA for all classes (greater than 90 %), except for Soy-Millet class (88 %). The main reason for this phenomenon is that class 8 presents a radiometric reflection similar to that of classes 5, 6, and 9, overtime, which leads to structural similarities between the classes mentioned for the bands of study. The above is because the classes 5,6,7,8 and 9, can be framed within the family of cereals and have similar growth stages. Table **4-2** also shows the UA for each class, which indicates that the reliability of our algorithm is mostly high for all classes except for the Fallow-Cotton (89 %) and Soy-Sunflower (88 %) classes because they are the classes less frequently present in the database. Despite this, our algorithm presents robustness against the imbalance between classes.

Table **4-3** shows the performance of the two proposed approaches compared to three state-of-the-art methods. The first two methods use decision trees [Chen et al., 2018] and random forest [Kastens et al., 2017] classifiers to discriminate between classes, but in their methodologies, they propose a classification in which similar classes come together to help classifiers distinguish between them. Our method, on the other hand, classifies the nine classes proposed in the database and still maintains a better performance to the mentioned methods. The third method shown creates vectors of high-dimension characteristics to nourish an SVM [Picoli et al., 2018], which surpasses our MILMKL proposal that although it has higher UA reliability, the precision provided by the PA measure is low. This behavior is because, for the task of discrimination in time series of satellite images, it is essential to preserve the sequentiality of the signal, a characteristic that is lost when comparing the time series with measurements of high-level structures, such as the MIL-based similarity representation. In this sense, we propose bags made up of instances extracted from each sensor in our proposal called MILMKL *. This new proposal maintains the sequentiality of the signal, improving accuracy and performance while maintaining reliability.

	1	2	3	4	5	6	7	8	9	UA
1 Cerrado	391	0	0	10	0	0	0	0	0	**0.97**
2 Fallow-Cotton	0	34	0	0	1	3	0	0	0	**0.89**
3 Forest	7	0	137	0	0	0	0	0	0	**0.95**
4 Pasture	2	0	1	360	3	0	0	4	0	**0.97**
5 Soy-Corn	0	0	0	0	361	20	0	20	0	**0.9**
6 Soy-Cotton	0	0	0	0	8	375	0	2	0	**0.97**
7 Soy-Fallow	0	0	0	0	0	0	88	0	0	**1.0**
8 Soy-Millet	0	0	0	0	20	1	0	207	2	**0.9**
9 Soy-Sunflower	0	0	0	0	5	0	0	2	51	**0.88**
PA	**0.98**	**1.0**	**0.99**	**0.97**	**0.91**	**0.94**	**1.0**	**0.88**	**0.96**	

Table **4-2**: Confusion matrix and UA/PA measures for the proposal MILMKL*

The importance of the proposed method is evident given the global problem of deforestation and misuse of natural resources, and since Colombia is particularly one of the most diverse countries in terms of plantations and forests. Then, we can extrapolate our method to different databases in many of the national territories, such as the Amazon rainforest and the department of Nariño.

Approach	*UA(%)*	*PA(%)*	*Acc(%)*
Decision tree classifier [Chen et al., 2018]	74	72	73.0
Random forest classifier [Kastens et al., 2017]	79	72	79.0
SVM classifier [Picoli et al., 2018]	94	93	94.0
Proposal MILMKL	**95**	92	93.8
Proposal MILMKL*	94	**96**	**94.8**

Table **4-3**: Achieved performance of proposed MILMKL and MILMKL* algorithm compared with land cover classification state-of-the-art approaches.

On the other hand, for the ADNI dataset, the obtained morphological measures from MRI scans are used to perform the patient-wise conversion, predicting the transition from the

MCI to AD states. In this regard, our proposed methodology aims at enhancing the classification performance while dealing, at the same time, with missing instances and improving the interpretability of the morphological information. This approach defines the missing instances as the missed follow up visits per patient or her/his dropping out of the study. To deal with missing instances, the suggested instance-based feature mapping is compared with the bag standardization approaches (augmented LD and MIL-LD). table **4-4** displays the classification performance achieved by the two considered classification machines. For purposes of comparison, each classifier is tuned independently, so that the third column illustrates their chosen optimal parameters. Thus, the best classification accuracy (71,3 %) is reached by Kk-nn with F1-score 64,7 %, followed by an SVM classifier (67,8 %) with F1-score 51,5 %. In this case, the obtained advantage is explained since the Kk-nn algorithm can be settled as a piecewise linear discriminant function mapped onto the RKHS, rather than as a simple linear function in SVM spaces. Also, table **4-4** shows that the intended instance-based feature mapping mostly outperforms the listwise method to face the missing data effect, employing either classifier. Thus, the augmented LD achieves a higher averaged sensitivity, but at the cost of more significant standard deviation. This issue can arise since removing the number of visits enlarges the estimated similarity between patients, affecting the generalization ability. In turn, the proposal increases the averaged F1-score and raises the estimation confidence, reducing its standard deviation due to the specificity improved by 20 %.

Here, interpretability is understood as the capability to infer the relevance of a brain structure for discriminating the reflected classes. So, we apply the MKL algorithm to the morphological measures that are grouped to enhance the interpretability of brain structures, resulting in a set of 61 instance-based feature mappings weighted by the convex combination of their respective kernels. Comparison against the following five state-of-the-art approaches is further accomplished: *i)* the multi-modal imputation for SVM-based classification [Ritter et al., 2015], *ii)* the domain transfer learning using MRI and PET [Cheng et al., 2015], *iii)* the biomarker learning using machine learning [Moradi et al., 2015], *iv)* the multi-scale feature extraction from MRI data [Hu et al., 2016], and *v)* the automatic feature-selection using genetic algorithms [Beheshti et al., 2017]. Thus, table **4-4** shows that the MKL-based combination of feature sets outperforms the other approaches in most of the evaluated cases. Specifically, the Kk-nn, which is fed by the combined kernel, reaches the best classification accuracy (77,8 %) in discriminating *sMCI* from *cMCI* patients. Besides, MKL produces better *F1-score* than the single-kernel in $\sim 12\,\%$ and $\sim 24\,\%$ points using Kk-nn and SVM-based classifiers, respectively. Note that the proposal decreases the standard deviation considerably, yielding a more reliable estimate of the involved performance metric. As an illustration, the appraised standard deviation for the single-kernel strategy specificity is more than twice higher. Therefore, MKL achieves a more balanced performance in comparison with the biomarker learning and the multiscale feature extraction, being both biased either towards specificity or sensitivity.

Approach	*Classifier*	*Acc*(%)	*F1-score*(%)	*Sen*(%)	*Spe*(%)
A-LD	K*k*-NN	67.2±20.9	50.7±28.6	50.0±35.3	69.1±23.9
	SVM	60.0±24.4	53.6±27.5	65.0±33.5	63.3±36.5
MIL-LD	K*k*-NN	64.7±17.7	36.8±36.9	35.0±41.8	76.9±16.6
	SVM	62.2±14.8	48.0±27.6	50.0±35.6	22.9±24.9
Proposed MILMKL	K*k*-NN	71.3±6.4	64.7±11.5	57.9±20.8	**81.5**±12.1
	SVM	67.8±6.9	51.5±21.0	44.2±28.0	**85.6±15.0**
	MKL + K*k*-NN	**77.8±3.7**	**76.8±5.1**	**75.9±12.8**	79.5±5.8
	MKL + SVM	**76.1±4.9**	**75.5±5.8**	**74.6±11.7**	77.4±2.1
Multi-modal imputation [Ritter et al., 2015]	SVM	73.4	-	74.0	72.0
Domain transfer learning [Cheng et al., 2015]	SVM	73.4	-	74.3	72.1
Biomarker learning [Moradi et al., 2015]	LDS	74.7	-	88.8	51.6
Multiscale feature extraction [Hu et al., 2016]	SVM	76.7	-	71.8	82.3
GA-based feature selection [Beheshti et al., 2017]	SVM	75.0	-	76.9	73.2

Table **4-4**: Achieved performance of proposed MILMKL algorithm compared with the baseline instance-based feature mapping and state-of-the-art approaches. Average and standard deviation are reported just for the proposal since the contrasting approaches do not communicate these values.

Since the proposed combination outperforms the straightforward instance-based feature mapping, the estimated mixing vector confidently identifies the relevant brain structures, meaning that they contribute mostly to the classification task. Furthermore, we introduce a relevance index that is directly proportional to the combined weight obtained by the MKL algorithm. Namely, the feature set extracted for each anatomical structure is combined, according to its discriminatory capability for supporting sMCI vs. cMCI diagnosis task and relying on the incorporated CKA similarity criterion. Figure **4-1** displays the MKL approach pipeline based on the computing of a combined kernel representation from multiple features mappings per brain area. In this case, each relevance index is extracted in such a way that makes clear their association with the brain anatomy and physiology. Thus, figure **4-2**(a) visually illustrates the brain structure ranked by the appraised relevance, showing that the regions with more relative importance are the Entorhinal cortex, Hippocampus, Banks of the superior temporal sulcus, Middle Temporal, and Medial Orbitofrontal. figure **4-2**(b) displays classification performance when successively including the relevance of brain structures, ranked from the highest to the lowest value, showing the contribution of brain region to the achieved performance. Particularly, including the Hippocampus and Medial Orbitofrontal cortex considerably enhances the accuracy that had been related before to the memory function [Shen et al., 2010].

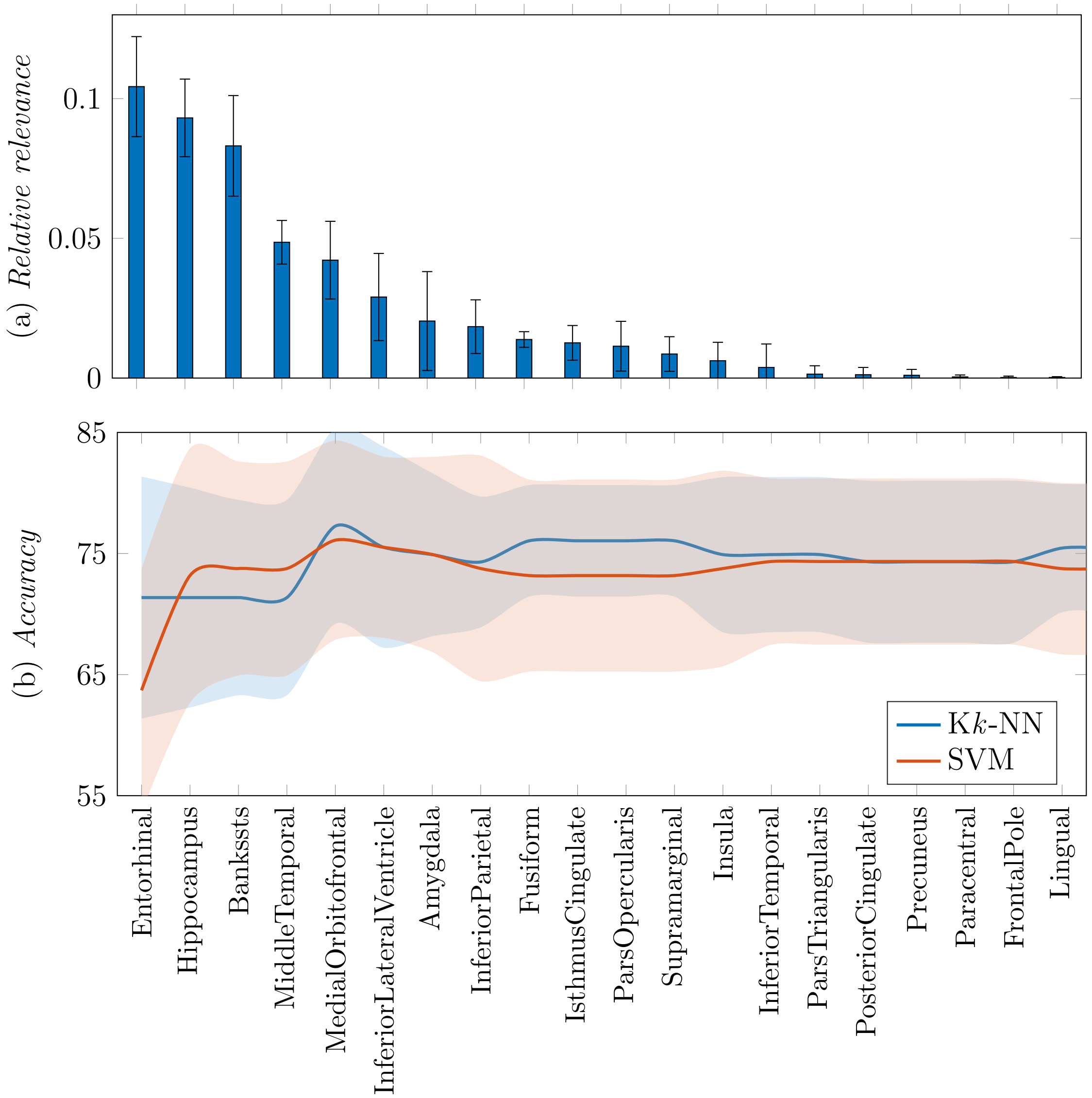


Figure **4-2**: Incremental learning results on the test data for the 20 more relevant brain structures

Ranking of brain structures, ordered by a decreasing value of relevance, enhances the discrimination performance between cMCI and sMCI states. Considering that the similarity estimation of the multi-instance learning hinders the relevance of input features, a structure-wise feature mapping is proposed and followed by a multiple-kernel learning, resulting in a general performance improvement as shown in table **4-4**. It is worth noting that MKL performs the best using only a feature set extracted from MRI scans, while the contrasted baseline works take information from other clinical tests and diagnoses. Moreover, the MKL performance improves as its estimation increases in confidence. Although combining the instance-based features and multiple kernel learning leads to the enhanced identification of the conversion concept from fewer data per subject, one aspect to remark is that the discussed MKL method demands data extracted from a higher amount of subjects.

5 Multiple-instance lasso regularization via embedded instances.

Another important approach in time series analysis is highly non-stationary sequential signals, such as the EEG. One of the main objects of study in the EEG time-series analysis is the search for a similarity measure, generally focusing on shaped-based relationships between them. While some of these approaches work well for short time-series data, the above methodology fails when the sequence is long [Lin and Li, 2009]. Further, exist several approaches that propose methodologies for extracting global features. However, it is not trivial how to determine relevant features and compute the similarity given these features[Ratanamahatana and Keogh, 2004]. This chapter introduces a methodology of feature selection based on L1-regularization for MIL-based similarity representation of EEG time-series named MILRES. The proposed approach is distributed as follows:

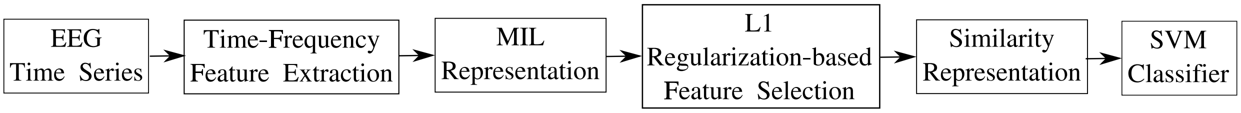


Figura **5-1**: General scheme of MILRES methodology

5.1. Materials and Methods

5.1.1. Dataset and Pre-processing

In order to validate our methodology, we consider EEG signals from the Database for Emotion Analysis using Physiological Signals (DEAP)[1] [Koelstra et al., 2012]. DEAP records physiological signal from 32 subjects (16 males and 16 females) while watching 40 videos, selected to evoke specific emotional states. For each subject, the EEG recording related to a video represents a trial lasting 63 seconds (summing up to 1280 trials). At each trial, the BioSemi ActiveTwo system records 32 channels (using the 10-20 system) for a three-seconds baseline period followed by a 60-seconds stimulus response. After the stimulus, each subject quantified the emotional response for valence, arousal, dominance, and linking in a continuous interval from 1 to 9. Since the two-dimensional valence-arousal model represents several emotional states [Lang, 1995], we carry out our experiment as two binary classification

[1] https://www.eecs.qmul.ac.uk/mmv/datasets/deap/

task, namely, high valence ($[6-9]$) vs. low valence ($[1-5]$), and high arousal vs. low arousal.

For the signal preprocessing, we apply a three-stage pipeline. Firstly, we downsampled the raw EEG signals to 128Hz and excluded the first three seconds of baseline. Secondly, we attenuated the electrooculographic artifacts and band-passed the signals from 4 to 45 Hz so reducing the high-frequency electromyographic noise [Koelstra et al., 2012]. Lastly, we referenced all channels to the common average and selected 22 symmetric ones (6 parietal, 12 frontal, 2 temporal, and 2 occipital) [Alazrai et al., 2018]. Therefore, the i-th trial becomes a matrix of $T = 7680$ time instants and $C = 22$ channels, $\boldsymbol{Z}_i \in \mathbb{R}^{T\times C}$, where $i \in [1, 40]$.

5.1.2. Time-Frequency Feature extraction

For each trial matrix, we compute the quadratic time-frequency features that are known to suitably perform in the emotion recognition task [Alazrai et al., 2018]. Particularly, the Choi-Williams transform (CWD) extracts the quadratic time-frequency distribution at 512 time instants for 1024 frequency bins as:

$$\begin{aligned} CWD_z\left(t,f\right) =& \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} WVD_z\left(\phi,\tau\right)\xi\left(\phi,\tau\right)e^{j2\pi(t\phi-f\tau)}\partial_\tau\partial_\phi, \\ \xi\left(t,f\right) =& \exp\left(-\frac{t^2f^2}{\alpha^2}\right), \end{aligned} \tag{5-1}$$

where $WVD_z\left(t,f\right)$ is the Wigner-Ville distribution of the signal $\boldsymbol{Z}$, $\xi\left(t,f\right)$ is a exponential kernel and α is a parameter that controls the suppression of the cross-terms fixed to 0,5 [Alazrai et al., 2018]. Then, we estimate the following set of 13 CWD-based frequency features at four-second sliding windows with 50 % overlap within the channel-wise CWD [Alazrai et al., 2018]: mean, variance, skewness, kurtosis, the sum of logarithmic amplitudes, median absolute deviation, inter-quartile range of the CWD, root mean square value, flatness, flux, spectral roll-off, normalized Renyi entropy, and energy concentration. As a result, each trial becomes a set of vectors (Bag) $\boldsymbol{B}_i = \{\boldsymbol{x}_{ij} \in \mathbb{R}^D : j = 1\dots N_i\}, i \in [1, M]$, being $D = 13\times C$, $N_i = 29$ the number of sliding windows in a trial and M is the number of trials.

5.1.3. L1 Regularization-based Feature Selection

The MIL framework defines the i-th EEG trial $\boldsymbol{B}_i$ as a bag composed of N_i instances extracted $\boldsymbol{x}_{ij}$, and the provided bag label as $l_i \in \{0, 1\}$. To estimate the bag label, the Multiple-Instance Logistic Regression with LASSO Penalty (MILR-LASSO) aggregates the approximated bags labels as $\tilde{l}_i = I\left(\sum_{j=1}^{N_i} l_{ij} > 0\right)$, where $I(x) = 1$ if $x > 0$ and $I(x) = 0$ in otherwise, and instance labels l_{ij} result from the logistic regression [Chen et al., 2016]:

$l_{ij} \sim \mathit{Bernoulli}\left(p_{ij}\right)$, being $p_{ij} = p\left(\beta_0 + \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}\right)$, $p\left(x\right) = 1/\left(1+e^{-x}\right)$, β_0 the bias term, and $\boldsymbol{\beta} \in \mathbb{R}^D$ the coefficient vector. Therefore, finding the regression paramters becomes a quadratic optimization problem [Chen et al., 2017]:

$$\min_{\beta_0,\boldsymbol{\beta}} \left(-Q_q\left(\beta_0, \boldsymbol{\beta} | \beta_0^t, \boldsymbol{\beta}^t\right) + \lambda \sum_{d=1}^{D} |\beta_d| \right), \tag{5-2}$$

$$Q\left(\beta_0, \boldsymbol{\beta} | \beta_0^t, \boldsymbol{\beta}^t\right) = \sum_{i=1}^{M}\sum_{j=1}^{N_i} l_i \gamma_{ij}^{\ t}\left(\beta_0 + \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}\right) - \log\left(1 + e^{\left(\beta_0 + \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}\right)}\right)$$

where Q_q is the quadratic approximation of Q, γ_{ij} is the conditional expectation given $l_i = 1$, $\gamma_{ij} = p_{ij}/(1 - \prod_{j=1}^{N_i} q_{ij})$, and $\beta_0^t, \boldsymbol{\beta}^t$ are the parameters at iteration t. We solved the optimization problem using the iterative coordinate decent algorithm [Friedman et al., 2010], such as:

$$\beta_0^{t+1} = \frac{Z_0}{\sum_{i=1}^{M}\sum_{j=1}^{N_i} w_{ij}^t},$$

$$\beta_d^{t+1} = \begin{cases} \left(Z_d - \lambda\right) / \sum_{i=1}^{M}\sum_{j=1}^{N_i} w_{ij}^t \left(\boldsymbol{x}_{ij}^d\right)^2 & \text{if} \quad S_d > \lambda \\ \left(Z_d + \lambda\right) / \sum_{i=1}^{M}\sum_{j=1}^{N_i} w_{ij}^t \left(\boldsymbol{x}_{ij}^d\right)^2 & \text{if} \quad S_d < -\lambda \\ 0 & \text{if} \quad |S_d| \leq \lambda \end{cases}, \tag{5-3}$$

$$d = 1, \ldots, D$$

where,

$$Z_0 = \sum_{i=1}^{M}\sum_{j=1}^{N_i} w_{ij}^t \left(u_{ij}^t - \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}^t\right),$$

$$Z_d = \sum_{i=1}^{M}\sum_{j=1}^{N_i} w_{ij}^t \boldsymbol{x}_{ij}^d \left(u_{ij}^t - \beta_0^t - \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}_{(d)}^t\right), \tag{5-4}$$

$\boldsymbol{\beta}_{(d)}^t$ represents $\boldsymbol{\beta}^t$ with its d-th element replaced by 0, $w_{ij}^t = p_{ij}^t q_{ij}^t$, $u_{ij}^t = \beta_0^t + \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}^t + (l_i\gamma_{ij}^t - p_{ij}^t)/(p_{ij}^t q_{ij}^t)$, $p_{ij}^t = 1/(1 + e^{-(\beta_0^t + \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}^t)})$, and λ is the LASSO regularization parameter. Next, the coefficients that exceed a threshold are chosen from the training bags, i.e. $\boldsymbol{\beta}^* = |\boldsymbol{\beta}| > \epsilon$.

Then, we compare each new itraining instance $\widetilde{\boldsymbol{x}} \in \mathbb{R}^{D'}$, $D' < D$ against all the bags to form the similarity matrix $\boldsymbol{S} \in \mathbb{R}^{N_c \times M}$ as below:

$$\boldsymbol{S} = \begin{bmatrix} s(\widetilde{\boldsymbol{x}}_1, \boldsymbol{B}_1) & \cdots & s(\widetilde{\boldsymbol{x}}_1, \boldsymbol{B}_M) \\ s(\widetilde{\boldsymbol{x}}_2, \boldsymbol{B}_1) & \cdots & s(\widetilde{\boldsymbol{x}}_3, \boldsymbol{B}_M) \\ \vdots & \ddots & \vdots \\ s(\widetilde{\boldsymbol{x}}_{N_c}, \boldsymbol{B}_1) & \cdots & s(\widetilde{\boldsymbol{x}}_{N_c}, \boldsymbol{B}_M) \end{bmatrix}, \tag{5-5}$$

being, $s(\widetilde{\boldsymbol{x}}_{j_c}, \boldsymbol{B}_i), j_c \in [1, N_c]$ the similarity function which holds the probability of an instance with $\boldsymbol{\beta}^*$ selected features belonging to i-th bag, N_c represent the total number of training instances concatenated for all training bags, and σ is the bandwidth of Gaussian kernel.

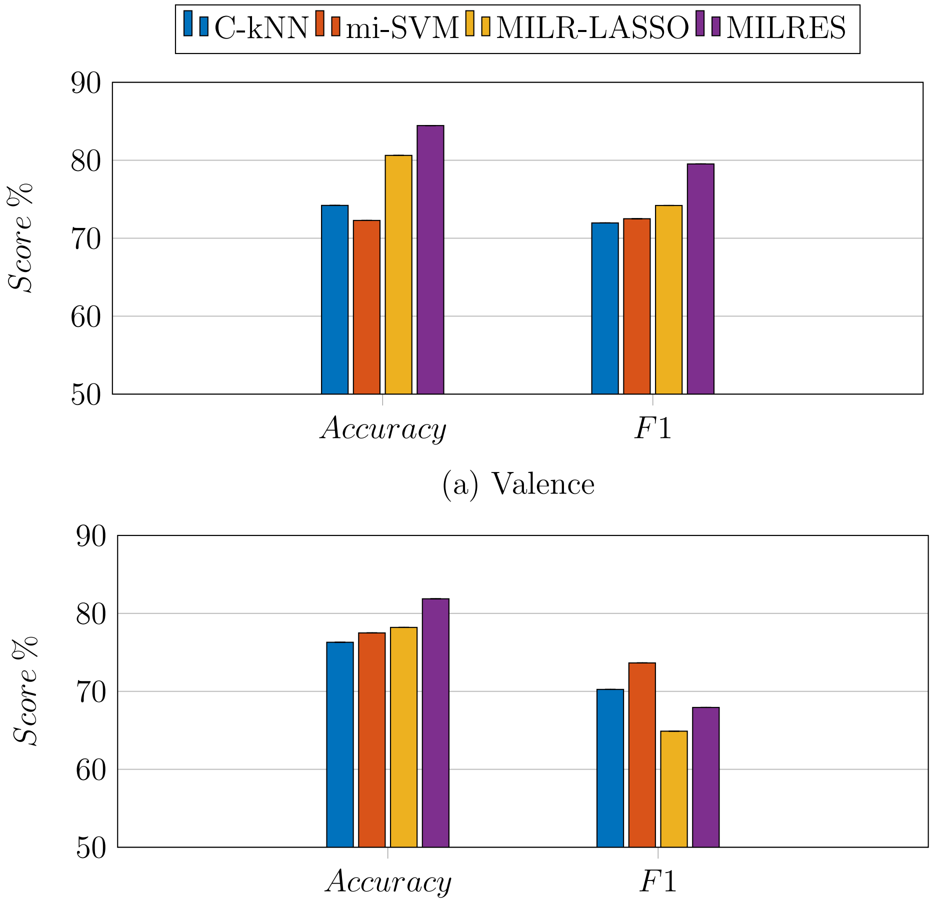
5.1.4. Classification and Performance Assessment

After the MIL representation, the vectors of matrix $\boldsymbol{S}$ feed a support vector machine classifier with a Gaussian kernel to discriminate the emotional states. To tune the regularization parameter λ, the threshold ϵ, and kernel bandwidth, we carried out a 5-fold cross-validation grid search. Due to class imbalance impacts classification results, we reported the F1-score as the performance measure along with the conventional accuracy rate. Besides,in order to evaluate the performance of the proposed methodology across the state-of-the-art methods, we compare our approach against performance results reported for Citation-kNN (C-kNN) [Wang and Zucker, 2000], MILR-LASSO [Chen et al., 2017], and mi-SVM [Zhang et al., 2018a] in the same classification tasks.

5.2. Results and Discussion

Figure **5-2** compares C-kNN, mi-SVM, MILR-LASSO, and the proposed MILRES in terms of their performed accuracy and F1-score for discriminating High vs Low valence and High vs Low arousal. For the valence dimension in figure **5-2**(a), MILRES outperforms the compared approaches from 3 % to 12 % in both, accuracy and F1-score. Regarding the arousal dimension in figure **5-2**(b), MILRES reaches a larger accuracy (81,9 %) than MILR-LASSO (78,2 %), mi-SVM (77,5 %), and C-kNN (76,3 %). However, mi-SVM (73,65 %) and C-kNN (70,25 %) better perform in F1-score than MILR-LASSO (64,89 %) and MILRES (67,94 %). Such a fact is due to the class imbalance in the arousal discrimination problem that implies a smaller number of training instances for the minority class. Therefore, the similarity representation biases towards the larger class, so increasing the average accuracy but reducing the F1-score.

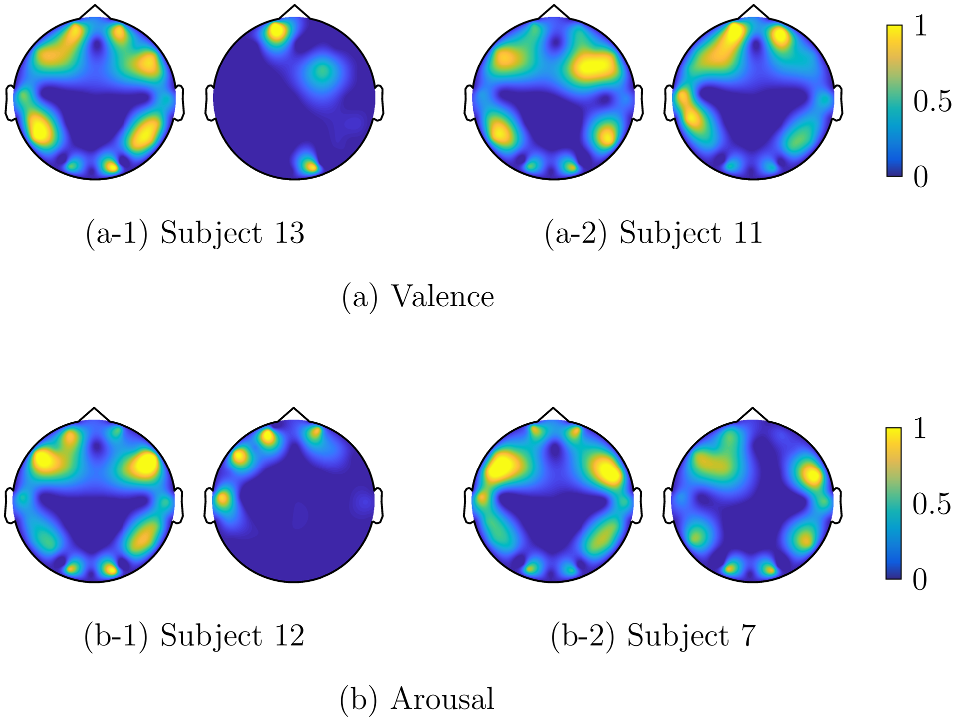
Since we compute the feature set in Section 5.1.2 at the channel level, MILRES allows interpreting the relevance of each EEG channel for discriminating emotional states according to its corresponding regression coefficients β_d. In this regard, the topographic plots in Figure **5-3** illustrate the channel-wise sum of absolute regression coefficients for four subjects before (left column of each subject) and after (right column of each subject) the instance selection stage. Our results indicate that the frontal, prefrontal, temporal, parietal and occipital regions contribute the most in identifying the emotional response, which agrees with previous studies [Zheng et al., 2017, Zhuang et al., 2017]. Particularly, the subjects 13 and 12 (Figure **5-3**(a-1) and Figure **5-3**(b-1) respectively) highly concentrate the relevant information in the frontal, prefrontal, and parietal areas, processing the emotional stimuli, self-reflection, and activation from pleasant and unpleasant emotions [Bermpohl et al., 2006]. Also, the proposed MILRES identifies discriminant information all over the 22 channels for subjects 11 and 7 (Figure **5-3**(a-2) and Figure **5-3**(b-2) respectively) that include parietal, temporal, and occipital. Those areas involve the emotional working memory [Rämä et al., 2001], decision making based on emotions [Deppe et al., 2005], experiencing emotional states [Pelletier et al., 2003], visual processing of emotional images [Lane et al., 1999], and emotional attachment [Gillath et al., 2005]. Therefore, our proposed approach not only suitably discriminates states within an emotional dimension but also identifies the brain areas involved in the process.



(a) Valence

(b) Arousal

Figure **5-2**: Comparison of classification accuracies and F1-scores obtained by C-kNN, mi-SVM, MILR-LASSO, and MILRES.



(a-1) Subject 13

(a-2) Subject 11

(a) Valence

(b-1) Subject 12

(b-2) Subject 7

(b) Arousal

Figure **5-3**: Channel-based spatial activation relevance for two subjects in valence (a) and arousal (b) using the proposed MILRES.

6 Enhanced Instance-level Time Series Representation based on MIL

Following the methodology proposed in the previous chapter, we address the problem of how to enhance the MIL-based representation of time series. This problem can be seen mainly in two ways: The first is an instance selection approach, where one o more instances are selected to represent the corresponding bag label [Fu et al., 2010]. The second viewpoint starts from the dictionary learning approaches which seek the best way to build dictionaries (bags in this case) that represent the data set [Wang et al., 2013b]. In this regard, we propose a methodology of instance-level bag optimization based on L1-regularization which holds two approaches called MILCSP and MIFB for the instance selectión and bag dictionary learning respectively. The following scheme describes the proposed methodology.

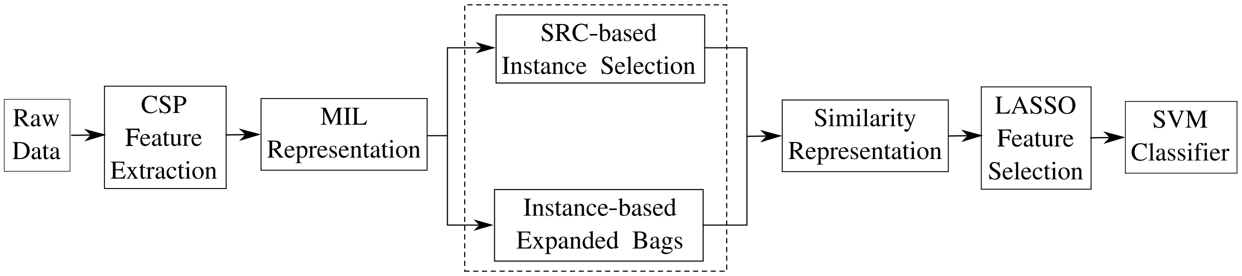


Figura **6-1**: General scheme of proposed methodology

6.1. Materials and Methods

6.1.1. Dataset and Pre-processing

MI Dataset 2a: This signal collection, publicly available at [1], was recorded from nine subjects using 22 EEG channel system sampled at 250 Hz, holding trials regarding one of four MI tasks: left hand, right hand, both feet, and tongue. All recordings were performed in six runs separated by short breaks so that each run held 48 trials (each one lasting 7 s). A short beep indicated the trial start, after which a fixation cross appeared on the black screen within the first 2 s. Then, an arrow (or cue) was shown during 1,25 s to indicate the left,

[1] http://www.bbci.de/competition/iv/

right, up, or down directions, stimulating to imagine a left hand, right hand, both feet, or tongue movement, respectively. Next, each subject performed one MI task within the time interval from 3,25 to 6 s, waiting for the cross to reappear again. The EEG data had been labeled (e.g., left hand and right hand) and the artifacts removed.

The raw EEG data are band-pass filtered using N_f overlapped bandwidths, resulting in $\boldsymbol{Z}_{if}{=}[\boldsymbol{z}_{if}{:}f \in N_f]$, where $\boldsymbol{z}_{if} \in \mathbb{R}^{T\times C}$ is each filtered EEG signal per bandwidth f and trial i. Namely, we use $N_f{=}17$ five-order overlapped bandpass Butterworth filters between 4 Hz and 40 Hz, having a bandwidth of 4 Hz and overlapping rate of 2 Hz as in [Zhang et al., 2015]. Then, the obtained filter-banked signals are time-windowed onto N_τ intervals, each one lasting τ, yielding $\boldsymbol{Z}_{if\tau}{=}[\boldsymbol{z}_{if\tau}{:}\tau \in N_\tau]$ where $\boldsymbol{z}_{if\tau} \in \mathbb{R}^{T'\times C}, T' < T$ is each time-segmented, filtered EEG.

6.1.2. Instance Construction unsig Common Spatial Pattern Feature Extraction

Given a trial matrix $\boldsymbol{Z}_i$ with labels $l_i \in \{-1, +1\}$, Common Spatial Patterns (CSP) find the linear transformation matrix $\boldsymbol{W} = \{\boldsymbol{w}_c \in \mathbb{R}^{C'} : c \in [1, C]\}$ [Blankertz et al., 2008], where each $\boldsymbol{w}$ is calculated to maximize the Rayleigh Quotient (RQ) computed for the mapped data variance between classes as follows:

$$\boldsymbol{w}^* = \max_{\forall \boldsymbol{w}} J(\boldsymbol{w}) = \frac{\boldsymbol{w}^\top \boldsymbol{\Sigma}^{(+1)} \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{\Sigma}^{(-1)} \boldsymbol{w}}, \quad \text{s.t.: } \|\boldsymbol{w}\|_2 = 1 \tag{6-1}$$

where notations $\|\cdot\|_p$ and $\mathbb{E}\{\cdot{:}\forall i\}$ stand for ℓ_p-norm and expectation operator across variable i, respectively. Matrix $\boldsymbol{\Sigma}^{(l)}{=}\mathbb{E}\left\{\boldsymbol{Z}^{l\top}_{if\tau}\boldsymbol{Z}^{l}_{if\tau}{:}\forall i \in R_l\right\}$ is the simplest estimate of class data variance computed at a frequency f and a time segment τ, being R_l the number of available trials per class l. Then, the EEG sample signal $\boldsymbol{Z}_{if\tau}$ is filtered through the learned spatial matrix $\boldsymbol{W} \in \mathbb{R}^{C'\times C}$ holding the C' transformation components. As result, the projected data $\widetilde{\boldsymbol{Z}}_{if\tau}{=}\boldsymbol{Z}_{if\tau}\boldsymbol{W}$ is computed using only the $2H$ representative terms of C' (namely, H first and H last rows). The extracted feature vector $\widetilde{\boldsymbol{z}}_{if\tau}$ is then calculated as below:

$$\widetilde{\boldsymbol{z}}_{if\tau} = \log\left(\operatorname{diag}(\operatorname{var}\{\boldsymbol{Z}_{if\tau}\})\right)/\operatorname{tr}\{\boldsymbol{Z}_{if\tau}\} \tag{6-2}$$

where var$\{\cdot\}$ and tr$\{\cdot\}$, denotes the variance and trace operator, respectively.

As result,for MIL representation, we redefine each column vector $\widetilde{\boldsymbol{z}}_{if\tau} \in \mathbb{R}^{C'}$ as an instance $\boldsymbol{x}_{ij}$, and in turn, we assemble a bag using the whole column vector set: $\boldsymbol{B}_i{=}\left[\boldsymbol{x}_{i1} \ldots \boldsymbol{x}_{N_\tau N_f}\right]$ with $\boldsymbol{B}_i \in \mathbb{R}^{2H\times N_\tau N_f}$.

6.1.3. Instance Selection based on Sparse Representation Classification

Based on the Sparse Representation Classification (SRC) [Shin et al., 2012], we propose an instance selection approach based on dictionary learning and L1-regularization named MILCSP. Firstly, we assume than an instance of i-th bag can be sparsely reconstructed as the linear combination of the elements within the i-th dictionary of the class l such as:

$$\begin{aligned} \boldsymbol{y}_{ij} &= \beta_1 \boldsymbol{a}_1 + \beta_2 \boldsymbol{a}_2 + \cdots + \beta_N \boldsymbol{a}_N \\ \boldsymbol{y}_{ij} &= \boldsymbol{A}_i \boldsymbol{\beta}_i \end{aligned} \tag{6-3}$$

Where the dictionary $\boldsymbol{A}_i$ holds the concatenated instances of other bags belonging to the class of $\boldsymbol{B}_i$, i.e., $\boldsymbol{A}_i = \{\boldsymbol{a}_n : \boldsymbol{a}_n = \boldsymbol{x}_{kj}; \forall l_k = l_i; k \neq i\}, n \in [1, N]$, being N the number of elements in the dictionary. Then, bag instances are ranked according to their results for regressing instances of the same class as follows:

$$\boldsymbol{\beta}_i^* = \arg\min_{\boldsymbol{\beta}_i} \left(\frac{1}{2} \|\boldsymbol{A}_i \boldsymbol{\beta}_i - \boldsymbol{x}_{ij}\| + \lambda_1 \|\boldsymbol{\beta}_i\|_1 \right), \tag{6-4}$$

where $\|.\|_1$ and $\|.\|$ denote the L1 and euclidean norm, respectively, $\boldsymbol{\beta}_i \in \mathbb{R}^N$ represents the vector of scalar coefficients for the linear regression, λ_1 is a positive regularization parameter which controls the sparsity of $\boldsymbol{\beta}_i$. Further, the residual criterion selects the instances with reconstruction error smaller than a threshold as [Wright et al., 2009].

$$\widehat{\boldsymbol{B}}_i = \{\widehat{\boldsymbol{x}}_{ij} \;:\; \|\boldsymbol{x}_{ij} - \boldsymbol{y}_{ij}\| < \mathbb{E}\{\|\boldsymbol{x}_{ik} - \boldsymbol{y}_{ik}\|\}; k = 1, \cdots, N_i\} \tag{6-5}$$

6.1.4. Instance-based Expanded Bags

On the other hand, intending to include the mutual influence of several time-window sizes and in order to explore different brain dynamics in the EEG signals, we enhance the bags representation through the time-window scaling in MIFB approach. Starting from Equation (6-2), we redefine the instances $\boldsymbol{x}_{ij}$ by concatenating all frequency components for a time-window as below:

$$\boldsymbol{x}_{ij}^{\tau} = \begin{bmatrix} \boldsymbol{z}_{i1\tau} \\ \boldsymbol{z}_{i2\tau} \\ \cdots \\ \boldsymbol{z}_{iN_f\tau} \end{bmatrix}, \qquad \boldsymbol{x}_{ij} \in \mathbb{R}^{2HN_f} \tag{6-6}$$

In this regard, we propose an iterative algorithm to expanded bags based on the addition of instances calculated from several window sizes using the incremental rate $\Delta\tau$. Then, the i-th expanded bag is defined as:

$$\boldsymbol{B}_i^{\tau} = \left[\boldsymbol{x}_{it}^{\tau_1}, \boldsymbol{x}_{it}^{\tau_2}, \ldots, \boldsymbol{x}_{it}^{p\Delta\tau}\right] \forall t, \tau = p\Delta\tau \tag{6-7}$$

In this regard, we propose a expanded similarity representation by compute similarity matrix for a time-window of length τ:

$$\boldsymbol{\Delta}_{pq} = [\boldsymbol{S}_{p\Delta\tau}\| \cdots \|\boldsymbol{S}_{N_\tau\Delta\tau} : \forall p, q], \quad p{=}[q, N_\tau],\ q{=}[\tau_{init}, N_\tau], \tag{6-8}$$

where p and q are variables that iterate over the interval $[\tau_{init}, N_\tau]$.Further, in order to find the optimal combination of windows, we propose a grid search as shown in Algorithm 1

Algorithm 1 Instance-based Expanded Bags - *MIFB*

```
Input: CSP filtered training matrix $\boldsymbol{Z}_{if\tau}$
Output: Optimal window combination bag $\boldsymbol{B}_i^*$
$\boldsymbol{\Delta}_{pq} = \phi$
for $q{=}[\tau_{init}, N_\tau]$ do
    for $p{=}[q, N_\tau]$ do
        $\boldsymbol{S}_{p\Delta\tau} = [s(\boldsymbol{x}_{it}^{p\Delta\tau}, \boldsymbol{B}_i^{p\Delta\tau})]$        Calculate similarity matrix        Section 3.2
        $\Delta_{pq} \leftarrow \boldsymbol{S}_{p\Delta\tau}$
        $\tilde{l}_i^{pq}$=SVM($\Delta_{pq}$)
$\boldsymbol{B}_i^*{=}\arg\max_{\forall pq} \mathbb{E}\left\{l_i - \tilde{l}_i^{pq}\right\}$
```

Due to the high computational cost, we propose the intensive version of poposed approach bringing together the stages of bag expansion and instance selection usig RQ defined in equation (6-1). Then, we define a function $h{:}\mathbb{R}^{C'\times N_f N_t}{\to}\mathbb{R}^{C'\times N'_f N'_t}$, with $N'_f N'_t < N_f N_t$, and the new optimal bag is as follows: $\boldsymbol{B}_i^*{=}\{\boldsymbol{x}_{ft}^{\tau} : h(\boldsymbol{B}_i^{\tau}){=}1, \quad \forall f, t, \tau\} \in \mathbb{R}^{C'\times N'_f N'_t}$

$$h(\boldsymbol{x}_{ift}^{\tau}){=}\begin{cases} 1, & J_{ft}^{\tau}{=}\dfrac{\boldsymbol{w}^{*\top}\boldsymbol{\Sigma}_{ft}^{(+1)}\boldsymbol{w}^{*}}{\boldsymbol{w}^{*\top}\boldsymbol{\Sigma}_{ft}^{(-1)}\boldsymbol{w}^{*}} > \epsilon, \\ \\ 0, & \text{otherwise} \end{cases} \tag{6-9}$$

6.1.5. Classification and Performance Assessment

As illustrated in figure **6-1**, the proposed methodology classification framework relies on a bag-based data representation obtained from a CSP-based feature set, appraising the following procedures: *i*) frequency-temporal feature extraction of EEG raw data based on CSP approach, *ii*) MIL representation, *iii*) Enhanced bags representation for MILCSP (Section 6.1.3) and MIFB (Section 6.1.4), *iv*) Instance-based feature mapping using similarity representation (Section 3.2), for which an optimized version named MIFB* usign feature selection in Section 5.1.3 is also considered; *v*) Feature selection performed by LASSO, and *vi*) Bag classification by an SVM algorithm.

For MILCSP approach, we apply CSP method to a set of 48 frequency-temporal segments obtained from both frequency filter bands and time windows for each trial. To obtain the frequency-temporal segments, each trial is filtered using 16 sliding filters of 4Hz between $6 - 40$Hz with 2Hz overlap. Then, each slide filtered trial is segmented using sliding one-second time windows with 50 % overlap [Miao et al., 2017]. In addition for MIFB approach, the investigated range of τ is adjusted to $[0.2–2.0]\,s$ to embrace the whole motor imagery period, while the slicing overlap $\Delta\tau$ is empirically adjusted to $100\,ms$.

After the instance-based feature mapping to the similarity space, a sparse regression is performed in order to select features and theregularization parameter is tuned according to the maximum classification accuracy of training for each subject through a thorough search. Further, due to the MIL algorithms often provide a large number of redundant or irrelevant features, which limits their application for large datasets, we include L1 regularization-based feature selection (Section 5.1.3) that improves further the performed accuracy, increasing the subject performance with a low signal-to-noise ratio. Finally, the classification stage is developed using a SVM with gaussian kernel through 10-fold cross-validation scheme, and our methodology is compared against three the state-of-the-art methods: TSGSP [Zhang et al., 2018b], SFBCSP [Zhang et al., 2015], and SFTRFRC [Miao et al., 2017].

6.2. Results and Discussion

The proposed MILCPS is based on both instance and feature selection as in Section 6.1.3 and Section 2.1, under a multiple instance learning framework. The importance of the two mentioned stages can be highlighted by analyzing the behavior of the subjects in time-frequency intervals shown in the Figure **6-2**. The instance selection and the feature selection with and without the instance selection are presented in Figure **6-2** for the subjects 3 (a, b, c) and 5 (d, e, f) respectively. The analysis shows a grid of 16 frequency bands for 3 time windows where the lighter colors represent the most often selected components in both instance selection and feature selection stages. The subject A05T (d, e, f) shows activity during the two final segments of time in frequencies corresponding to high β band that is present in some subjects as is presented in [Ahn and Jun, 2015]. These results evidencing that significant features appear at the end of MI interval and the importance of the time-frequency segmentation and instance-features selection presented. According to the subject A03T (a, b, c), while the obtained classification accuracy by TSGSP, SFBCSP, SFTOFSRC and the proposed MILCSP method (see Table **6-1**) is high, the selected features shows activity in the mu(μ) which is mainly linked to the motor cortex activity.

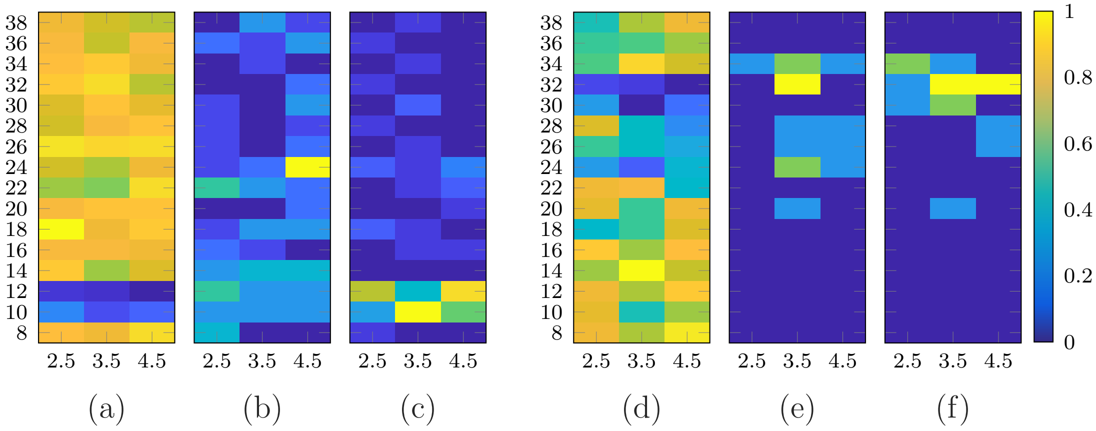


Figure **6-2**: Relative frequency of the time-frequency components for instance selection (a, d),feature selection without (b, e) and with (c, f) instance selection .

Besides, in MIFB we investigate the influence of τ (the relationship window between time series) on the produced spectro-temporal dynamics that are the most relevant in discriminating between classes. The upper plots in figure **6-3** display the accuracy achieved by estimating the higher-level structure similarity for each one of the available instances of time-frequency representation, that is, all combination of pq (see Algorithm 1), for which the graphical meaning is drawn to get a better understanding of the proposed bag expansion. Because of the symmetry of matrix accuracy in τ, only its upper part is reported and ranked in decreasing order of accuracy achieved by each subject.

In terms of distinguishing between MI tasks, the measured similarity matrix values allow extracting the higher-level structure dynamics of relevance, facilitating an accuracy enhancement over a wide range of τ. Further, we rely on the LASSO fits estimated by the feature selection task in Section 2.1, which increases the model interpretability by eliminating irrelevant variables that are not associated with the response variable and this way also reducing the overfitting [Roth, 2004, Fonti and Belitser, 2017]. Namely, besides information about feature relevance, some indication is given about the degree up to which a feature is relevant or can be replaced by others. Nevertheless, the solutions tend to be not consistent estimations of the underlying “true” weight vector computed through Lasso regularization, regarding its exact value as quoted in [Pfannschmidt et al., 2019]. As shown in the bottom rows of figure **6-3**, we compute the normalized absolute Lasso weights at each time instant. It is worth noting that computation of LASSO fits directly through the features calculated from CSP as in [Miao et al., 2017], does not provide an understandable representation of brain dynamics at each time instant since the optimal vector of Lasso weights holds the extracted CSP features, but contributing across the whole MI period. Instead, the learned sparse vector from the bag-of-instances representations reveals a dynamic behavior that somehow resembles an elicited ERP waveform, rising in the beginning and declining in the closing periods.

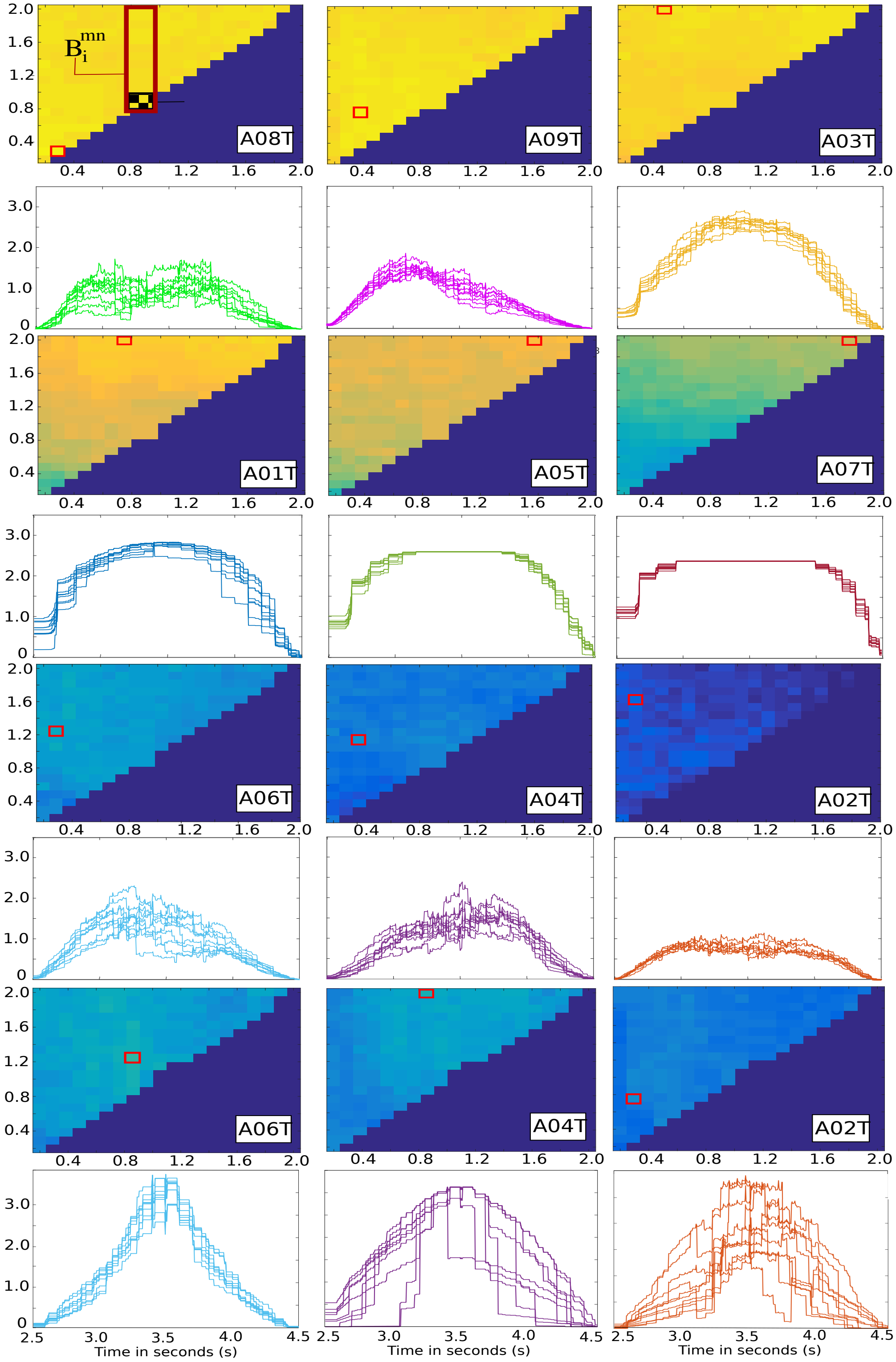


Figure **6-3**: Sparse feature selection of expanded bags representations for each subject. Top rows: Accuracy performed by window lengths of instances. Bottom rows: Temporal dynamics of the absolute LASSO weights performed within the motor imagery period. Each time series is a cross-validated fold.

The two bottom lines in figure **6-3** represent the rise of accuracy performed by optimizing the higher-level structure similarity of bag-based representation through the procedure in Section 5.1.3. As a result, the LASSO fits may increase prominently and remain constant over extensive window lengths. Table **6-1** displays the accuracy of the compared LASSO-based CSP algorithms reported for each subject, showing that all of them are overperformed by the proposed MILCSP and MIFB methods using an SVM instance classifier, at least, in terms of the average across the whole subject set. Furthermore, by optimizing the higher-level structure similarity of bag-based representation using feature selection in Section 5.1.3, the performed accuracy increases further, improving most of the subject with a low signal-to-noise ratio. Additionally, table **6-2** shows the computational cost of our approaches highlighting the improvement provided by the intensive version.

Subject	TSGSP	SFBCSP	SFTOFSRC	*MILCSP*	*MIFB*	*MIFB**
A08T	95.8	99.3	98.5	99.3	**99.6**	98.9
A09T	81.3	**98.3**	95.7	96.6	97.8	98.2
A03T	93.8	98.5	97.8	96.3	97.7	**99.2**
A01T	87.0	91.3	92.7	94.2	96.0	**97.0**
A05T	90.4	90.0	92.2	89.9	93.3	**95.3**
A07T	91.4	93.2	81.2	75.2	90.0	**95.9**
A06T	63.9	67.3	64.8	72.6	**75.9**	74.2
A04T	74.3	64.4	67.4	**74.4**	70.7	69.5
A02T	64.7	57.9	60.9	63.9	64.6	**66.6**
Average	82.5	82.7	83.5	**84.7±6**	**87.3±4.8**	**88.3±5.1**

Table **6-1**: Comparison of accuracy performed by the proposed MILCSP and MIFB methods against TSGSP [Zhang et al., 2018b], SFBCSP [Zhang et al., 2015] and SFTOF-SRC [Miao et al., 2017], using an SVM bag classifier. The best accuracy of each subject is marked in black. Notation * stands for the MIFB method that includes the instance selection proposed in Section 5.1.3, on procedure additionally.

	MILCSP	*MIEB*	*MIEB**	*Intensive*
Time×Subj	$N_\tau \times 12$h	36h	$N_\lambda \times 36$h	**1h**
Complexity	$O(N_\tau n^2)$	$O(n^2)$	$O(N_\lambda n^2)$	$\boldsymbol{O(n)}$
Accuracy	84.7±6	87.3±4.8	88.3±5.1	**87.3±5.1**

Cuadro **6-2**: Computational cost of our approaches including intensive version

7 Conclusions and Future Work

7.1. Conclusions

- We propose a classification methodology that integrates MIL-based similarity representation and multiple kernel learning for dealing with image time-series and enhancing models interpretability, named *MILMKL*. The main advantage of our method is that it allows combining various sets of features extracted from the analysis data in order to use several information sources to improve the performance of the models for classification tasks, even for multi-class tasks. Another benefit of kernel combinations is the provided discrimination ranking of the analysis structures (Bands for satellite images and Brain structures for AD prediction), allowing to enhance the interpretability of the extracted features.

- We propose *MILRES* methodology, a temporal analysis of EEG signals based on a multi-instance framework including LASSO regularization for feature selection at the instance level and the embedded instance selection for the similarity representation of trials. Joining the MIL representation and the feature selection possess two main advantages: First, the instance representation as an overlapping temporal segmentation allows each time-segment to be analyzed individually to account for the appearance of stimulus response. Second, the feature selection allows identifying information improving the discrimination of emotional states with interpretability from the brain physiology.

- Intending to build patterns of neural activation with improved class separation, we propose an instance-based enhanced bags representation with two approaches, *MILCSP* and *MIFB*. To this purpose, exploiting the baseline short-time CSP feature extraction, two approaches are introduced, the first based on time-frequency instance selection, and the expanded bag representation, which combine instances lasting more extensive window lengths. The proposed enhanced time series representation promotes in motor imagery the following two contributions:
  - *Accuracy improvement of bi-conditional tasks*. Effectiveness of the conventional CSP extraction is very affected by the time-frequency window of EEG segments due to the significant inter- and intra-subject variation. To cope with this issue,

we propose to build enhance the bag representations using the instance selected or a combination of instances calculated from several windows length. As a result, the designed bag using the higher-level structures allows capturing the structural dynamics of EEG data more carefully and therefore increases the classification accuracy, overperforming the baseline sparse CSP-based systems reported in the literature.

  - *Better understanding of dynamic brain behavior.* A better understanding of dynamic brain behavior through the learned LASSO fits. For the designed an enhanced bag representation, the model interpretability is increased since the sparse feature selection eliminates irrelevant variables, which are not associated with the response variable. Thus, the learned sparse vector from the MIL-based similarity representations reveals a dynamic behavior that somehow resembles an elicited ERP waveform, rising in the beginning and declining in the closing periods.

7.2. Future Work

- As a future research direction, We plan to improve the kernel representation through supervised optimization methods in order to improve the combination. Further, we also plan to extend our approach by increasing the number of sources for information extraction to improve relevance analysis. Finally, we intend to test our methodology with other databases of different tasks, such as the analysis of multimodal medical records.

- As a future work, we plan to extend the MIL framework to multi-class and regression problems for modeling emotional dimensions at a finer level. Also, we will work on a temporal relevance analysis that provides information about the stimulus-response for education and neuromarketing applications.

- As future work, to improve the robustness across trials, the authors plan to explore more powerful bag-of-patterns representations, using the disgregation/selection of filter-banked components and testing other distances between high-level structures of time series. Further, the computational complexity must be minimized, encouraging validation of the proposed methodology on more extensive EEG databases with a higher number of electrodes, multiple labels, and larger populations. Besides, in order to identify intricate nonlinear structures, we propose to connect neural networks to the representation based on MIL.

Bibliography

[Aghabozorgi and Teh, 2014] Aghabozorgi, S. and Teh, Y. W. (2014). Stock market co-movement assessment using a three-phase clustering method. *Expert Systems with Applications*, 41(4):1301–1314.

[Ahn and Jun, 2015] Ahn, M. and Jun, S. C. (2015). Performance variation in motor imagery brain–computer interface: a brief review. *Journal of neuroscience methods*, 243:103–110.

[Alazrai et al., 2018] Alazrai, R., Homoud, R., Alwanni, H., and Daoud, M. (2018). Eeg-based emotion recognition using quadratic time-frequency distribution. *Sensors*, 18(8):2739.

[Álvarez-Meza et al., 2014] Álvarez-Meza, A. M., Cárdenas-Peña, D., and Castellanos-Dominguez, G. (2014). Unsupervised kernel function building using maximization of information potential variability. In Bayro-Corrochano, E. and Hancock, E., editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 335–342, Cham. Springer International Publishing.

[Bagnall et al., 2003] Bagnall, A., Janacek, G., and Zhang, M. (2003). Clustering time series from mixture polynomial models with discretised data.

[Beheshti et al., 2017] Beheshti, I., Demirel, H., and Matsuda, H. (2017). Classification of alzheimer's disease and prediction of mild cognitive impairment-to-alzheimer's conversion from structural magnetic resource imaging using feature ranking and a genetic algorithm. *Computers in Biology and Medicine*, 83:109 – 119.

[Bermpohl et al., 2006] Bermpohl, F., Pascual-Leone, A., Amedi, A., Merabet, L. B., Fregni, F., Gaab, N., Alsop, D., Schlaug, G., and Northoff, G. (2006). Attentional modulation of emotional stimulus processing: an fmri study using emotional expectancy. *Human brain mapping*, 27(8):662–677.

[Berndt and Clifford, 1994] Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA.

[Blankertz et al., 2008] Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., and Müller, K. R. (2008). Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Processing Magazine*, 25(1):41–56.

[Buckner et al., 2004] Buckner, R. L., Head, D., Parker, J., Fotenos, A. F., Marcus, D., Morris, J. C., and Snyder, A. Z. (2004). A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: Reliability and validation against manual measurement of total intracranial volume. *NeuroImage*, 23(2):724–738.

[Cai and Ng, 2004] Cai, Y. and Ng, R. (2004). Indexing spatio-temporal trajectories with chebyshev polynomials. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pages 599–610. ACM.

[Chakraborty et al., 1992] Chakraborty, K., Mehrotra, K., Mohan, C. K., and Ranka, S. (1992). Forecasting the behavior of multivariate time series using neural networks. *Neural networks*, 5(6):961–970.

[Chen and Ng, 2004] Chen, L. and Ng, R. (2004). On the marriage of lp-norms and edit distance. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 792–803. VLDB Endowment.

[Chen et al., 2017] Chen, P.-Y., Chen, C.-C., Yang, C.-H., Chang, S.-M., and Lee, K.-J. (2017). milr: Multiple-instance logistic regression with lasso penalty. *The R Journal*, 9(1):446–457.

[Chen et al., 2016] Chen, R.-B., Cheng, K.-H., Chang, S.-M., Jeng, S.-L., Chen, P.-Y., Yang, C.-H., and Hsia, C.-C. (2016). Multiple-instance logistic regression with lasso penalty. *arXiv preprint arXiv:1607.03615*.

[Chen et al., 2006] Chen, Y., Bi, J., and Wang, J. Z. (2006). Miles: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1931–1947.

[Chen et al., 2018] Chen, Y., Lu, D., Moran, E., Batistella, M., Dutra, L. V., Sanches, I. D., da Silva, R. F. B., Huang, J., Luiz, A. J. B., and de Oliveira, M. A. F. (2018). Mapping croplands, cropping patterns, and crop types using modis time-series data. *International journal of applied earth observation and geoinformation*, 69:133–147.

[Cheng et al., 2015] Cheng, B., Liu, M., Zhang, D., Munsell, B. C., and Shen, D. (2015). Domain Transfer Learning for MCI Conversion Prediction. *IEEE Transactions on Biomedical Engineering*, 62(7):1805–1817.

[Cortes et al., 2012] Cortes, C., Mohri, M., and Rostamizadeh, A. (2012). Algorithms for Learning Kernels Based on Centered Alignment. *Journal of Machine Learning Research*, 13:795–828.

[Deppe et al., 2005] Deppe, M., Schwindt, W., Kugel, H., Plassmann, H., and Kenning, P. (2005). Nonlinear responses within the medial prefrontal cortex reveal when specific implicit information influences economic decision making. *Journal of Neuroimaging*, 15(2):171–182.

[Dietterich et al., 1997] Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71.

[Faloutsos et al., 1994] Faloutsos, C., Ranganathan, M., and Manolopoulos, Y. (1994). *Fast subsequence matching in time-series databases*, volume 23. ACM.

[Fonti and Belitser, 2017] Fonti, V. and Belitser, E. (2017). Feature selection using LASSO. *Amsterdam Research Paper in Business Analytics*.

[Foody, 2002] Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote sensing of environment*, 80(1):185–201.

[Friedman et al., 2010] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

[Fu et al., 2010] Fu, Z., Robles-Kelly, A., and Zhou, J. (2010). Milis: Multiple instance learning with instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):958–977.

[Fujita et al., 2012] Fujita, A., Severino, P., Kojima, K., Sato, J. R., Patriota, A. G., and Miyano, S. (2012). Functional clustering of time series gene expression data by granger causality. *BMC systems biology*, 6(1):137.

[Gillath et al., 2005] Gillath, O., Bunge, S. A., Shaver, P. R., Wendelken, C., and Mikulincer, M. (2005). Attachment-style differences in the ability to suppress negative thoughts: exploring the neural correlates. *Neuroimage*, 28(4):835–847.

[Golay et al., 1998] Golay, X., Kollias, S., Stoll, G., Meier, D., Valavanis, A., and Boesiger, P. (1998). A new correlation-based fuzzy logic clustering algorithm for fmri. *Magnetic Resonance in Medicine*, 40(2):249–260.

[Gönen and Alpaydın, 2011] Gönen, M. and Alpaydın, E. (2011). Multiple Kernel Learning Algorithms. *Journal of Machine Learning Research*, 12:2211–2268.

[Gönen and Alpaydın, 2013] Gönen, M. and Alpaydın, E. (2013). Localized algorithms for multiple kernel learning. *Pattern Recognition*, 46(3):795–807.

[Guam and Jiang, 2007] Guam, H.-S. and Jiang, Q.-S. (2007). Cluster financial time series for portfolio. In *2007 international conference on wavelet analysis and pattern recognition*, volume 2, pages 851–856. IEEE.

[Gullo et al., 2012] Gullo, F., Ponti, G., Tagarelli, A., Tradigo, G., and Veltri, P. (2012). A time series approach for clustering mass spectrometry data. *Journal of Computational Science*, 3(5):344–355.

[Hanssens, 1980] Hanssens, D. M. (1980). Market response, competitive behavior, and time series analysis. *Journal of Marketing Research*, 17(4):470–485.

[Hu et al., 2016] Hu, K., Wang, Y., Chen, K., Hou, L., and Zhang, X. (2016). Multi-scale features extraction from baseline structure MRI for MCI patient classification and AD early diagnosis. *Neurocomputing*, 175(PartA):132–145.

[Iglesias and Kastner, 2013] Iglesias, F. and Kastner, W. (2013). Analysis of similarity measures in times series clustering for the discovery of building energy patterns. *Energies*, 6(2):579–597.

[Ji et al., 2013] Ji, M., Xie, F., and Ping, Y. (2013). A dynamic fuzzy cluster algorithm for time series. In *Abstract and Applied Analysis*, volume 2013. Hindawi.

[Jiang et al., 2014] Jiang, Z., Bai, W., and Bin, W. (2014). Social network users clustering based on multivariate time series of emotional behavior. *The Journal of China Universities of Posts and Telecommunications*, 21(2):21–31.

[Kakizawa et al., 1998] Kakizawa, Y., Shumway, R. H., and Taniguchi, M. (1998). Discrimination and clustering for multivariate time series. *Journal of the American Statistical Association*, 93(441):328–340.

[Kastens et al., 2017] Kastens, J. H., Brown, J. C., Coutinho, A. C., Bishop, C. R., and Esquerdo, J. C. D. (2017). Soy moratorium impacts on soybean and deforestation dynamics in mato grosso, brazil. *PloS one*, 12(4):e0176168.

[Kawagoe and Ueda, 2002] Kawagoe, K. and Ueda, T. (2002). A similarity search method of time series data with combination of fourier and wavelet transforms. In *Proceedings Ninth International Symposium on Temporal Representation and Reasoning*, pages 86–92. IEEE.

[Keogh, 2004] Keogh, E. (2004). Data mining and machine learning in time series databases. *Tutorial in ICML.*

[Keogh et al., 2004] Keogh, E., Lonardi, S., and Ratanamahatana, C. A. (2004). Towards parameter-free data mining. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 206–215. ACM.

[Keogh and Pazzani, 1998] Keogh, E. J. and Pazzani, M. J. (1998). An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. In *Kdd*, volume 98, pages 239–243.

[Koelstra et al., 2012] Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., and Patras, I. (2012). Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31.

[Korn et al., 1997] Korn, F., Jagadish, H. V., and Faloutsos, C. (1997). Efficiently supporting ad hoc queries in large datasets of time sequences. In *Acm Sigmod Record*, volume 26, pages 289–300. ACM.

[Kumar et al., 2002] Kumar, M., Patel, N. R., and Woo, J. (2002). Clustering seasonality patterns in the presence of errors. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 557–563. ACM.

[Kurbalija et al., 2012] Kurbalija, V., von Bernstorff, C., Burkhard, H.-D., Nachtwei, J., Ivanović, M., and Fodor, L. (2012). Time-series mining in a psychological domain. In *Proceedings of the Fifth Balkan Conference in Informatics*, pages 58–63. ACM.

[Lane et al., 1999] Lane, R. D., Chua, P. M., and Dolan, R. J. (1999). Common effects of emotional valence, arousal and attention on neural activation during visual processing of pictures. *Neuropsychologia*, 37(9):989–997.

[Lang, 1995] Lang, P. J. (1995). The emotion probe: studies of motivation and attention. *American psychologist*, 50(5):372.

[Latecki et al., 2005] Latecki, L. J., Megalooikonomou, V., Wang, Q., Lakaemper, R., Ratanamahatana, C. A., and Keogh, E. (2005). Elastic partial matching of time series. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 577–584. Springer.

[Li et al., 2009] Li, W.-J. et al. (2009). Mild: Multiple-instance learning via disambiguation. *IEEE Transactions on Knowledge and Data Engineering*, 22(1):76–89.

[Lin et al., 2012] Lin, J., Khade, R., and Li, Y. (2012). Rotation-invariant similarity in time series using bag-of-patterns representation. *Journal of Intelligent Information Systems*, 39(2):287–315.

[Lin and Li, 2009] Lin, J. and Li, Y. (2009). Finding structural similarity in time series data using bag-of-patterns representation. In *International conference on scientific and statistical database management*, pages 461–477. Springer.

[Liu et al., 2014] Liu, S., Maharaj, E. A., and Inder, B. (2014). Polarization of forecast densities: A new approach to time series classification. *Computational Statistics & Data Analysis*, 70:345–361.

[Maron and Lozano-Pérez, 1998] Maron, O. and Lozano-Pérez, T. (1998). A Framework for Multiple-Instance Learning. In Jordan, M. I., Kearns, M. J., and Solla, S. A., editors, *Advances in Neural Information Processing Systems 10*, pages 570–576. MIT Press.

[Miao et al., 2017] Miao, M., Wang, A., and Liu, F. (2017). A spatial-frequency-temporal optimized feature sparse representation-based classification method for motor imagery EEG pattern recognition. *Medical and Biological Engineering and Computing*, 55(9):1589–1603.

[Moradi et al., 2015] Moradi, E., Pepe, A., Gaser, C., Huttunen, H., and Tohka, J. (2015). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *NeuroImage*, 104:398–412.

[Mühling et al., 2012] Mühling, M., Ewerth, R., Zhou, J., and Freisleben, B. (2012). Multimodal video concept detection via bag of auditory words and multiple kernel learning. In *International Conference on Multimedia Modeling*, pages 40–50. Springer.

[Nanopoulos et al., 2001] Nanopoulos, A., Alcock, R., and Manolopoulos, Y. (2001). Feature-based classification of time-series data. *International Journal of Computer Research*, 10(3):49–61.

[Pelletier et al., 2003] Pelletier, M., Bouthillier, A., Lévesque, J., Carrier, S., Breault, C., Paquette, V., Mensour, B., Leroux, J.-M., Beaudoin, G., Bourgouin, P., et al. (2003). Separate neural circuits for primary emotions? brain activity during self-induced sadness and happiness in professional actors. *Neuroreport*, 14(8):1111–1116.

[Peugh and Enders, 2004] Peugh, J. L. and Enders, C. K. (2004). Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. *Review of Educational Research*, 74(4):525–556.

[Pfannschmidt et al., 2019] Pfannschmidt, L., Jakob, J., Biehl, M., Tino, P., and Hammer, B. (2019). Feature relevance bounds for ordinal regression. *CoRR*, 1902.07662.

[Picoli et al., 2018] Picoli, M. C. A., Camara, G., Sanches, I., Simões, R., Carvalho, A., Maciel, A., Coutinho, A., Esquerdo, J., Antunes, J., Begotti, R. A., et al. (2018). Big earth observation time series analysis for monitoring brazilian agriculture. *ISPRS journal of photogrammetry and remote sensing*, 145:328–339.

[Pyatnitskiy et al., 2014] Pyatnitskiy, M., Mazo, I., Shkrob, M., Schwartz, E., and Kotelnikova, E. (2014). Clustering gene expression regulators: new approach to disease subtyping. *PLoS One*, 9(1):e84955.

[Rämä et al., 2001] Rämä, P., Martinkauppi, S., Linnankoski, I., Koivisto, J., Aronen, H. J., and Carlson, S. (2001). Working memory of identification of emotional vocal expressions: an fmri study. *Neuroimage*, 13(6):1090–1101.

[Ratanamahatana and Keogh, 2004] Ratanamahatana, C. A. and Keogh, E. (2004). Making time-series classification more accurate using learned constraints. In *Proceedings of the 2004 SIAM international conference on data mining*, pages 11–22. SIAM.

[Ritter et al., 2015] Ritter, K., Schumacher, J., Weygandt, M., Buchert, R., Allefeld, C., and Haynes, J.-D. (2015). Multimodal prediction of conversion to alzheimer's disease based on incomplete biomarkers. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(2):206–215.

[Roth, 2004] Roth, V. (2004). The generalized LASSO. *IEEE Transactions on Neural Networks*, 15(1):16–28.

[Sadahiro and Kobayashi, 2014] Sadahiro, Y. and Kobayashi, T. (2014). Exploratory analysis of time series data: Detection of partial similarities, clustering, and visualization. *Computers, Environment and Urban Systems*, 45:24–33.

[Schreiber and Schmitz, 1997] Schreiber, T. and Schmitz, A. (1997). Discrimination power of measures for nonlinearity in a time series. *Physical Review E*, 55(5):5443.

[Shen et al., 2010] Shen, L., Qi, Y., Kim, S., Nho, K., Wan, J., Risacher, S. L., and Saykin, A. J. (2010). Sparse bayesian learning for identifying imaging biomarkers in ad prediction. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6363 LNCS(PART 3):611–618.

[Shin et al., 2012] Shin, Y., Lee, S., Lee, J., and Lee, H. N. (2012). Sparse representation-based classification scheme for motor imagery-based brain-computer interface systems. *Journal of Neural Engineering*, 9(5).

[Shumway, 2003] Shumway, R. H. (2003). Time-frequency clustering and discriminant analysis. *Statistics & probability letters*, 63(3):307–314.

[Stock and Watson, 1988] Stock, J. H. and Watson, M. W. (1988). Variable trends in economic time series. *Journal of economic perspectives*, 2(3):147–174.

[Theodoridis and Koutroumbas, 2009] Theodoridis, S. and Koutroumbas, K. (2009). Pattern recognition. 2003. *Elsevier Inc.*

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

[Verbesselt et al., 2010] Verbesselt, J., Hyndman, R., Newnham, G., and Culvenor, D. (2010). Detecting trend and seasonal changes in satellite image time series. *Remote sensing of Environment*, 114(1):106–115.

[Wang et al., 2013a] Wang, J., Liu, P., She, M. F., Nahavandi, S., and Kouzani, A. (2013a). Bag-of-words representation for biomedical time series classification. *Biomedical Signal Processing and Control*, 8(6):634–644.

[Wang and Zucker, 2000] Wang, J. and Zucker, J.-D. (2000). Solving multiple-instance problem: A lazy learning approach.

[Wang et al., 2006] Wang, X., Smith, K., and Hyndman, R. (2006). Characteristic-based clustering for time series data. *Data mining and knowledge Discovery*, 13(3):335–364.

[Wang et al., 2013b] Wang, X., Wang, B., Bai, X., Liu, W., and Tu, Z. (2013b). Max-margin multiple-instance dictionary learning. In *International Conference on Machine Learning*, pages 846–854.

[Wismüller et al., 2002] Wismüller, A., Lange, O., Dersch, D. R., Leinsinger, G. L., Hahn, K., Pütz, B., and Auer, D. (2002). Cluster analysis of biomedical image time-series. *International Journal of Computer Vision*, 46(2):103–128.

[Wright et al., 2009] Wright, J., Yang, a. Y., Ganesh, a., Sastry, S. S., and Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 31(2):210–227.

[Xiang et al., 2014] Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P. M., Ye, J., Initiative, A. D. N., et al. (2014). Bi-level multi-source learning for heterogeneous block-wise missing data. *NeuroImage*, 102:192–206.

[Zhang et al., 2018a] Zhang, X., Wang, Y., Zhao, S., Liu, J., Pan, J., Shen, J., and Ding, T. (2018a). Emotion recognition based on electroencephalogram using a multiple instance learning framework. In *International Conference on Intelligent Computing*, pages 570–578. Springer.

[Zhang et al., 2018b] Zhang, Y., Nam, C. S., Zhou, G., Jin, J., Wang, X., and Cichocki, A. (2018b). Temporally constrained sparse group spatial patterns for motor imagery BCI. *IEEE Transactions on Cybernetics*, 49(9):1–11.

[Zhang et al., 2015] Zhang, Y., Zhou, G., Jin, J., Wang, X., and Cichocki, A. (2015). Optimizing spatial patterns with sparse filter bands for motor-imagery based brain-computer interface. *Journal of Neuroscience Methods*, 255:85–91.

[Zheng et al., 2017] Zheng, W.-L., Zhu, J.-Y., and Lu, B.-L. (2017). Identifying stable patterns over time for emotion recognition from eeg. *IEEE Transactions on Affective Computing*.

[Zhou et al., 2012] Zhou, Z.-H., Zhang, M.-L., Huang, S.-J., and Li, Y.-F. (2012). Multi-instance multi-label learning. *Artificial Intelligence*, 176(1):2291–2320.

[Zhu et al., 2006] Zhu, Q., Yeh, M.-C., and Cheng, K.-T. (2006). Multimodal fusion using learned text concepts for image categorization. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 211–220. ACM.

[Zhuang et al., 2017] Zhuang, N., Zeng, Y., Tong, L., Zhang, C., Zhang, H., and Yan, B. (2017). Emotion recognition from eeg signals using multidimensional information in emd domain. *BioMed research international*, 2017.