Multiple Instance Learning for Emotion Recognition using Physiological Signals

Luca Romeo, Andrea Cavallo, Lucia Pepa, Nadia Bianchi-Berthouze, Massimiliano Pontil

Abstract—The problem of continuous emotion recognition has been the subject of several studies. The proposed affective computing approaches employ sequential machine learning algorithms for improving the classification stage, accounting for the time ambiguity of emotional responses. Modeling and predicting the affective state over time is not a trivial problem because continuous data labeling is costly and not always feasible. This is a crucial issue in real-life applications, where data labeling is sparse and possibly captures only the most important events rather than the typical continuous subtle affective changes that occur. In this work, we introduce a framework from the machine learning literature called Multiple Instance Learning, which is able to model time intervals by capturing the presence or absence of relevant states, without the need to label the affective responses continuously (as required by standard sequential learning approaches). This choice offers a viable and natural solution for learning in a weakly supervised setting, taking into account the ambiguity of affective responses. We demonstrate the reliability of the proposed approach in a gold-standard scenario and towards real-world usage by employing an existing dataset (DEAP) and a purposely built one (Consumer). We also outline the advantages of this method with respect to standard supervised machine learning algorithms.

Index Terms—Emotion Recognition, Multiple Instance Learning, Time Ambiguity, Physiological signals, Support Vector Machine, Diverse Density

1 Introduction

Recent years have witnessed a growing number of datasets ([1]–[11]) used to train learning systems capable of recognising automatically the affective states of people engaged with technology while accomplishing a given task. The aim of these systems is to build a technology able to provide personalised support or experiences while taking into account people's psychological needs or enabling an affective communication channel between people and artificial agents.

Most of the initial datasets were built using acted expressions or through well-controlled studies, where specific emotions were elicited [1], [2]. This type of process enabled a full labeling of each frame of the dataset.

As affective computing continues to mature, naturalistic datasets are becoming available. Some of these datasets are being fully labeled manually by experts [3], [7], [12] or naive raters (AVEC, [8]–[10], [13]), and different forms of inter-rater reliability have been used. Despite the importance that manual labeling has had in the field of machine learning, in many cases, such labeling is not feasible or scalable. This is more and more the case as naturalistic data are gathered in real-life situations where people do not want to be video recorded (to enable labeling) for privacy reasons or where recording is not feasible (e.g. on the move rather than in front of a camera) [14]. In addition, the continuous monitoring of people might result in datasets of unlimited dimensions, making manual labeling too costly, if not impossible.

Some affective states are also difficult to capture by an external observer. For example, in [15], many physiotherapists did not consider it feasible to judge the intensity of pain in patients, as people respond to and express pain levels in different ways and pain may be affected by other factors, such as anxiety (e.g. in the context of chronic pain rather than acute pain), individual pain threshold and individual pain tolerance. Self-report is often a solution to label naturalistic datasets [4]–[6], [16], [17]; however, it is not always possible to continuously interrupt people to self-report their states. This could create frustrating disruptions to people's tasks (e.g. [12]) and in the case of patients it may even lead to unhealthy catastrophising on one's condition [18].

These issues have led to a growing number of datasets that are only sparsely labeled [4]–[6], [11], [17], [19], [20] or in which some of the states are sparsely annotated. External raters (e.g. an observer) are often asked to focus on or detect what is salient in a particular situation, rather than to identify any change in emotional states [21]. Detecting salient states is important, e.g. physiotherapists may focus on detecting increased anxiety in their patients during an exercise rather than all the specific affective changes that may occur. Furthermore, this reduces annotation cost and at the same time it increases inter-rater reliability [15]. Finally, by focusing on specific states, the raters' task becomes more viable [4]. Indeed, by constraining the number and types of states to be attended to, raters can be trained to discriminate between them [22].

Sparse annotation is also the outcome of self-reporting. The constrains discussed above for the external-observer apply to self-reporting too as self-reports can be considered as a case of self-observation. In addition, for the reasons discussed above, self-

- *L. Romeo is with Department of Information Engineering Università Politecnica delle Marche, Ancona, Italy, and Cognition, Motion and Neuroscience and Computational Statistics and Machine Learning, Istituto Italiano di Tecnologia, Genova, Italy.*
A. Cavallo is with Dipartimento di Psicologia, Università di Torino, and C'MoN Unit, Istituto Italiano di Tecnologia, Genoa, Italy.
L. Pepa is with Department of Information Engineering Università Politecnica delle Marche, Ancona, Italy
N. Bianchi-Berthouze is with the UCL Interaction Centre, University College London, UK.
M. Pontil is with Istituto Italiano di Tecnologia, Genova, Italy, and Department of Computer Science, University College London, UK.
Corresponding to L. Romeo (E-mail: l.romeo@univpm.it).

reports can only be gathered sparsely. For example, labels are often self-reported only at the end of short events, e.g. a video viewing [4], [5] or a game (e.g. affective state of entertainment [19])). In longer data collection (e.g. over a day or more), opportunistic sampling is often the only viable practice. It consists of sparsely reporting the affective states; participants are time to time asked to report their affective state or the most salient one within the last period of time. The time when to self-report the affective state is generally triggered by either a specific person's behaviour (e.g. during the use of an app) or at specific moments during the day (e.g. before going to sleep) and in some cases also at random moments or when the person feels to report. In all these cases, the labels are then used to classify entire segments of behavioural or physiological data gathered around the time the label was reported.

Label sparsity creates challenges for the machine learning community: labels represent only those captured and possibly most salient affective events within the labeled time window [4]–[6] rather than the typical continuous subtle affective changes that may occur within it [8]–[10]. When these affective events represent only a small part of the window, their labels become noisy as they are instead modeled as if they were representative of each frame within the labeled period.

Starting from this motivation, the aim of the present work is to model the time intervals that better capture the presence of the self-reported emotions, without the need to label the affective states continuously (as required by sequential learning approaches). In this context, we introduce a framework from the machine learning literature called Multiple Instance Learning (MIL) [23]. This framework offers a viable and natural solution for learning in a weakly supervised setting by taking into account the temporal ambiguity of the emotional response. In such a weakly supervised setting, each label only specifies the presence or absence of a specific affective state within the given period without localising when it exactly occurred or was expressed. This work contributes to the *affective computing field* for the following reasons:

- it introduces the MIL framework for capturing the temporal ambiguity of the occurrence of an affective state [24], [25]. The current approach seeks to detect the most prominent emotional events rather than the continuous affective changes that occur within a labeled period.
- it proposes an application of three MIL-based methods for the emotion recognition task and demonstrates a significantly improved prediction performance over the best standard supervised machine learning approaches.
- it measures and demonstrates the reliability of the proposed approach in real-world usage, where the unobtrusiveness of the device and the lower *accuracy* of the sensors provide a challenging scenario to machine learning.

The paper is organised as follows. In Section 2, we review previous related works. The MIL framework is outlined in Section 3. Section 4 describes the two datasets employed for the analysis, and experimental results are presented in Section 5. The conclusions and future research directions are presented in Section 6.

2 Related works

MIL has attracted much attention from the research community, especially in recent years, as the amount of data needed to learn difficult tasks has increased exponentially [26]. The annotation of large datasets requires a highly costly effort (not always feasible) to label them correctly and continuously. MIL can reduce this burden, as it is able to operate in a weakly supervised setting [26], [27]. Current MIL methods have been proven useful in a variety of domains, including bioinformatics [23], medical image analysis [28], text processing [29], educational scenarios [30] and object recognition and tracking studies [31].

MIL has been used in previous studies on affective computing, although with different data modalities than those considered in the present paper and, most importantly, with the aim to address a different problem than the temporal sparse labeling. Specific methods based on MIL have been used to model the spatial ambiguity of the emotion in an image [32]–[34] or in text data [35], [36]. Considering that different regions of an image frame may correspond to different emotions or that different regions of an image frame may contribute differently to a general emotion portrayed in the image, the authors in [32]–[34] aim to identify peaks in the spatial dimension of the image frame that correspond to the dominant expression label assigned to it, i.e. the peaks are sparse inside each block of the image frame (intra-frame annotation). Similarly, MIL was used for the affective opinion classification of users' reviews (bags) by assigning the label associated with a review to each sentence (instance) in that review (bag) [35], [36]. Some approaches [34], [37] tried to extend this formulation to model the temporal variation of facial expressions between subsequent image frames, with supervised learning algorithms (e.g. Hidden Markov Model [34] or Support Vector Machine (SVM) [37]) requiring the continuous labeling of each image frame. However, in contrast current approaches [32]–[34], [37], we propose to capture the time ambiguity of the emotion where only weak labels at the sequence level are available (inter-frame annotation), i.e. only time windows are labeled rather than each frame within those time windows. This means that data labeling is sparse and possibly describes only the dominant affective state for each labeled time window.

Most related to our work are the papers [38]–[45] that proposed the application of MIL for capturing the time ambiguity of pain [38]–[42], affective music response [43], behavioural expressions [44] and vocal interaction [45]. As discussed below, the main differences with our work lie in the (i) multiple instance algorithms we propose, (ii) the nature of the employed predictors (e.g. physiological signals rather than facial expressions, music) and (iii) the type of affective response (e.g. affective dimensions rather than pain).

In [38], [39], the MIL-boost method [46] was applied to the problem of automatic pain recognition from videos. They represented each video as a bag containing multiple time segments that are modeled using the MIL extension of the gradient-boosting framework. The authors in [38], [39] encapsulated the temporal dynamics by representing the data not as individual frames but as segments. As we shall see in the experimental section, our SVM-based MIL method performs favourably over MIL boost. A different approach was used in [40], where the authors proposed a structured latent variable model for learning sub-events with weakly supervised data. In [41], [42] instead, the authors imposed an ordinal constraint on the instance labels for modeling the weak relation between instance and sequence labels introducing the Multi-Instance Dynamic Ordinal Random Fields [47] for solving the regression task. The structured models of these studies require multiple parameter vectors that act at different temporal positions, impose learning constraints on the temporal order of such vectors [40] or need to encapsulate dynamic information in undirected graphical regression models [41], [42]. On the contrary, our work

aims to model the most prominent emotional sub-event without assuming an ordinal and defined structure of the instances in the presence of sparse labeling.

MIL was used in [43] to automatically recognise the affective content of a piece of music using a generative approach based on a hierarchical Bayesian model. Each song is associated with a bag, and the temporal audio segments form the corresponding instances. On the contrary, here, we present three discriminative MIL methods that solve the classification problem directly.

Recently, MIL was employed for behavioural coding [44] during problem-solving discussions. The authors' aim in [44] was to reveal the local/global nature of behaviours, estimating the level of ambiguity presented via a particular channel (i.e. acoustic, lexical and visual channels) through the application of the Incremental Diverse Density (DD) method [48], [49], followed by an SVM. Similarly, in [45], the authors proposed different DD classifiers based on expectation maximisation that can be used for affective state recognition in married couples' interactions. Although the methodology applied in [44] and [45] is similar to the methodology presented in this paper (named Expectation Maximization Diverse Density [EMDD]-SVM), we go further by applying three other MIL methodologies. These methods represent natural extensions of standard supervised learning methods, namely SVMs and boosting, whereas the DD-based methods are specific learning algorithms that do not compare favorably in the limiting case of the standard supervised learning setting [29].

Finally, we comment on the differences between our approach and those based on the Bag of Words (BoW) representation, a prominent technique in computer vision and document classification. This methodology represents a given input with the frequency histogram relative to a prescribed set of words. Standard BoW models only focus on the number of words (occurrences) while ignoring the spatial-temporal information within the input. Hence, BoW based models may not be able to localise similar physiological responses occurring at different relative periods of time in an input sequence. To overcome this limitation, a sequential BoW model [50] was proposed for human action classification; this modeling techinque captures the temporal ambiguity by segmenting the entire action into sub-actions. Furthermore, spatiotemporal feature extraction within a BoW framework was proposed by [51] for depression analysis using upper body expressions from video clips. The main differences between the above mentioned approaches and ours lie in: (i) the different nature of the task (they aim to solve a computer vision-based task) and the different nature of predictors (they employed a histogram of spatial-temporal word occurrence for each video clip); (ii) the method proposed by [51] is not tailored to extract the most prominent histogram of head movement within each video. We have provided evidence of how our method can be exploited to localise the most prominent emotional response (see Section 5.5); (iii) the final prediction in [50] was obtained by assigning a certain weight and salience to each sub-action in order to mitigate the time ambiguity. In particular, the weight and salience are determined in advance according to the sub-actions discrimination evaluated by training data. Instead, in our approach, the local emotional response is directly learned by the MIL-based model by capturing the most prominent emotional events.

3 Multiple Instance Learning

The background about the MIL methodology is provided in Section 3.1. The employed notation is described in Section 3.2. The proposed Expectation Maximization Diverse Density – SVM (EMDD-SVM) is described in Section 3.3, while the maximum pattern margin (mi-SVM) and maximum bag margin (MI-SVM) formulation are reported in Section 3.5 and Section 3.6 respectively, starting from the definition of Single Instance Learning-SVM (SIL-SVM) and Normalised Set Kernel-SVM (NSK-SVM) (see Section 3.4).

3.1 Background

In the MIL paradigm, the learner receives a set of *bags* along with the corresponding labels. Each bag contains multiple instances. In this paradigm, the data are assumed to have some ambiguity about how the labels are assigned: a bag is labeled as negative if all of its instances are negative, while it is labeled as positive if there is at least one positive instance. The MIL problem was originally formalised for *drug activity prediction* by [23], where the authors developed a multiple instance algorithm for learning Axis-Parallel Rectangles (APRs). Subsequently, [49] introduced the DD framework for solving the person identification problem. The combination of the EM algorithm with the diverse density algorithm [52] improved the computation time and the robustness against the number of irrelevant features. Furthermore, several adaptations of SVMs were proposed for MIL. In the NSK configuration of [53], a traditional SVM was trained with bags represented as the sum of all its instances, normalised by the L1- or L2-norm, while in the Statistical Kernel set [53], every bag was transformed into a feature vector representation (i.e. the maximum and minimum value across all instances in the bag). Other SVM approaches, called, respectively, mi-SVM and MI-SVM, follow an heuristic approximation outlined in [29]. The mi-SVM aims to label the instances in the positive bags using the learned decision hyperplane. Specifically, if a positive bag contains no instances labeled as positive, the instance that gives the maximum value of the decision function for that bag is relabeled as positive. While in the MI-SVM, for every positive bag, the learned decision function is used to select the bag instance that gives the maximum value.

3.2 Notation

In the following subsections, we propose the application of MIL to the emotion recognition task during a multimedia interaction using the physiological features.

We use the following notation:

- We let N^+ and N^- be the number of positive and negative training bags, respectively.
- We let $B_1^+, \ldots, B_{N^+}^+$ and $B_1^-, \ldots, B_{N^-}^-$ be the set of positive and negative training bags, respectively. We let $\mathscr{B}$ be the training set formed by all such bags.
- We let x_{ij}^+ and x_{ij}^- be the set of vector instances in the i-th positive and negative training bag, respectively, where j is the index running over the instances in each bag.
- We let L be the number of instances for each bag.

In the present study, each bag is assumed to contain the same amount of instances (L). The rationale behind this choice is to provide a naturalistic setting, where each instance is represented by the features computed by the same amount of samples (i.e. fixed window length size). However, the different duration of the video stimuli and the latency of physiological response can lead to analyze a signal with different length for each trial. In this scenario the window size can be adapted according to the different

length of the signal. For example, starting with two different trials with different duration (i.e., 2 min and 4 min), each window size can be changed respectively to 24 s and 48 s in order to retrieve the same number of instances (i.e., L=5) within each Bag across different trials. On the one hand, the different number of samples for each window might lead to bias in the features set. On the other hand, a fixed window size across each trial leads to a different number of instances for each bag. For instance, a window size fixed to 24 s would lead to L=5 and L=10 for a video of duration 2 min and 4 min respectively. In the MIL paradigm each bag B is allowed to have a different size, which means that the L can vary among the bags in the dataset [27]. This formulation allows the feature set to be computed by considering the same amount of samples across different instances without introducing any bias. The proposed MIL-based algorithms (EMDD-SVM, mi-SVM and MI-SVM) readily generalise to a different number of instances for each bag (i.e. L_i^+ and L_i^-) by taking into account the different response latencies of physiological signals for different types of induced stimuli. As confirmation of this point, we have provided the general formulation of the proposed MIL-algorithms.

3.3 EMDD-SVM

The key idea behind EMDD is the DD concept and EM algorithm. DD is a measure of the intersection of the positive bags minus the union of the negative bags. Then, by maximising DD, we can look for both the intersection point (the desired *concept*) and the set of feature weights. Indeed, the DD at a point h in the feature space is a probabilistic measure of both how far the negative instances are from h and how many different positive bags have an instance near h. Specifically, the DD of a particular concept h is defined as:

$$DD(h) \equiv P(h|\mathscr{B}). \tag{1}$$

The application of Bayes Rule leads to the finding of the maximum likelihood estimation:

$$\hat{h} = \arg\max_{h\in H}[P(h|\mathscr{B})] = \arg\max_{h\in H}\left[\frac{P(\mathscr{B}|h)P(h)}{P(\mathscr{B})}\right] \tag{2}$$

where H is the hypothesis space. Then, assuming independence of the bag instances, uniform prior of the instances and reapplying Bayes rule, leads to:

$$\begin{aligned}\hat{h} &= \arg\max_{h\in H}\left[\prod_{i=1}^{N^+}P(B_i^+|h)\prod_{i=1}^{N^-}P(B_i^-|h)\right] \\ &= \arg\max_{h\in h}\left[\prod_{i=1}^{N^+}P(h|B_i^+)\prod_{i=1}^{N^-}P(h|B_i^-)\right].\end{aligned} \tag{3}$$

The posterior probability is estimated using the *noisy-or* approximation [54]:

$$\begin{aligned}P(h|B_i^+) &= P(h|x_{i1}^+,\ldots,x_{iL_i^+}^+) = 1-\prod_{j=1}^{L_i^+}(1-P(h|x_{ij}^+)) \\ P(h|B_i^-) &= P(h|x_{i1}^-,\ldots,x_{iL_i^-}^-) = \prod_{j=1}^{L_i^-}(1-P(h|x_{ij}^-))\end{aligned} \tag{4}$$

where, for every vector x, we defined $P(h|x) = \exp(-\|x-h\|^2)$.

While the learning concept points in the instance space, we can also find the best scaling for each k feature that maximises DD. Then, the Euclidean distance $\|x_{ij}-h\|^2$ becomes the weighted distance $\sum_k s_k^2\|x_{ij}^{(k)}-h^{(k)}\|^2$. The optimisation of DD returns both a location c and a scaling vector s that belong to the hypothesis space H ($H=\{c,s\}$). The employed optimisation procedure is comprised of multiple gradient-based optimizations: a line search and a quasi-Newton search.

The intuition behind the EMDD [52] algorithm is to start with some initial condition of the target point h by trying points on the positive bag. Afterwards in the E-step, the hypothesis h is used to pick one instance from each bag that is most likely to be the one responsible for the label given to the bag. In the second step (M-step), the two-step gradient ascent search [54] of the standard DD algorithm was implemented to find a new h' that maximises $DD(h)$. Because our goal was to classify the self-reported emotional events, categorised in the form of low/high levels of *valence* and *arousal* and triggered by each specific video, we propose an adaptation of the EMDD algorithm. The EMDD algorithm was repeated using a different set of initial instances picked from 10 different positive bags. These 10 positive bags were chosen randomly within the training set. The final optimal-scaled concept point h was then the maximum (in terms of DD) of each scaled concept point computed using a new set of initial values. Then, once the scaled concept has been learned, features based on this concept are used to train a linear SVM model. Nearest concept features define the features mapping for each bag as being the minimum distance of any of the instances in that bag to the scaled concepts as follows:

$$\begin{aligned}\phi(B_i^+) &= \min_{1\le j\le L_i^+}(\sum_k s_k^2\|x_{ij}^{+\,(k)}-\hat{h}^{(k)}\|^2) \\ \phi(B_i^-) &= \min_{1\le j\le L_i^-}(\sum_k s_k^2\|x_{ij}^{-\,(k)}-\hat{h}^{(k)}\|^2).\end{aligned} \tag{5}$$

The optimisation of SVM hyperparameter (i.e. box constraint) was performed by the implementation of a grid search and the optimisation of the *macro-F1* score (F_1) in a nested Cross Validation evaluation according to [55]. The *macro-F1* provides the unweighted mean of the F1 score computed for each label. It is a consistent metric because it does not take into account the potential imbalance between labels. Because of this consistency, the *macro-F1* score is widely used in affective computing fields to evaluate the performance of ML models [4], [56]. For the purpose of our study, the Box Constraint was picked inside the subset $\{0.1, 0.5, 1, 5, 25, 100\}$. The subset was constructed around the sub-optimal value picked by a random search. As evidence of our choice, the best-selected value of the hyperparameter always falls in the middle of the considered range.

3.4 SIL-SVM and NSK-SVM

The SVM aims to choose the decision boundary to maximise the margin, which is defined as the smallest distance between the decision boundary and any of the samples, see [57] for more information. The SIL approach transforms the MIL problem into a standard supervised learning problem by assigning the bag's label to all instances inside the bag. Then, a standard SVM is applied to the collection of labeled instances. That is, it solves the

optimisation problem:

$$\begin{aligned}\min_{\boldsymbol{w},b,\xi} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + C\left(\sum_{i=1}^{N^+}\sum_{j=1}^{L_i^+}\xi_{ij}^+ + \sum_{i=1}^{N^-}\sum_{j=1}^{L_i^-}\xi_{ij}^-\right)\\ \text{s.t.} \quad & \\ & \boldsymbol{w}^T x_{ij}^+ + b \geq 1 - \xi_{ij}^+ \\ & \boldsymbol{w}^T x_{ij}^- + b \leq -(1 - \xi_{ij}^-) \\ & \xi_{ij}^-, \xi_{ij}^+ \geq 0\end{aligned} \tag{6}$$

where C represents the Box Constraint.

Instead, in the NSK [53], a bag is represented as the sum of all its instances, normalised by its L1- or L2-norm. The resulting representation is used to train a standard SVM by solving the optimisation problem:

$$\begin{aligned}\min_{\boldsymbol{w},b,\xi} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + C\left(\sum_{i=1}^{N^+}\xi_i^+ + \sum_{i=1}^{N^-}\xi_i^-\right)\\ \text{s.t.} \quad & \\ & \boldsymbol{w}^T \frac{\sum_j x_{ij}^+}{|B_i^+|} + b \geq 1 - \xi_i^+ \\ & \boldsymbol{w}^T \frac{\sum_j x_{ij}^-}{|B_i^-|} + b \leq -(1 - \xi_i^-) \\ & \xi_i^-, \xi_i^+ \geq 0.\end{aligned} \tag{7}$$

3.5 mi-SVM

The mi-SVM problem for MIL [29] is given by:

$$\begin{aligned}\min_{y_{ij}} \min_{\boldsymbol{w},b,\xi} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + C\left(\sum_{i=1}^{N^+}\sum_{j=1}^{L_i^+}\xi_{ij}^+ + \sum_{i=1}^{N^-}\sum_{j=1}^{L_i^-}\xi_{ij}^-\right)\\ \text{s.t.} \quad & \\ & y_{ij}(\boldsymbol{w}^T x_{ij}^+ + b) \geq 1 - \xi_{ij}^+ \\ & \boldsymbol{w}^T x_{ij}^- + b \leq -(1 - \xi_{ij}^-) \\ & \xi_{ij}^-, \xi_{ij}^+ \geq 0\end{aligned} \tag{8}$$

where the variable y_{ij} is a binary variable associated with the instances in the positive bags and is bound to satisfy the constraint that $y_{ij} = 1$ for at least one $j \in \{1, \ldots, L\}$.

Notice that the mi-SVM optimisation problem (8) is a mixed integer programming problem that can only be tackled with heuristic methods. In particular, we used the optimisation heuristic proposed in [29]. The mi-SVM starts by training the SIL-SVM described above. This is followed by a relabeling of the instances in the positive bags using the SIL decision hyperplane. Hence, if a positive bag contains no instances labeled as positive, the instance that gives the maximum margin of the decision hyperplane is relabeled as positive. This relabeling procedure is repeated, retraining a new SVM model until no labels are changed.

3.6 MI-SVM

The MI-SVM is an alternative way of applying the maximum margin approach to the MIL scenario. The underlying idea is to learn a model that assigns a margin larger than one to at least one instance in the positive bags. This leads to the optimisation problem:

$$\begin{aligned}\min_{\boldsymbol{w},b,\xi} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + C\left(\sum_{i=1}^{N^+}\xi_i^+ + \sum_{i=1}^{N^-}\sum_{j=1}^{L_i^-}\xi_{ij}^-\right)\\ \text{s.t.} \quad & \\ & \max_j \ \boldsymbol{w}^T x_{ij}^+ + b \geq 1 - \xi_i^+ \\ & \boldsymbol{w}^T x_{ij}^- + b \leq -(1 - \xi_{ij}^-) \\ & \xi_{ij}^-, \xi_i^+ \geq 0.\end{aligned} \tag{9}$$

The first constraint in this optimisation problem is not convex. By introducing an extra variable $s(i)$ for each bag, it is possible to convert the above formulation into a mixed integer programming problem. Then, in the bag-centered formulation, only one pattern per positive bag will determine the margin of the bag. These patterns can be identified as a witness of the entire bag. Hence, the MI-SVM formulation can be formulated as in [29]:

$$\begin{aligned}\min_{s} \min_{\boldsymbol{w},b,\xi} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + C\left(\sum_{i=1}^{N^+}\xi_i^+ + \sum_{i=1}^{N^-}\sum_{j=1}^{L_i^-}\xi_{ij}^-\right)\\ \text{s.t.} \quad & \\ & \boldsymbol{w}^T x_{is(i)}^+ + b \geq 1 - \xi_i^+ \\ & \boldsymbol{w}^T x_{ij}^- + b \leq -(1 - \xi_{ij}^-) \\ & \xi_{ij}^-, \xi_i^+ \geq 0.\end{aligned} \tag{10}$$

As in the mi-SVM, this mixed integer programme is hard to solve efficiently (even for datasets of moderate size); therefore, we employ the heuristic algorithm proposed in [29]. Note that the MI-SVM approach effectively ignores the negative instances in positive bags, and only one instance in the positive bags contributes to the optimisation of the hyperplane. On the other hand, in the mi-SVM, the negative instances in the positive bags, as well as multiple positive instances from one bag, can be the support vectors. At last, we note that, unlike the EMDD-SVM, both the mi-SVM and MI-SVM are able to identify positive instances belonging to positively labeled bags. It follows that a specific self-reported emotional event that occurred while watching a specific video can also be localised.

4 Datasets

We tested the three MIL-based approaches with two datasets in which the affective states were self-reported and sparsely annotated. The three MIL-based approaches were designed for modeling the temporal ambiguity of the affective response in a weakly supervised scenario. Thus, we used the physiological signals as unidimensional data predictors for testing the effectiveness and the reliability of our methodology. However, the algorithms may be generalized to handle bi-dimensional data (e.g., facial expressions) modeling both spatial and time dimension with sparse intra-frame and inter-frame annotations.

The first experiment used high-quality data from the Database for Emotion Analysis Using Physiological signals dataset (DEAP) [4]. The database includes the possibility to classify 3-D emotion dimensions induced by music videos shown to different users. We aimed to investigate how our proposed MIL-based affective state classification methods act when the physiological signal is

gathered by accurate sensors in a controlled environment (gold-standard scenario). In this setting, the user was monitored with an accurate set of sensors and the labeling of the perceived emotion was based on participants' self-reports. The second experiment aimed to test the reliability of our approach in a scenario closer to the real-world applications. Data were collected by an unobtrusive smartwatch sensor (worn on the subject's wrist) in a less-controlled environment, where the first author in collaboration with psychologists designed the data collection and the labeling procedure. The smartwatch sensor collects physiological signals without requiring additional sensors placed in different body parts (e.g. in the DEAP dataset the galvanic skin response [GSR] was measured by positioning two electrodes on the distal phalanges of the middle and index fingers). The use of unobtrusive consumer devices improves usability and acceptability on the one hand but worsens the quality of recorded data on the other hand.

4.1 DEAP dataset

In the DEAP experiment [4], 32 healthy participants were asked to watch 40 music videos. The video selection was performed by a semi-automated method, with the main goal of minimising bias. From the initial candidate of 120 stimuli (60 videos collected from a database and 60 manually collected to maximise the clearness of the emotional reaction for each of the quadrants), the final 40 test music videos were chosen by using a web-based subjective emotion assessment interface. Each of the 40 resulting videos lasted 1 min. The experiment was performed in two laboratory rooms with controlled illumination. First, a 2 min baseline signal was recorded, during which a fixation cross was displayed to the participant, who was asked to relax during this period. Then, the 40 videos were presented in 40 trials, each consisting of the following steps:

1) The trial number was displayed to inform the participant of their progress for 2 s.
2) Baseline recordings of 5 s (fixation cross).
3) Music video of 1 min.

After each video, participants were asked to self-report their emotional experience in four dimensions: *valence*, *arousal*, *dominance* and *liking*. The ratings for *valence*, *arousal* and *dominance* could range from 1 to 9 and were collected through the Self-Assessment Manikins (SAM)[58]. For the *liking* scale, thumbs down/thumbs up symbols were used. Full-scalp EEG and 13 peripheral physiological signals, including GSR, respiration amplitude, skin temperature (ST), blood volume pressure by plethysmograph (BVP), electromyograms of zygomaticus and trapezius muscles and electrooculogram (EOG) were recorded at a sampling rate of 512 Hz.

4.2 Consumer dataset

Twenty-nine volunteers (14 females, age range 20-30 years) were recruited at the Department of Information Engineering (Polytechnic University of Marche, Ancona Italy). None of them had a history of neurological disorders. They were asked to watch six movie clips (each lasting 4 min) and self-report through the SAM their emotional experience of each video in two dimensions: *valence* and *arousal*. Figure 1 shows the flow chart of the data recording setup. The experiment was composed by two different sessions performed on different days. Both sessions started with a resting phase (baseline stage), where subjects were asked to relax for 10 minutes lying on a couch without falling asleep. Then, during the first session (experimental stage I), three movie clips were presented to each participant. These videos were chosen with the purpose of eliciting positive *valence* (i.e. happiness or satisfaction) and to accustom subjects to the procedure. The movie clips proposed in the second session (experimental stage II) were instead selected with the aim of eliciting negative *valence* (i.e. sadness or fear). At the end of each movie clip, subjects self-reported (emotion self-assessment stage) the *valence* and *arousal* experienced in response to the stimulus by means of nine-point SAM scale [58] in a similar way to what was done in [4], [56]. The goal of this experiment was to provide a reliable prediction of the self-reported emotional response independently from the order and the nature of the video participants were presented with.

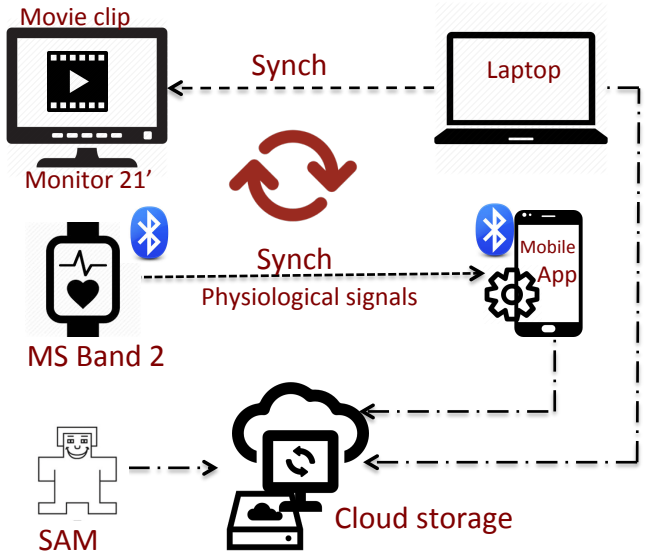


Fig. 1: A flow description of the Consumer dataset experiment: flow chart of the data recording setup. The physiological signals were collected from smartwatch sensors (Microsoft Band 2 [MS Band 2] [59]). A mobile application was implemented to gather the physiological signals via Bluetooth. The data collection was properly synchronised to the movie clips displayed on the laptop. Before each trial, the smartwatch was connected to the laptop to synchronise the clock time via the Microsoft Band Synch application [60].

We collected physiological signals from the sensors of a smartwatch (Microsoft Band 2 [59]) worn on each participant's wrist. A mobile application was implemented to gather the physiological signals via Bluetooth. The data collection was properly synchronised to the movie clips displayed on the laptop. Before each trial, the smartwatch was connected to the laptop to synchronise the clock time via the Microsoft Band Synch application [60]. Three physiological signals were recorded: the inter-beat interval (IBI) of heart rate (HR) acquired with an event-based sampling, the GSR sampled at the frequency of 5 Hz and the ST sampled at the frequency of 0.03 Hz.

The following aspects of the Consumer dataset contribute to being close to a real-world application: (i) physiological measurements were collected through a smartwatch, (ii) a mobile application was implemented to gather the physiological signals

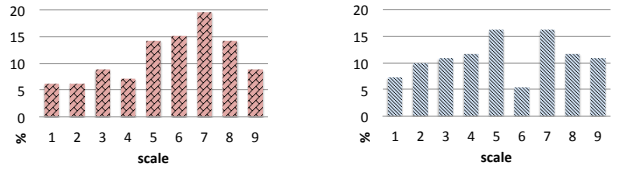
from smartwatch to mobile phone; (iii) the subject is sitting while watching TV movie without any specific constraint.

5 Modeling and Results

This section presents our experiments on the DEAP and Consumer datasets. We first describe the data processing stage and the experimental setup. Then, we present the experimental results and statistical analyses [1].

5.1 Data processing

The features were extracted to be as similar as possible for both datasets. The original self-reports of both *arousal* and *valence* were binarised by thresholding at level 5 (midpoint) following the other state-of-the-art works [4], [56] related to the DEAP dataset. Figure 2 shows the percentage histograms of the original self-reports for the *arousal* (see Figure 2a) and *valence* (see Figure 2b) dimensions.



(a) *arousal*. (b) *valence*.

Fig. 2: The percentage histogram of the original self-reports of both *arousal* (a) and *valence* (b).

Hence, we defined "low/negative" values as below 5 and "high/positive" values as above 5 for both the *arousal* and *valence* dimensions respectively. Data processing was implemented to extract salient features for each physiological signal according to the computed ML model. For the standard supervised learning approach, we considered the synchronously recorded physiological signals over each video as a single instance represented by a row-vector of d-features (bag-level features). For the MIL models, the physiological signals were segmented using the timestamps of the visualised videos. In particular, we evaluated two settings:

- 3 instances per bag: each synchronously recorded physiological signal was segmented in three windows each containing $ns/2$ samples (WS) and overlapped by $WS/2$;
- 5 instances per bag: each synchronously recorded physiological signal was segmented in five windows each containing $ns/3$ samples (WS) and overlapped by $WS/2$;

where ns is the total number of samples in a physiological signal. Although we modeled the single bag as a set of multiple instances represented by the extracted features, we did not assume explicitly an ordinal and defined structure of the instance.

5.1.1 DEAP dataset

All the acquired physiological measurements were downsampled to 128 Hz and then segmented into 60 s trials. The 5 s pre-trial baseline was removed and thus not considered for further analyses. The preprocessing and the feature extraction stages were performed with the Matlab Toolbox for Emotional feAture extraction from Physiological signals (TEAP) [61]. Table 1 shows the physiological signals and the resulting 46 features used for the data analysis. The ST was not considered in the analysis, due to the presence of several outliers [61].

1. MATLAB code used in the experiments will be made available at the following GitHub link: https://github.com/whylearning22/EmotionPrediction.

TABLE 1: Features extracted from physiological signals: DEAP dataset. The employed physiological signals are galvanic skin response (GSR), blood volume pressure (BVP), respiration (RESP) and two channel electromyography signal (EMG). The preprocessing and the feature extraction stage were performed with the Matlab Toolbox for Emotional feAture extraction from Physiological signals (TEAP) [61]. The admissible rise time range for finding the number of GSR peaks was tuned from .1 to 4 sec while the amplitude threshold was tuned to 20 ohm.

Signal	Extracted Features
GSR	1) average peak amplitude; 2) average rising time; 3) number of peaks per second; 4) average; 5) standard deviation; 6) 1^{st} quartile; 7) 3^{rd} quartile.
BVP	8) average; 9) standard deviation of Inter-beat interval (IBI); 10) average of IBI; 11-15) multi-scale-entropy of IBI [2]; 16) spectral power in $0-0.1$ Hz band; 17) spectral power in $0.1-0.2$ Hz band; 18) spectral power in $0.2-0.3$ Hz band; 19) spectral power in $0.3-0.4$ Hz band; 20) energy ratio between the frequency bands $0-0.08$ Hz and $0.15-0.5$ Hz; 21) spectral power of IBI in $0.01-0.08$ Hz band (LF); 22) spectral power of IBI in $0.08-0.15$ Hz band (MF); 23) spectral power of IBI in $0.15-0.5$ Hz band (HF); 24) energy ratio of IBI between MF and LF+HF.
RESP	25) average; 26) standard deviation; 27) kurtosis; 28) skewness; 29-35) 7 spectral power in the bands from 0 to 2.5 Hz; 36) main frequency.
EMG	37-42) average; 38-43) standard deviation; 39-44) kurtosis; 40-45) skewness; 41-46) spectral power over 20 Hz.

5.1.2 Consumer dataset

The zero-order hold interpolation was applied to resample all physiological signals (i.e. GSR, IBI and ST) at 5 Hz while preserving the information of the acquired signal. The GSR samples that were higher than a 300 μS threshold were considered outliers and replaced using the cubic spline interpolation. The 300 μS threshold was selected according to [62] by computing the maximum acceptable absolute deviation around the median of the GSR signal. All the data were then smoothed using a five-sample moving average filter to reduce high-frequency noise. Because the *accuracy* of the smartwatch data is not comparable with respect to signals gathered from gold standard sensors (see [4]), the extracted features corresponded to a subset of those computed by the major reference works published in the affective computing literature [4], [61], [63], [64] (see Table 2). Seven features of IBI ($26-32$) were computed with respect to values recorded during the baseline stage.

5.2 Modeling setup and measures

The extracted features represent the predictors of the proposed MIL-based algorithm described in Section 3. The two classification tasks perform a binary prediction of a low/high level of *arousal* and negative/positive level of *valence*. The assessment of the MIL model was performed according to the following measures:

- *accuracy*: the percentage of correct predictions;
- *confusion matrix*: the square matrix that shows the type of error in a supervised paradigm;
- *macro-F1*: the harmonic mean of precision and recall averaged over all output categories;

TABLE 2: Features extracted from physiological signals: Consumer dataset

Signal	Extracted Features
GSR	1) average; 2) standard deviation; 3) average of the derivative; 4) root mean square of the derivative.
IBI	5) standard deviation; 6) standard deviation of the first difference; 7) root mean square of the first differences; 8) number of the absolute values of the first differences samples greater than 50 ms (NN50); 9) number of the first differences samples greater than 50 ms (dNN50); 10) number of the first differences samples lower than 50 ms (aNN50); 11) number of the absolute values of the first differences samples greater than 50 ms normalized over the number of samples; 12) average of the absolute values of the first differences; 13) average of the absolute values of the first differences of the normalized signal; 14) average of the absolute values of the second differences; 15) the means of the absolute values of the second differences of the normalized signals; 16) spectral power in $0.04-0.15$ Hz band (LF); 17) spectral power in $0.15-0.4$ Hz band (HF); 18) energy ratio between LF and HF; 19) energy ratio between LF and LF+HF; 20) energy ratio between HF and LF+HF; 21) spectral power in $0.04-0.4$ Hz band (LF+HF); 22) Poincare plot feature $SD1^2$; 23) Poincare plot feature $SD2^2$; 24) Poincare plot feature $SD1^2/SD2^2$; 25) average (IBI^{mean}); 26) $NN50-NN50_{baseline}$; 27) $dNN50-dNN50_{baseline}$; 28) $aNN50-aNN50_{baseline}$; 29) $IBI^{mean}-IBI^{mean}_{baseline}$; 30) standard deviation of the IBI with the mean value calculated from the baseline IBI; 31) $SD1^2/SD1^2_{baseline}$; 32) $SD2^2/SD2^2_{baseline}$;
Skin Temperature	33) average; 34) maximum

- *receiver operating characteristic (ROC)*: is designed by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. It illustrates the performance of a binary classifier as its discrimination threshold is varied.

The macro-F1 score deviates from normality according to the one-sample Kolmogorov-Smirnov test (D = 0.268, $p < .05$). Hence, to test for a significant difference from chance-level classification, the one-sided Wilcoxon signed rank test was performed comparing the *macro-F1* distribution with respect to level of chance (.5). Accordingly, the statistical comparison between the best proposed MIL-based approach and the best standard supervised learning competitor was also performed by means of a non-parametric one-sided Wilcoxon signed rank test (significance level = 0.05).

In both settings, we computed a *user-specific* and a *user-independent* MIL model. Due to the high inter-subject variability that affects the physiological signals [4], a user-independent setup (i.e. Leave-One-Subject-Out [LOSO] evaluation) did not achieve satisfactory performance (*accuracy* not above chance level) for the DEAP dataset and was not reported in this study. The difficulty of the *user-independent* task for the DEAP dataset is also confirmed in [56], where the authors obtained satisfactory results only when exploiting the EEG signal, as well. Furthermore, they implemented a multiple kernel learning/multi-task approach for combining the different natures of the employed predictors and for modeling multiple users together, respectively. Future work will focus on addressing individual differences in the DEAP dataset, which is not the current focus on this paper. For the DEAP dataset, we thus used Leave-One-Video-Out (LOVO) cross-validation. The LOVO evaluation was implemented for each subject following the study of [4], while the overall results were presented considering the average performances over all participants.

For the second dataset (Consumer dataset), we used a 10-fold Cross Validation (10-CV) evaluation over video and a LOSO evaluation for the *user-independent* setup. The 10-CV evaluation was implemented in the second experiment, dividing all videos in 10 folds and selecting nine folds for training and one fold for testing. This procedure allowed to take into account the small number of observations for each subject. On the other hand, the LOSO evaluation for the Consumer dataset was performed with an iterative selection of one subject for testing and the other subjects for training.

5.3 Results

5.3.1 Experiment 1: DEAP dataset

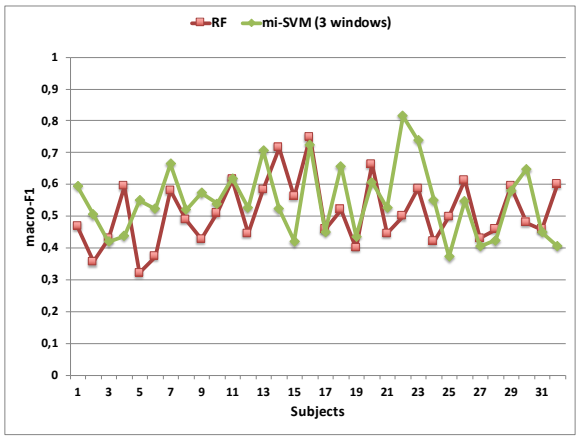
Table 3 reports the outcome of the latest studies related to the emotion approaches using the DEAP dataset and the physiological signals as predictors. The Naive Bayes (NB), the linear SVM and the Random Forest (RF) classifiers are standard supervised ML algorithms used as comparison. The NB was employed in [4], [65], while the RF achieved the best performance among other tested supervised classifiers, such as Decision Tree and K-Nearest Neighbors. Accordingly, both the SVM and RF were widely used [8]–[10], [56] for solving emotion recognition tasks using other biosignals (e.g. electroencephalogram) and facial expressions. In addition, from the machine learning viewpoint, the applied MIL methods are a theoretical extension of linear SVM. Following [55], the optimal number of trees for the RF classifier was picked from the subset $\{20,50,100,200\}$ via nested Cross Validation. Moreover, we compared the proposed approach with respect to the Milboost methodology proposed in [38], [39] and originated from the algorithm proposed by Viola et al. [46]. The number of weak learners was set to 100. Although this hyperparameter value has been originally set according to [39], further experiments for different values ($\{20,50,100,200\}$) of this hyperparameter confirmed that 100 weak learners is the best choice for solving our task.

TABLE 3: State of the art: Emotion recognition using physiological signals of the DEAP dataset. PHY: physiological signals, FS: fisher score, NB: naive bayes, FFS: Forward Features Selection, LOVO: Leave One Video Out evaluation, CV: Cross Validation evaluation

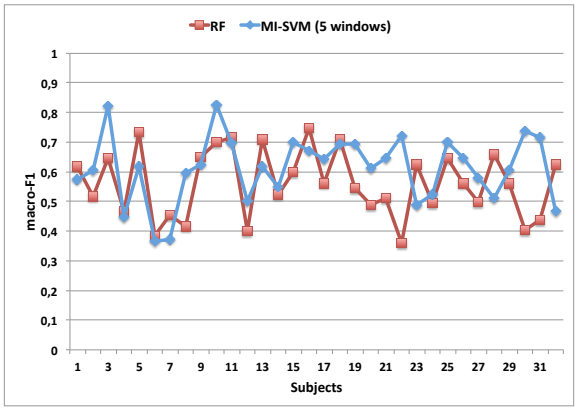
Ref	Year	Method	Signals	Procedure	Labels	F1	ACC
[4]	2012	FS+NB	PHY	LOVO	arousal	.53	.57
					valence	.61	.63
					liking	.54	.59
[66]	2014	HMM	PHY	5-CV	arousal		.55
					valence		.58
			EOG+EMG+PHY	5-CV	arousal		0.63
					valence		.54
[65]	2015	FFS + NB	PHY	LOVO	arousal		.59
					valence		.61

Table 4 shows the average accuracies and *macro-F1* scores of *user-specific* setup (LOVO evaluation) over participants for the standard ML algorithms and the MIL methods for each task (i.e., *arousal*, *valence*) and setting (i.e. $L = 3$ and $L = 5$).

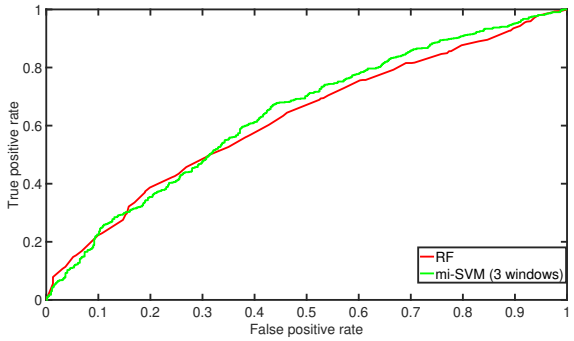
Concerning the estimation of the two *arousal* levels, the mi-SVM method with $L = 3$ showed the highest *macro-F1* (*macro-F1* = 0.546, ACC = 0.611), while the EMDD-SVM approach with $L = 3$ had the lowest *macro-F1* (*macro-F1*=0.463, ACC = 0.559). The MI-SVM method with $L = 5$ had the best outcome (*macro-F1* = 0.612, ACC = 0.636) for recognising the *valence* level, while



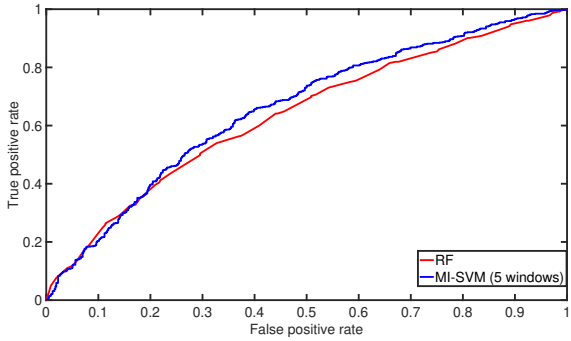
(a) The *macro-F1* score of the mi-SVM with $L = 3$ and the standard RF approach for each participant for the *arousal* task.



(b) The *macro-F1* score of the MI-SVM with $L = 5$ and the standard RF approach for each participant for the *valence* task.



(c) ROC curve of the mi-SVM with $L = 3$ and the standard RF approach over all participants for the *arousal* task.



(d) ROC curve of the MI-SVM with $L = 5$ and the standard RF approach over all participants for the *valence* task.

Fig. 3: Experimental results 1: DEAP dataset.

the EMDD-SVM approach with $L = 5$ has the worst performance (*macro-F1* = 0.526, ACC = 0.566).

For the *valence* task, the *macro-F1* was significantly higher ($p < .05$) than chance level (.5) for both the standard and MIL methods. The *macro-F1* obtained from the EMDD-SVM with $L = 5$ did not reach significance ($p = .350$). On the other hand, for the estimation of *arousal* level, only the *macro-F1* of the mi-SVM approach with $L = 3$ achieved a $p < .05$, while none of the other approaches implemented performed significantly better than chance.

The *macro-F1* of the MI-SVM approach with $L = 5$ is significantly higher than RF ($W = 366; Z = 1.898; p < .05$) and SVM ($W = 390; Z = 2.347; p < .01$) for the estimation of *valence*, while the *macro-F1* of the mi-SVM approach with $L = 3$ is not signficantly higher than RF ($W = 340; Z = 1.412; p = .079$) and SVM ($W = 296; Z = 0.589; p = .278$) for the prediction of *arousal*.

We chose to compare *macro-F1* scores, the confusion matrices and the ROC of the best MIL approach with those obtained from the best standard supervised learning competitor (i.e. RF) for both the *arousal* and *valence* tasks.

Figure 3a shows the *macro-F1* scores of the mi-SVM with three windows per video and the standard RF approach for each participant for the *arousal* task. When compared to standard RF, the mi-SVM with $L = 3$ showed higher *macro-F1* scores in 17 out of 32 participants (i.e. participants 1-2, 5-10, 12-13, 18-19, 21-24, 30).

TABLE 4: Average accuracies (ACC) and *macro-F1* (F1) of the *user-specific* setup (LOVO evaluation) over participants for the MIL algorithms. For comparison, we give the results of standard NB, SVM and RF. Stars indicate whether the *macro-F1* distribution over subjects is significantly higher than chance level (i.e. *macro-F1* = 0.5) according to the one-sided Wilcoxon signed rank test ($** = p < .01, * = p < .05$).

	Arousal		**Valence**	
Algorithm	ACC	F1	ACC	F1
Standard				
NB	0.572	0.514	0.595	0.577**
SVM	0.591	0.539	0.581	0.557**
RF	0.601	0.511	0.598	0.562**
MIL				
$L = 3$				
mil-Boost	0.608	0.521	0.600	0.547**
mi-SVM	**0.611**	**0.546***	0.622	0.595**
MI-SVM	0.594	0.533	0.577	0.556*
EMDD-SVM	0.559	0.463	0.587	0.544*
$L = 5$				
mil-Boost	0.605	0.532	0.604	0.560**
mi-SVM	0.583	0.512	0.621	0.593**
MI-SVM	0.585	0.530	**0.636**	**0.612****
EMDD-SVM	0.589	0.501	0.566	0.526

Table 5 shows the confusion matrices of the mi-SVM and the RF approach over all participants for the *arousal* task. Notice that the standard *Recall* of the mi-SVM (0.71) was higher than RF (0.65), while the *Precision* of RF (0.65) was comparable with that obtained by the mi-SVM (0.65).

TABLE 5: Confusion matrices (rows are the true classes) of the mi-SVM with $L = 3$ and the RF approach over all participants for the *arousal* task.

mi-SVM

	low	*high*
low	0.47	0.53
high	0.29	0.71

RF

	low	*high*
low	0.53	0.47
high	0.35	0.65

The overall ROC for the mi-SVM and RF is depicted in Figure 3c. The area under curve (AUC) of the mi-SVM (AUC = 0.639) is higher than RF (AUC = 0.624).

Concerning the estimation of the *valence* level, the performance of the MI-SVM with $L = 5$ is higher than the RF method for 19/32 participants (i.e. participant 2, 3, 8, 10, 12, 14-15, 17, 19, 20-22, 24-27, 29-31) (see Figure 3b).

Table 6 shows the confusion matrices of the MI-SVM and the RF approaches over all participants for the *valence* task. The MI-SVM reaches higher standard *Precision* and *Recall* (0.67 and 0.68) values than the ones obtained for RF (0.63 and 0.65).

TABLE 6: Confusion matrices (rows are the true classes) of the MI-SVM with $L = 5$ and the RF approach over all participants for the *valence* task

MI-SVM

	neg	*pos*
neg	0.58	0.42
pos	0.32	0.68

RF

	neg	*pos*
neg	0.53	0.47
pos	0.35	0.65

Figure 3d shows the overall ROC for MI-SVM and the RF method. The AUC for the MI-SVM (AUC = 0.658) is higher than RF (AUC = 0.636).

5.3.2 Experiment 2: Consumer dataset

Evaluation over video. Table 7 shows the average accuracies and *macro-F1* scores of the *user-specific* setup (10-CV evaluation) of the MIL algorithm and the standard methods for the two dimensions of emotional state: *arousal* and *valence*. For the classification of *arousal* level, the mi-SVM approach with $L = 3$ revealed the highest *macro-F1* (*macro-F1* = 0.637, ACC = 0.703), while the RF classifier and the EMDD-SVM model with $L = 3$ had the lowest *macro-F1* (respectively, *macro-F1* = 0.473, ACC = 0.594 and *macro-F1* = 0.424, ACC = 0.630). The RF model showed the best performance (*macro-F1* = 0.682, ACC = 0.686), while the mi-SVM method with $L = 3$ was the best competitor (*macro-F1* = 0.662, ACC = 0.669) for the *valence* classification.

TABLE 7: Average accuracies (ACC) and *macro-F1* (F1) of the *user-specific* setup (10-CV evaluation) over 10-fold for the MIL algorithms. For comparison, we give the results of standard NB, SVM and RF. Stars indicate whether the *macro-F1* distribution over the 10-fold is significantly higher than chance level (i.e. *macro-F1* = 0.5) according to the one-sided Wilcoxon signed rank test ($** = p < .01, * = p < .05$).

	Arousal		**Valence**	
Algorithm	ACC	F1	ACC	F1
Standard				
NB	0.546	0.530	0.623	0.618 **
SVM	0.641	0.552	0.591	0.569
RF	0.594	0.473	**0.686**	**0.682****
MIL				
$L = 3$				
mil-Boost	0.668	0.520	0.657	0.651**
mi-SVM	**0.703**	**0.637***	**0.669**	**0.662****
MI-SVM	0.643	0.554	0.618	0.599
EMDD-SVM	0.630	0.424	0.583	0.612
$L = 5$				
mil-Boost	0.652	0.589	0.632	0.618*
mi-SVM	0.640	0.600*	0.628	0.617
MI-SVM	0.685	0.590	0.656	0.636*
EMDD-SVM	0.714	0.591	0.629	0.602*

The standard SVM approach had the worst outcome (*macro-F1* = 0.591, ACC = 0.569).

Concerning the classification of *valence* level, all the MIL models except the mi-SVM with $L = 5$, the MI-SVM and the EMDD-SVM with $L = 3$ showed a *macro-F1* significantly higher ($p < .05$) than chance level (.5), while for the standard methods (i.e. NB, SVM and RF), the *macro-F1* of the NB and RF overcame significantly ($p < .05$) the chance level (.5). On the other hand, for the estimation of *arousal*, only the mi-SVM with $L = 3$ and with $L = 5$ revealed a *macro-F1* significantly higher ($p < .05$) than chance level (.5).

The *macro-F1* of the mi-SVM approach with $L = 3$ is significantly higher than RF ($W = 40.5; Z = 2.075; p < .05$), but it is not significantly higher than SVM ($W = 33; Z = 1.185; p = .125$) for the estimation of *arousal*. The *macro-F1* of mi-SVM approach with $L = 3$ is not significantly higher than SVM ($W = 36; Z = 1.540; p = .065$) and RF ($W = 10; Z = -0.766; p = .778$) for the prediction of *valence*.

Figure 4a shows the *macro-F1* scores of each fold for both the mi-SVM with $L = 3$ and the standard RF methods. The classification of *arousal* level was identical for both methods in fold 5, while the mi-SVM was superior to SVM for 8 out of 10 folds (i.e. 1-4, 6-8 and 10).

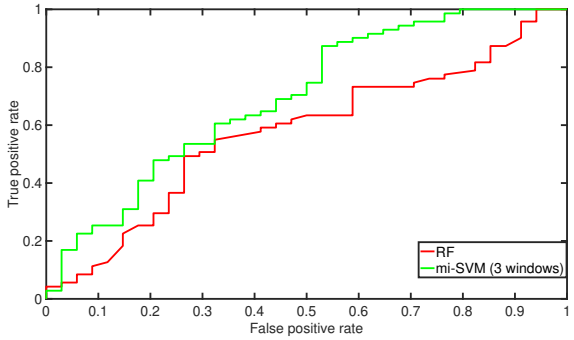
Table 8 shows the confusion matrices of the mi-SVM with $L = 3$ and the standard RF approach over all folds for the *arousal* task. The standard *Recall* and *Precision* of the mi-SVM (0.82 and 0.76, respectively) are greater than the ones of RF (0.77 and 0.67, respectively).

The ROCs for the mi-SVM $L = 3$ and RF are depicted in Figure 4c. The area under the curve (AUC) of the mi-SVM (AUC = 0.697) is higher than the ones of RF (AUC = 0.576).

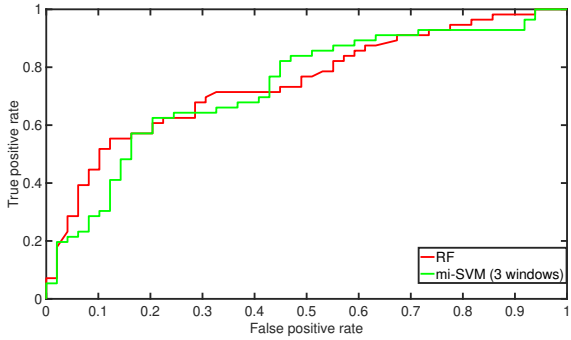
For the *valence* task, compared to standard RF, the mi-SVM with $L = 3$ yielded a higher *macro-F1* in three out of 10 folds (i.e. fold 3, 5 and 7) and a comparable trend in three out of 10 folds (see Figure 4b).

(a) The *macro-F1* score of the mi-SVM with $L = 3$ and the standard RF approach for each fold for the *arousal* task.

(b) The *macro-F1* score of the mi-SVM with $L = 3$ and the standard RF approach for each participant for the *valence* task.



(c) ROC curve of the mi-SVM with $L = 3$ and the standard RF approach over all folds for the *arousal* task.



(d) ROC curve of the mi-SVM with $L = 3$ and the standard RF approach over all folds for the *valence* task.

Fig. 4: Experimental results 2: Consumer dataset.

TABLE 8: Confusion matrices (rows are the true classes) of the mi-SVM with $L = 3$ and the RF approach over all participants for the *arousal* task

mi-SVM

	low	*high*
low	0.47	0.53
high	0.18	0.82

RF

	low	*high*
low	0.21	0.79
high	0.23	0.77

TABLE 9: Confusion matrices (rows are the true classes) of the mi-SVM with $L = 3$ and the RF approach over all participants for the *valence* task

mi-SVM

	neg	*pos*
neg	0.73	0.27
pos	0.39	0.61

RF

	neg	*pos*
neg	0.71	0.29
pos	0.34	0.66

Table 9 shows the confusion matrices of the mi-SVM and the standard RF approach over all folds for the *valence* task. The mi-SVM showed a lower standard *Recall* than RF (0.61 vs 0.66). On the other hand, the standard *Precision* of the mi-SVM (0.72) is comparable than those of RF (0.73).

Figure 4d shows the ROC curves for the mi-SVM and the standard RF method. The AUCs of the mi-SVM (AUC = 0.733) and RF (AUC = 0.747) were comparable.

LOSO evaluation. Table 10 shows the average accuracies of the *user-independent* setup (LOSO evaluation) of the MIL algorithm and the standard methods for the two dimensions' emotional state: *arousal* and *valence*. For the classification of *arousal* level, the EMDD-SVM with $L = 5$ revealed the highest *accuracy* (ACC = 0.691), while the NB classifier had the lowest *accuracy* (ACC = 0.500). The MI-SVM method with $L = 3$ showed the best performance (ACC = 0.680) for the *valence* classification, while the EMDD-SVM approach with $L = 5$ had the worst outcome (ACC = 0.547).

For what concerns the classification of *valence* level, the MI-SVM and the mil-Boost with $L = 3$ showed an *accuracy* significantly higher ($p < .05$) than chance level (.5), while for the standard methods (i.e. NB, SVM and RF), only the *accuracy* of the RF overcame significantly ($p < .05$) the chance level (.5). On the other hand, for the estimation of *arousal*, all the MIL methods except the MI-SVM with $L = 5$ revealed a *macro-F1* significantly higher ($p < .05$) than chance level (.5). Both the *accuracy* of SVM and of RF overcame significantly ($p < .05$) the chance level.

TABLE 10: Average accuracies (ACC) of the *user-independent* setup (LOSO evaluation) for the MIL algorithms. For comparison, we give the results of standard NB, SVM and RF. Stars indicate whether the *accuracy* distribution over the n subjects is significantly higher than chance level (i.e. ACC = 0.5) according to the one-sided Wilcoxon signed rank test ($** = p < .01, * = p < .05$)

	Arousal	**Valence**
Algorithm	ACC (std)	ACC (std)
Standard		
NB	0.500 (0.372)	0.595 (0.324)
SVM	0.647∗ (0.333)	0.571 (0.328)
RF	0.667∗∗ (0.315)	0.639∗ (0.290)
MIL		
3 windows		
mil-Boost	0.655∗ (0.365)	0.679∗∗ (0.296)
mi-SVM	0.655∗ (0.351)	0.579 (0.300)
MI-SVM	0.657∗∗ (0.277)	**0.680**∗∗ (0.264)
EMDD-SVM	0.682∗∗ (0.326)	0.561 (0.285)
5 windows		
mil-Boost	0.621∗ (0.334)	0.556 (0.331)
mi-SVM	0.637∗ (0.348)	0.563 (0.279)
MI-SVM	0.598 (0.319)	0.645∗ (0.306)
EMDD-SVM	**0.691**∗∗ (0.296)	0.547 (0.329)

5.4 Window setting

The performance of the MIL-based approaches is influenced by the structure of the bag (window setting): different (i) window sizes (WSs) and (ii) overlap factors (OFs). WS refers to the amount of samples within each window, while OF represents the amount of overlap between two consecutive windows. For instance, OF = 1/2 means that the amount of overlap between two consecutive windows is WS/2. We have measured the sensitivity of the algorithm to different window settings using the DEAP dataset (gold-standard scenario, see Figure 5), which is comprised of 40 music videos of 1 min for each of the 32 participants. Because all the acquired physiological signals were downsampled to 128 Hz, each unsegmented physiological signal comprised 7680 samples ($ns = 7680$). Note that $WS = 7680$ corresponds to taking a SIL approach, while a small WS (e.g. $WS = 1280$ and $WS = 640$) did not achieve a macro-F1 score above chance level ($p < .05$).

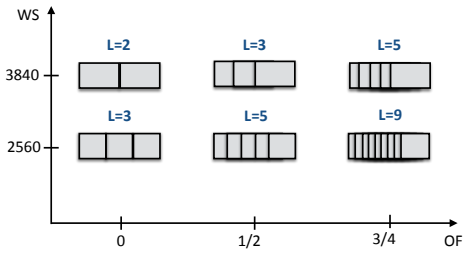
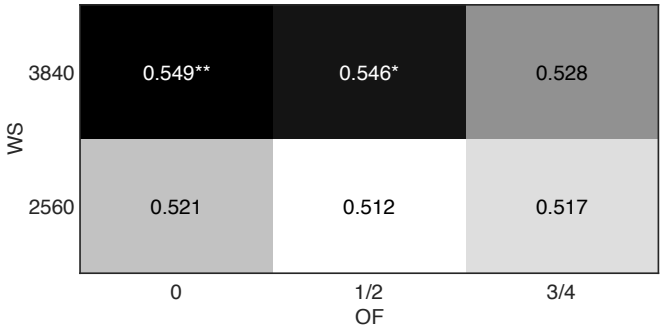
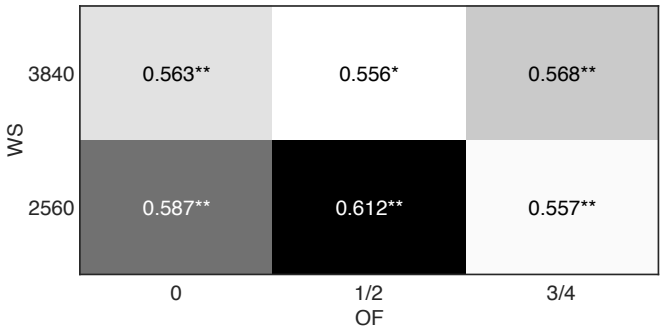


Fig. 5: Different explored window settings to measure the sensitivity of the algorithm in the DEAP dataset. WS refers to the window size (i.e. the amount of samples within each window), OF refers to the overlap factor (i.e. the amount of overlap between two consecutive windows) and L refers to the number of instances of each bag.

The *macro-F1* scores of the best MIL-based approaches are summarised in Figure 6. The mi-SVM with $WS = 3840$ and $OF = 0$ achieved the best performance for solving the *arousal* task. The *macro-F1* of the mi-SVM approach with this WS is significantly higher ($p < .05$) than chance level (0.5). The MI-SVM with $WS = 2560$ and $OF = 1/2$ achieved the best *macro-F1* score for solving the *valence* task. Meanwhile, the *macro-F1* of the MI-SVM approach with this WS is significantly higher ($p < .05$) than chance level (0.5).



(a) The *macro-F1* score of the mi-SVM with different window settings for solving the *arousal* task.



(b) The *macro-F1* score of the MI-SVM with different window settings for solving the *valence* task.

Fig. 6: The *macro-F1* scores of the best MIL-based approach for different WSs for the DEAP dataset. Stars indicate whether the *accuracy* distribution over the 32 subjects is significantly higher than chance level (i.e. *macro-F1*=0.5) according to the one-sided Wilcoxon signed rank test ($** = p < .01, * = p < .05$).

5.5 Pattern localisation

The mi-SVM and MI-SVM can be exploited to localise the most prominent emotional response inside each bag (i.e. predict the label of each instance). In particular, we show the results of the best MIL-based approach for predicting self-reported *arousal* (mi-SVM with $L = 3$, see Figure 7a and Figure 7b) and *valence* (MI-SVM with $L = 5$, see Figure 7c and Figure 7d) for a single movie of the DEAP dataset for localising the most prominent emotional responses inside each bag.

In the performed analysis, only the subjects who reported also instances of a high level of arousal (respectively 22 subjects for the video 6 and 20 subjects for the video 30) and a positive level of *valence* (respectively 30 subjects for the video 19 and 22 subjects for the video 27) were considered. The main goal of this analysis was to demonstrate the potential of the MIL methods to localise—once the emotional response was present—the most prominent emotional event within each bag. For this reason, we considered all trials where the emotional response should be present strongly in terms of a self-reported high level of arousal and positive level of valence. However, because the experimental

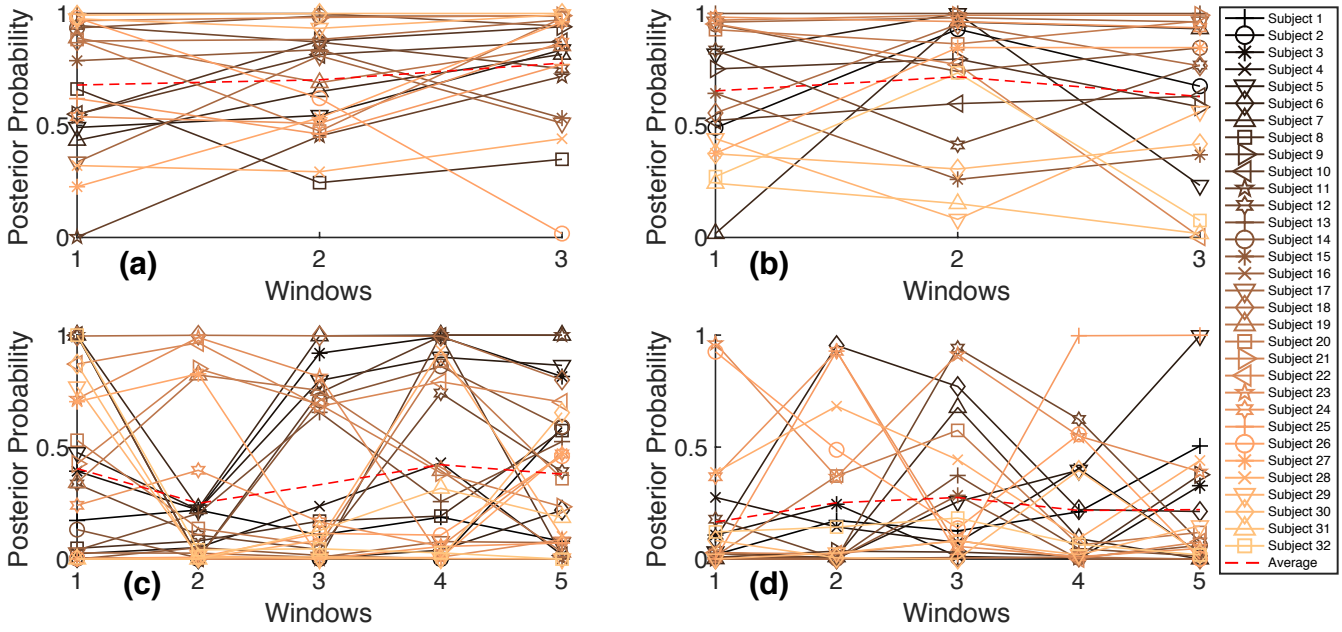


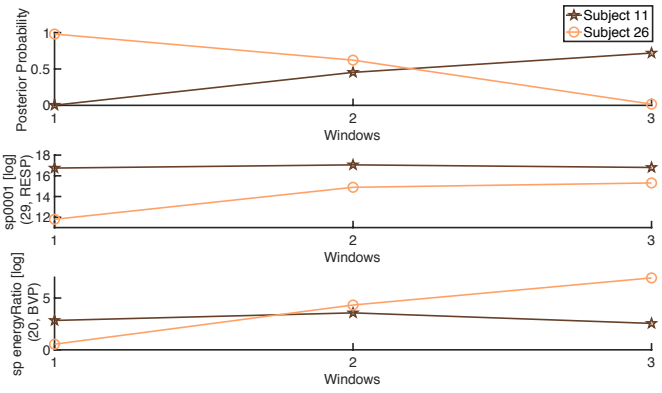
Fig. 7: Pattern localisation: DEAP dataset. The posterior probability of each instance of the mi-SVM with $L = 3$ for solving the *arousal* tasks (a) and (b). The results show different subjects while watching video 6 (a) and 30 (b). The posterior probability of each instance of the MI-SVM with $L = 5$ for solving the *valence* tasks (c) and (d). The results show different subjects while watching video 19 (c) and 27 (d). The dashed line shows the posterior probability averaged over the considered subjects.

results were computed for the overall DEAP dataset, the prediction of the label of each instance can be generalised for all trials and for all subjects, regardless of whether they reported either a low/negative or high/positive level of arousal and valence, respectively. The margin of the predicted mi-SVM and MI-SVM responses for each instance was mapped into a [0-1] interval (posterior probability) by using a sigmoid function, without changing the original error function. The mapping was performed according to [67], adding a post-processing step where the sigmoid parameters were learned with regularised binomial maximum likelihood. Feature importance was computed according to the absolute value of the mi-SVM and MI-SVM coefficients.

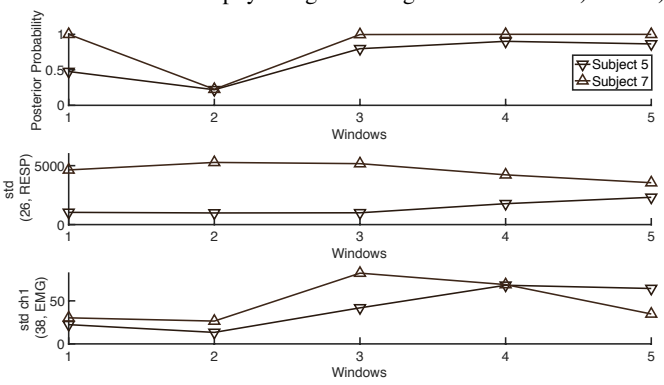
The changes in posterior probability of each instance of the MIL methods can be correlated to physiological changes. Figure 8 shows the posterior probability and the related physiological changes across each instance for solving the *arousal* task for subjects 11 and 26 while watching video 6 (see Figure 8a) and for solving the *valence* task for subjects 5 and 7 while watching video 19 (see Figure 8b). The videos were chosen from among the 40 as those that have been rated with a high level of *arousal* and positive level of *valence* by more than 50% of the subjects. Subjects 11 and 26 self-reported a high level of *arousal* for video 6 while subjects 5 and 7 self-reported a positive level of *valence* for video 19. Although this deepening can be performed for all subjects, the sparse model coefficients obtained for these four subjects might help to interpret the results better in terms of any correlation between the posterior probability of each instance of the MIL methods and the physiological signals. The mi-SVM selected only 10 out of 46 features with a cumulative feature importance greater than 50% for both subjects 11 and 26. Consequently, features 20 (BVP [sp energyRatio]) and 29 (RESP [sp001]) have been identified as two of the most discriminative features in common to both models for the *arousal* task. Similarly, in the *valence* task, the MI-SVM selected only 11 out of 46 features with a cumulative feature importance greater than 50% for both subjects 5 and 7. Thus, features 26 (RESP [std]) and 38 (EMG [std ch1]) have been used as two of the most discriminative features in common with both models. The different ways in which the MIL method (i.e. mi-SVM) chooses the most prominent instance for subjects 11 and 26 for predicting a high level of *arousal* may be associated with different changes in RESP (sp001) and BVP (sp energyRatio) features (see Figure 8a). Likewise, the similar ways in which the MIL method (i.e. MI-SVM) chooses the most prominent instance for subjects 5 and 7 for predicting a positive level of *valence* may be associated with a similar trend in RESP (std) and EMG (std ch1) features.

5.5.1 Feature importance

We expand on the previous analyses by considering whether there is a sort of agreement on the way the MIL method selects the most important features to discriminate a high level of *arousal* and positive level of *valence* for different subjects while watching videos 6 and 19, respectively (see Figure 9). Thus, based on the pattern localisation analysis, we focused only on those subjects who self-reported a high level of *arousal* and a positive level of *valence*. For each subject, the absolute coefficient values were normalised from 0 to 1. For both tasks, the MIL methods agreed to consider the main frequency of the RESP signal (36) as the less relevant feature. Concerning the *arousal* task, there is no substantial agreement on the most relevant features across different subjects. However, on average, the most important features are the average RESP (25), the average EMG signal (42) and the spectral power over 20 Hz of the EMG signal (46) (see Figure 9a). On the contrary, for the *valence* task, there is a high agreement across subjects concerning how MI-SVM selects the spectral power over 20 Hz of the EMG signal (41) as one of the most relevant features



(a) The posterior probability of each instance of the mi-SVM with $L = 3$ for solving the *arousal* task for subjects 11 and 26 while watching video 6 and the related physiological changes for features 29) and 20).



(b) The posterior probability of each instance of the MI-SVM with $L = 5$ for solving the *valence* task for subjects 5 and 7 while watching video 19 and the related physiological changes for features 26) and 38).

Fig. 8: Focus on pattern localisation: the posterior probability and the related physiological changes for each instance.

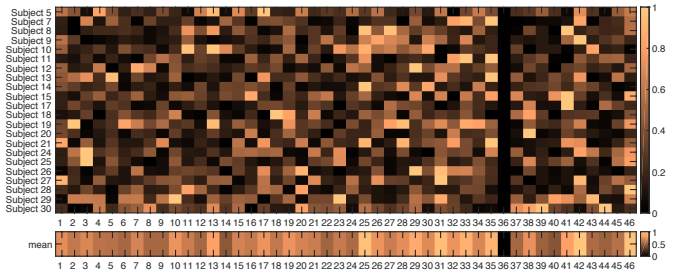
and the average GSR (4) as one of the less relevant features. On average, the most important features for solving the *valence* task are the spectral power over 20 Hz of the EMG signal (41), the skewness of the EMG signal (45) and the number of peaks of the GSR signal (3) (see Figure 9b).

6 Discussion and Conclusions

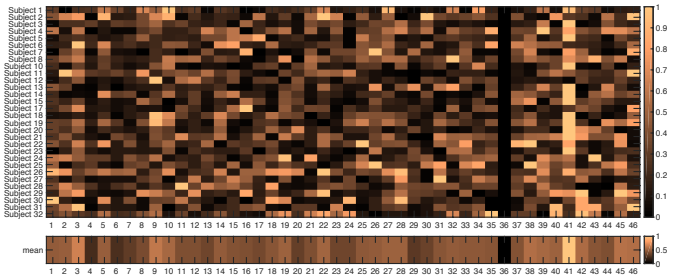
This section presents the discussion and conclusions of our work. We first discuss the presented results and suggest possible interesting directions for future work. Then, we provide the conclusions.

6.1 Reliability of the proposed approach

The MIL approach is able to model the ambiguity of emotional response in sparse labeling contexts, while standard supervised approaches (e.g. [4]–[6], [16], [17]) considered that emotion events are pervasive through a given time window. The better effectiveness of the proposed MIL approaches was confirmed, especially in a gold-standard scenario (DEAP dataset) using a *user-specific* setup for solving the *valence* and *arousal* task (see Table 4). Accordingly, the classification of the *arousal* is more challenging and the MIL is not always statistically superior to the best standard supervised ML. However, the MIL approaches disclose higher reliability for solving the *arousal* task in an experiment closer to



(a) The most discriminative features selected by mi-SVM with $L = 3$ for solving the *arousal* task for video 6.



(b) The most discriminative features selected by MI-SVM with $L = 5$ for solving the *valence* task for video 19.

Fig. 9: Feature importance: the most discriminative features across subjects. The feature importance was computed according to the absolute value of the mi-SVM and MI-SVM coefficients. For each subject, the absolute coefficient values were normalised from 0 to 1.

the real-world usage by using a purposely built dataset (Consumer dataset) (see Table 7). The MIL approaches are able to generalise across the unseen subject in the real-world scenario (Consumer dataset, *user-independent* setup, LOSO evaluation). However, the high difficulty of this task due to intrinsic inter-subject variability leads to the high variance of the extracted results (see Table 10). In particular, the obtained results (see Table 4, Table 7 and Table 10) suggest that the MIL algorithms are able to learn the most prominent emotional events that lead to better displaying the presence of self-reported emotions. Differently, the standard supervised learning approaches (e.g. [4]–[6], [16], [17]) are not always able to model the affective change that occurs. The labels are used to annotate the entire segment of physiological data, and the algorithm learns a global emotional response that is not always representative of the self-reported emotion. In other words, the proposed MIL approaches act at the bag level to classify the overall emotional response taking into account the local response, while the standard supervised learning approaches do not consider the local information represented by every single instance, but only provide a global response averaged over the entire video sequence.

6.2 Valence vs Arousal recognition tasks

The arousal recognition task discloses a relatively lower *accuracy* compared to *valence* classification (see Table 4 and Table 7). This is in line with the results obtained by [4] and with recent findings in the field of cognitive sciences. These findings confirm how the self-reports of arousal hardly predict its physiological values [68], [69]. However, the mi-SVM approach is above ($p < .05$) chance level (i.e. *macro-F1* = .5) in the gold-standard scenario

(DEAP), although the gain with respect to the standard supervised classifiers is not statistically superior.

The proposed MIL SVM extensions (i.e. mi-SVM, MI-SVM) seem to be more reliable with respect to the EMDD-SVM and mil-Boost, especially for the gold-standard scenario (see Table 4). However, they suffer from the unbalanced setting related to the arousal task in the Consumer dataset (i.e. 32% low *arousal* vs 68% high *arousal*) (see Figure 2a). In particular, the MIL SVM-based approaches are more sensitive than the EMDD-SVM (i) to the higher noise present in the self-report arousal response [68], [69], (ii) to the unbalanced setting (see Figure 2a) and (iii) to the higher variability/noise of the *user-independent* setup (LOSO evaluation) (see Table 10). This may be explained by the implicit feature selection performed by the EMDD algorithm. The scaling vector (s_k^2) gives weight to the features in order to maximize DD. This pre-processing step could lead to a decrease in the generalisation error across unseen subjects, especially in the presence of a high unbalanced dataset [70]. This fact is confirmed by the LOSO evaluation results (see Table 10), where the EMDD-SVM achieved the best *accuracy* for solving the *arousal* task.

6.3 Towards the real world usage

The benefits of the MIL apply to real-life application. Despite the lower *accuracy* of the sensors and the lesser consistency of the labeling procedure, the proposed MIL methodology is able to predict reliably (above chance level) across unseen trials the *valence* state (mi-SVM, *macro-F1* = .662) and at the same time provide a reliable estimation for the *arousal* state (mi-SVM, *macro-F1* = .637) (see Table 7). The performance for the arousal task is also statistically higher than that obtained by standard supervised ensemble algorithms widely used in the affective computing scenario (i.e. RF) [8]–[10]. Accordingly, the MIL approach is able to predict reliably (above chance level) (see Table 10) the *valence* state (MI-SVM, *accuracy* = .680) and *arousal* state (EMDD-SVM, *accuracy* = .691) of subjects left out from the training phase (user-independent setup). The largest improvement obtained by the MIL-based methods for the arousal task can be motivated by (i) the longer duration of the movie (i.e. 4 min) with respect to the DEAP dataset (i.e. 1 min) and (ii) the higher level of noise during the labeling procedure. Because the assumption of the standard supervised learning approach is that the emotion is pervasive within the window of time during which the label was gathered, this can lead to an increase in noise when the emotional stimuli have a longer duration. On the other hand, the proposed MIL approaches aim to capture only the most prominent event without assuming that the emotional response is constant over each frame. As a consequence, the physiological pattern can be captured with a higher resolution in time with respect to the overall sequence, leading to a better discrimination of self-reported emotion.

6.4 Structure of the bag

The performance of the MIL method is influenced by the structure of the bag. Hence, changing the size of each window can help to better model the local nature of the emotional response. The experimental comparison performed with different window lengths ($WS = 640$, $WS = 1280$, $WS = 2560$ and $WS = 3840$) and overlap factors ($OF = 0$, $OF = 1/2$ and $OF = 3/4$) (see Figure 6) confirms this hypothesis in all the performed analyses. Thus, a high number of instances representing a short time window of physiological response may lead to an increase in time resolution to capture the most prominent event. On the other hand, a too-short time window of the physiological signal (e.g. $WS = 1280$ and $WS = 640$) leads to the extraction of a lower frequency resolution of some salient features (e.g. the frequency spectrum of the IBI of the BVP signal). Accordingly, a low number of longer instances may lose salient information related to the local physiological response. This is especially confirmed for solving the *valence* task in a SIL instance learning scenario and with $L = 2$ (i.e. $OF = 0$ and $WS = 3840$) (see Figure 6b). The choice of the number of instances for each physiological signal (synchronously recorded over each video) can change for each subject and depends on several factors, such as the type and duration of stimuli. Although we demonstrated experimentally how the optimal number of instances falls within the implemented experimental setting ($L = 2$, $L = 3$ and $L = 5$) (see Figure 6a and Figure 6b), an intriguing future direction, the author aims might consist of selecting the correct number of instances maximising information theory metrics, such as entropy and mutual information.

6.5 Localisation of the most prominent emotional response and selection of the most relevant features

The mi-SVM and MI-SVM methods are natural MIL extensions of the standard supervised SVM. However, although both methodologies were used for the classification task at the bag level, they were also conceived for predicting instances' labels. Thus, the mi-SVM and MI-SVM were exploited not only for discriminating the affective response but also for localising the most prominent emotional responses inside each bag. The pattern localisation results highlight how the predictive power of each instance is not constant, thus leading to the discovery of the time interval of the video stimuli, where the physiological pattern of interest is more strongly displayed for discriminating the self-reported positive *valence* and the high *arousal* level (see Figure 7). Moreover, we found that in some cases, there is a sort of temporal correlation about the localisation of the self-reported emotion for a different subject while watching the same movie (see Figure 7c and Figure 8b).

A further analysis highlighted how the different behaviours of the MIL methods in choosing the most prominent instances can be associated with different changes in the physiological features (see Figure 8). Accordingly, in some cases where there is a sort of agreement concerning how the MIL localises the most prominent event between two different subjects, the physiological features reflect this agreement by showing a similar trend (see Figure 8b). Starting from this temporal analysis, we go further to determine whether the MIL methods agree on which discriminative features are the best, despite their different choices of instances (see Figure 9). Overall, for the *arousal* task, there is no agreement on the feature importance across subjects. This may further support the high variability between subjects and, hence, the higher difficulty in solving the *arousal* task (see Figure 9a). The agreement increases in the *valence* task, where the MI-SVM selects the EMG signal feature (i.e. spectral power over 20 Hz of the EMG signal [41]) as the most discriminative feature (see Figure 9b). The feature importance results highlight how the EMG-based features contain higher discriminative information to solve the *valence* (see Figure 9b) task.

Continuous human emotion recognition based on sequential learning approaches requires a costly labeling procedure and a

huge amount of data [24], which is not always feasible and scalable in real-life scenarios, where available labels are instead sparse and possibly describe only the most important events within a window of time. Despite the sparsity of these labels, many approaches (e.g. [4]–[6], [16], [17]) considered every single label as representative of each data frame that falls within the window of time during which the label was gathered. According to the length of such time window, it is likely that changes in the affective states may have instead occurred [71]–[73] with the label representing often noise rather than ground truth. Hence, supervision is provided only for the entire set of the frame (i.e. bag/video), and the individual label of the instances enclosed in the video is not provided. Thus, the self-reported overall emotional response cannot be always identified by a single observation, but by the interaction or the combination of several instances that are able to capture a local emotional response over time.

The work presented in this paper aimed to solve this challenge by:

- introducing the MIL methodologies for modeling only the most prominent emotional events rather than the continuous affective changes that occur;
- improving the emotion recognition performance with respect to standard supervised classifiers widely used for the categorical emotion recognition task;
- encouraging the application of this methodology in a setup closer to the real-world scenario, i.e. using an unobtrusive smartwatch sensor (worn on the subject's wrist) that is more prone to noise but acceptable for everyday usage.

Future works may be devoted to extending the MIL-based approach in a different scenario. This would involve modeling, discriminating and localising self-reported emotion. Following this direction we plan to consider experiments in which it might be possible to validate the predicted instance labels from MIL with a sort of ground truth labels collected in a fully annotated dataset. Identifying when the emotional response occurs within the labeled time window could further inform the personalisation that technology can provide to the person. For example, in the case of physical rehabilitation, knowing when a person becomes anxious during an exercise could help the rehabilitation system to understand in what part of the exercise routine the person may psychologically struggle. This could lead to the proposal of a simplified version of that exercise part to expose the patient gradually to the movement as a physiotherapist would do [74] or to provide other types of support, such as breathing reminders to reduce anxiety and tension. Similarly, companies interested in evaluating their advertisements could understand better what part of an advertisement is more effective and what is not [75].

Another interesting future direction would be to extend the methodology into a multi-instance multi-label formulation [76], where the emotional response is described by multiple instances and associated with multiple class labels. Accordingly, the MIL approach could be extended to map the dimensional perspectives of the emotion. The natural extension is to formulate the problem as a multiple instance regression task [27], [77], [78].

Finally, it would be interesting to integrate our approach with approaches aimed to tackle the spatial ambiguity due to sparse labeling. This is highly important for multi-dimensional signals, such as videos [32]–[34] and texts [35], [36], as discussed in Section 2, and also motion capture data [79], where emotional states within a frame are represented by expression peaks in certain parts of that frame (e.g. the smile and the eyes regions rather than the full facial expressions or the bending of the body trunk rather than the full body configuration [80]). Rather than simply combining MIL that works at the spatial level with MIL that works at the temporal level, it would be interesting to explore how both spatial and temporal information can be integrated to solve both ambiguities.

In conclusion, our work aimed to propose a new way of dealing with the problem of sparse labeling over time. The results are highly promising and open new possible directions for tackling the modeling of real-life datasets.

ACKNOWLEDGMENTS

Authors would like to thank Federica Verdini and Lucio Ciabattoni for their support in the recruitment of participants and data collection and Marianna Capecci for her advice during the Consumer experiment and visions for possible future use of this design protocol not discussed in this paper. This work was supported in part by EPSRC Grant Number EP/P009069/1.

REFERENCES

[1] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner *et al.*, "The humaine database: addressing the collection and annotation of naturalistic and induced emotional data," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2007, pp. 488–500.

[2] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[3] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The unbc-mcmaster shoulder pain expression archive database," in *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*. IEEE, 2011, pp. 57–64.

[4] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[5] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

[6] Y. Liu and O. Sourina, "Eeg databases for emotion recognition," in *International Conference on Cyberworlds (CW)*. IEEE, 2013, pp. 302–309.

[7] M. S. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng, A. Kemp, M. Shafizadeh, A. C. Elkins, N. Kanakam, A. de Rothschild, N. Tyler, P. J. Watson, A. C. d. C. Williams, M. Pantic, and N. Bianchi-Berthouze, "The automatic detection of chronic pain-related expression: requirements, challenges and the multimodal emopain dataset," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 435–451, 2016.

[8] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "Avec 2015: The 5th international audio/visual emotion challenge and workshop," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1335–1336.

[9] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

[10] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 3–9.

[11] Y. Cho, N. Bianchi-Berthouze, and S. J. Julier, "Deepbreath: Deep learning of breathing patterns for automatic stress recognition using low-cost thermal imaging in unconstrained settings," in *Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 456–463.

[12] J. J. Rivas, F. Orihuela-Espina, L. Palafox, N. Berthouze, M. d. C. Lara, J. Hernández-Franco, and E. Sucar, "Unobtrusive inference of affective states in virtual rehabilitation from upper limb motions: A feasibility study," *IEEE Transactions on Affective Computing*, pp. 1–1, 2018.

[13] D. McDuff, R. Kaliouby, T. Senechal, M. Amr, J. Cohn, and R. Picard, "Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 881–888.

[14] A. Singh, N. Bianchi-Berthouze, and A. C. Williams, "Supporting everyday function in chronic pain using wearable technology," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2017, pp. 3903–3915.

[15] T. Olugbade, N. Berthouze, N. Marquardt, and A. Williams, "Human observer and automatic assessment of movement related self-efficacy in chronic pain: from exercise to functional activity," *IEEE Transactions on Affective Computing*, pp. 1–1, 2018.

[16] W. Wen, G. Liu, N. Cheng, J. Wei, P. Shangguan, and W. Huang, "Emotion recognition based on multi-variant correlation of physiological signals," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 126–140, 2014.

[17] A. Sano and R. W. Picard, "Stress recognition using wearable sensors and mobile phones," in *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2013, pp. 671–676.

[18] T. A. Olugbade, A. Singh, N. Bianchi-Berthouze, N. Marquardt, M. S. H. Aung, and A. C. D. C. Williams, "How can affect be detected and represented in technological support for physical rehabilitation?," *ACM Trans. Comput.-Hum. Interact.*, vol. 26, no. 1, pp. 1:1–1:29, Jan. 2019.

[19] G. N. Yannakakis and J. Hallam, "Entertainment modeling through physiology in physical play," *International Journal of Human-Computer Studies*, vol. 66, no. 10, pp. 741–755, 2008.

[20] H. J. Griffin, M. S. H. Aung, B. Romera-Paredes, C. McLoughlin, G. McKeown, W. Curran, and N. Bianchi-Berthouze, "Perception and automatic recognition of laughter from whole-body motion: Continuous and categorical perspectives," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 165–178, 2015.

[21] J. M. Zacks and B. Tversky, "Event structure in perception and conception." *Psychological bulletin*, vol. 127, no. 1, p. 3, 2001.

[22] R. Vallacher and D. M. Wegner, "What do people think they're doing? action identification and human behavior," *Psychological Review*, vol. 94, pp. 3–15, 01 1987.

[23] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1, pp. 31–71, 1997.

[24] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int. J. Synth. Emot.*, vol. 1, no. 1, pp. 68–99, Jan. 2010.

[25] J. D. Velásquez, "Modeling emotions and other motivations in synthetic agents," in *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*, ser. AAAI'97/IAAI'97, 1997, pp. 10–15.

[26] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329 – 353, 2018.

[27] F. Herrera, S. Ventura, R. Bello, C. Cornelis, A. Zafra, D. Sánchez-Tarragó, and S. Vluymans, "Multi-instance regression," in *Multiple Instance Learning*. Springer, 2016, pp. 127–140.

[28] T. Tong, R. Wolz, Q. Gao, R. Guerrero, J. V. Hajnal, D. Rueckert, A. D. N. Initiative *et al.*, "Multiple instance learning for classification of dementia in brain mri," *Medical image analysis*, vol. 18, no. 5, pp. 808–818, 2014.

[29] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in neural information processing systems*, 2003, pp. 577–584.

[30] A. Zafra, C. Romero, and S. Ventura, "Multiple instance learning for classifying students in learning management systems," *Expert Systems with Applications*, vol. 38, no. 12, pp. 15 020–15 031, 2011.

[31] B. Babenko, M.-H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 8, pp. 1619–1632, 2011.

[32] T. Rao, M. Xu, H. Liu, J. Wang, and I. Burnett, "Multi-scale blocks based image emotion classification using multiple instance learning," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 634–638.

[33] A. Ruiz, J. Van de Weijer, and X. Binefa, "Regularized multi-concept mil for weakly-supervised facial behavior categorization." in *BMVC*, 2014.

[34] C. Wu, S. Wang, and Q. Ji, "Multi-instance hidden markov model for facial expression recognition," in *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 1. IEEE, 2015, pp. 1–6.

[35] N. Pappas and A. Popescu-Belis, "Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language processing (EMNLP)*, 2014, pp. 455–466.

[36] Y. Zhang, A. C. Surendran, J. C. Platt, and M. Narasimhan, "Learning from multi-topic web documents for contextual advertisement," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 1051–1059.

[37] T. Simon, M. H. Nguyen, F. De La Torre, and J. F. Cohn, "Action unit detection with segment-based svms," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 2737–2744.

[38] K. Sikka, A. Dhall, and M. Bartlett, "Weakly supervised pain localization using multiple instance learning," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–8.

[39] K. Sikka, A. Dhall, and M. S. Bartlett, "Classification and weakly supervised pain localization using multiple segment representation," *Image and Vision Computing*, vol. 32, no. 10, pp. 659–670, 2014.

[40] K. Sikka, G. Sharma, and M. Bartlett, "Lomo: Latent ordinal model for facial analysis in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5580–5589.

[41] A. Ruiz, O. Rudovic, X. Binefa, and M. Pantic, "Multi-instance dynamic ordinal random fields for weakly-supervised pain intensity estimation," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 171–186.

[42] A. Ruiz, O. Rudovic, X. Binefa, and M. Pantic, "Multi-Instance Dynamic Ordinal Random Fields for Weakly-supervised Facial Behavior Analysis," *ArXiv e-prints*, Feb. 2018.

[43] B. Wu, E. Zhong, A. Horner, and Q. Yang, "Music emotion recognition by multi-label multi-layer multi-instance multi-view learning," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 117–126.

[44] J. Gibson, A. Katsamanis, F. Romero, B. Xiao, P. Georgiou, and S. Narayanan, "Multiple instance learning for behavioral coding," *IEEE Transactions on Affective Computing*, 2015.

[45] C.-C. Lee, A. Katsamanis, M. P. Black, B. R. Baucom, P. G. Georgiou, and S. S. Narayanan, "Affective state recognition in married couples' interactions using pca-based vocal entrainment measures with multiple instance learning," in *Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 31–41.

[46] C. Zhang, J. C. Platt, and P. A. Viola, "Multiple instance boosting for object detection," in *Advances in neural information processing systems*, 2006, pp. 1417–1424.

[47] H. Hajimirsadeghi, J. Li, G. Mori, M. Zaki, and T. Sayed, "Multiple instance learning by discriminative training of markov networks," in *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'13. Arlington, Virginia, United States: AUAI Press, 2013, pp. 262–271.

[48] J. Gibson and S. Narayanan, "Learning multiple concepts with incremental diverse density," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4558–4562.

[49] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Advances in neural information processing systems*, 1998, pp. 570–576.

[50] H. Liu, H. Tang, W. Xiao, Z. Guo, L. Tian, and Y. Gao, "Sequential bag-of-words model for human action classification," *CAAI Transactions on Intelligence Technology*, vol. 1, no. 2, pp. 125–136, 2016.

[51] J. Joshi, R. Goecke, G. Parker, and M. Breakspear, "Can body expressions contribute to automatic depression analysis?" in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, April 2013, pp. 1–7.

[52] Q. Zhang and S. A. Goldman, "Em-dd: An improved multiple-instance learning technique," in *Advances in neural information processing systems*, 2002, pp. 1073–1080.

[53] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola, "Multi-instance kernels," in *In Proc. 19th International Conf. on Machine Learning*. Morgan Kaufmann, 2002, pp. 179–186.

[54] O. Maron, "Learning from ambiguity," Ph.D. dissertation, Massachusetts Inst. Techno., Cambridge, MA, USA, 1998.

[55] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, no. Jul, pp. 2079–2107, 2010.

[56] M. Kandemir, A. Vetek, M. Gönen, A. Klami, and S. Kaski, "Multi-task and multi-view learning of user state," *Neurocomputing*, vol. 139, pp. 97–106, 2014.

[57] V. Vapnik, *The Nature of Statistical Learning Theory*, ser. Information Science and Statistics. Springer New York, 1999. [Online]. Available: https://books.google.it/books?id=sna9BaxVbj8C

[58] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[59] "Microsoft band sdk," last accessed 21 August 2018. [Online]. Available: http://developer.microsoftband.com

[60] "Microsoft band synch application," last accessed 30 May 2019. [Online]. Available: https://www.microsoft.com/it-it/download/details.aspx?id=44579

[61] M. Soleymani, F. Villaro-Dixon, T. Pun, and G. Chanel, "Toolbox for emotional feature extraction from physiological signals (teap)," *Frontiers in ICT*, vol. 4, 2017.

[62] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *Journal of Experimental Social Psychology*, vol. 49, no. 4, pp. 764–766, 2013.

[63] M. Brennan, M. Palaniswami, and P. Kamen, "Do existing measures of poincare plot geometry reflect nonlinear features of heart rate variability?" *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 11, pp. 1342–1347, 2001.

[64] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[65] C. Godin, F. Prost-Boucle, A. Campagne, S. Charbonnier, S. Bonnet, and A. Vidal, "Selection of the most relevant physiological features for classifying emotion," *Emotion*, vol. 40, p. 20, 2015.

[66] C. A. Torres-Valencia, H. F. Garcia-Arias, M. A. A. Lopez, and A. A. Orozco-Gutiérrez, "Comparative analysis of physiological signals and electroencephalogram (eeg) for multimodal emotion recognition using generative models," in *XIX Symposium on Image, Signal Processing and Artificial Vision (STSIVA)*. IEEE, 2014, pp. 1–5.

[67] J. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," *Advances in Large Margin Classifiers*, pp. 61–74, 2000.

[68] A. Kron, M. Pilkiw, J. Banaei, A. Goldstein, and A. K. Anderson, "Are valence and arousal separable in emotional experience?" *Emotion*, vol. 15, no. 1, p. 35, 2015.

[69] A. Kron, A. Goldstein, D. H.-J. Lee, K. Gardhouse, and A. K. Anderson, "How are you feeling? revisiting the quantification of emotional qualia," *Psychological science*, vol. 24, no. 8, pp. 1503–1511, 2013.

[70] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Special issue on learning from imbalanced data sets," *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 1–6, 2004.

[71] R. Cai, C. Zhang, C. Wang, L. Zhang, and W.-Y. Ma, "Musicsense: contextual music recommendation using emotional allocation modeling," in *Proceedings of the 15th ACM international conference on Multimedia*. ACM, 2007, pp. 553–556.

[72] L. Lu, D. Liu, and H.-J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 1, pp. 5–18, 2006.

[73] E. M. Schmidt and Y. E. Kim, "Prediction of time-varying musical mood distributions using kalman filtering," in *9th International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2010, pp. 655–660.

[74] A. Singh, S. Piana, D. Pollarolo, G. Volpe, G. Varni, A. Tajadura-Jiménez, A. C. Williams, A. Camurri, and N. Bianchi-Berthouze, "Go-with-the-flow: tracking, analysis and sonification of movement and breathing to build confidence in activity despite chronic pain," *Human–Computer Interaction*, vol. 31, no. 3-4, pp. 335–383, 2016.

[75] A. Bleier and M. Eisenbeiss, "Personalized online advertising effectiveness: The interplay of what, when, and where," *Marketing Science*, vol. 34, no. 5, pp. 669–688, 2015.

[76] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li, "Multi-instance multi-label learning," *Artificial Intelligence*, vol. 176, no. 1, pp. 2291–2320, 2012.

[77] S. Ray and D. Page, "Multiple instance regression," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01, 2001, pp. 425–432.

[78] D. R. Dooly, Q. Zhang, S. A. Goldman, and R. A. Amar, "Multiple-instance learning of real-valued data," *Journal of Machine Learning Research*, vol. 3, no. Dec, pp. 651–678, 2002.

[79] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 15–33, 2013.

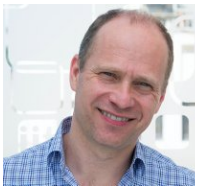
[80] T. A. Olugbade, N. Bianchi-Berthouze, N. Marquardt, and A. C. Williams, "Pain level recognition using kinematics and muscle activity for physical rehabilitation in chronic pain," in *International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 243–249.

Luca Romeo received a Ph.D. degree in computer science from the Department of Information Engineering (DII), Università Politecnica delle Marche, in 2018. His Ph.D. thesis was on "applied machine learning for human motion analysis and affective computing". He is currently a PostDoc Researcher with DII and he is affiliated with the Unit of Cognition, Motion and Neuroscience and Computational Statistics and Machine Learning, Fondazione Istituto Italiano di Tecnologia Genova. His research topics include Machine learning applied to biomedical applications, affective computing and motion analysis.

Andrea Cavallo is Assistant Professor at the Department of Psychology of the University of Turin. He received his master's degree in Clinical Psychology in 2008 from the University of Padua and a Ph.D. in Neuroscience in 2013 from the University of Turin. His research focuses on the processes that transform external sensory information into the correspondent internal motor representation (i.e. motor cognition). He published 36 papers in peer-reviewed journals (e.g. NeuroImage, Cerebral Cortex, Physics of Life Reviews). He is currently working on new approaches to investigate motor-related components in both clinical and non-clinical populations. As Principal Investigator, he is involved in the Find.Aut (Finding the motor signature of Autism) project, with the aim of studying early motor anomalies in people with Autism Spectrum Disorder.

Lucia Pepa Lucia Pepa, born in Jesi (AN) on 10 August 1988, received in 2010 the first-level degree in Biomedical Engineering (cum laude), in 2012 the Master Engineering degree in Electronic Engineering (cum laude), in 2016 a Ph.D. degree in E-learning – Technology Enhanced Learning from the Universita' Politecnica Marche, Italy. She is currently a postdoc researcher at Università Politecnica delle Marche and her primary research interests involve affective computing and movement analysis through consumer electronics devices.

Nadia Bianchi-Berthouze is a professor in Affective Computing and Interaction. She has pioneered the field of Affective Computing from both the machine learning and HCI perspectives. She has published more than 200 papers in affective computing, human-computer interaction and pattern recognition. She has investigated affect-aware technology in real-life contexts: e.g. EPSRC-funded Emo&Pain and H2020-funded EnTimeMent on affective-aware rehabilitation technology; EPSRC-funded Digital Sensoria on biosensors to capture subjective responses to tactile experiences; and H2020 HU-MAN Manufacturing to measure stress in the industry workfloor contexts.



Massimiliano Pontil is a Senior Researcher at Istituto Italiano di Tecnologia and Professor of Computational Statistics and Machine Learning in the Department of Computer Science at University College London. His research interests are in the areas of machine learning with a focus on regularisation methods, convex optimisation and statistical learning theory. He has been on the programme committee of the main machine learning conferences, including COLT, ICML and NIPS, he is an Associate Editor of the Machine Learning Journal, an Action Editor for the Journal of Machine Learning Research and he is on the Scientific Advisory Board of the Max Planck Institute for Intelligent Systems Germany.