This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2023.3297270

Multi-Rater Consensus Learning for Modeling Multiple Sparse Ratings of Affective Behaviour

Luca Romeo, Temitayo Olugbade, Massimiliano Pontil, Nadia Bianchi-Berthouze

Abstract—The use of multiple raters to label datasets is an established practice in affective computing. The principal goal is to reduce unwanted subjective bias in the labelling process. Unfortunately, this leads to the key problem of identifying a ground truth for training the affect recognition system. This problem becomes more relevant in a sparsely-crossed annotation where each rater only labels a portion of the full dataset to ensure a manageable workload per rater. In this paper, we introduce a Multi-Rater Consensus Learning (MRCL) method which learns a representative affect recognition model that accounts for each rater's agreement with the other raters. MRCL combines a multitask learning (MTL) regularizer and a consensus loss. Unlike standard MTL, this approach allows the model to learn to predict each rater's label while explicitly accounting for the consensus among raters. We evaluated our approach on two different datasets based on spontaneous affective body movement expressions for pain behaviour detection and laughter type recognition respectively. The two naturalistic datasets were chosen for the different forms of labelling (different in affect, observation stimuli, and raters) that they together offer for evaluating our approach. Empirical results demonstrate that MRCL is effective for modelling affect from datasets with sparsely-crossed multi-rater annotation.

Index Terms—Multiple labels, multitask learning, sparse raters ratings, body expressions, pain behaviour, protective behaviour, laughter types.

1 Introduction

The use of multiple raters to label datasets is an established practice in affective computing [1]. There are two strategies typically used. Where possible, each rater is asked to label every instance of the dataset in what we hereafter refer to as the *fully-crossed annotation design*. Unfortunately, this strategy is not usually practical, especially for large datasets. To ensure that each rater is assigned a manageable workload, labelling tasks are often designed such that each rater only labels a portion of the full dataset [2] and there are different sets of raters across data instances in the dataset. We hereafter refer to this as *sparsely-crossed annotation design*. The common challenge for both scenarios is how to decide the ground truth to train the learning model. The standard method is majority voting. This method is limited in its capture of the disagreement between raters. This limitation becomes even clearer in sparsely-crossed annotated datasets where the number of raters per instances is small and hence without a modal label that is representative at the population level. Other methods aim to predict the rating distribution and aggregate the predictions. While this approach may gain from the modeling process of each rater, it may struggle to generalize well under the condition of small number of instances per rater [3]. Another typical approach has been to avoid the use of a unique ground truth and instead focus on modeling the distribution of the labels over an instance [4]. This approach may be interesting in modeling population perception of an expression; however, it may be limited in informing the action to be taken in the case of technology-based intervention (e.g., how to tailor suggestions for movement execution on the basis of a distribution of pain behaviour interpretations).

In this paper, we address the above limitations by proposing an alternative method specifically formulated for the case of a sparsely-crossed annotation design. Rather than simply identifying a possible ground truth, our method creates an affective behaviour recognition model that is more representative of individual behaviour by accounting for each specific rater's agreement with the other raters across all the instances that the given rater labelled.

We explore our proposed method using two naturalistic datasets, pain behaviour detection and laughter type recognition datasets, with clinician and naive raters respectively. The pain behaviour dataset was annotated in a sparsely-crossed design to minimize the burden on expert raters. The laughter dataset, being instead annotated using a fully-crossed design, allowed for simulation of various sparsely-crossed subsets to further evaluate the benefits and shortcomings of our approach. Additional rationale behind the selection of the two datasets is that they are together representative of the different levels of interrater agreement observed in naturalistic datasets. The reasons for these different levels of agreements are mainly due to the difference in amount of contextual information available to the raters during the labelling process, and to the level of expertise of and common ground between the raters.

Our approach makes a methodological contribution, that is the combination of multitask learning (MTL) with a novel consensus loss. Specifically, we employ a variance regularizer to model the dependencies across multiple tasks (where the tasks are raters) and encourage the model to maximise a consensus loss in the validation stage. This is unlike standard MTL where the loss function decomposes across tasks. Our approach is also in contrast to standard methods which either only learn the majority vote across multiple raters or only take into account each rater

- *L. Romeo is with Università degli Studi di Macerata, Macerata, Italy and Fondazione Istituto Italiano di Tecnologia, Genova, Italy.*
  T. Olugbade is with University of Sussex, and also has a honorary research fellow position at University College London (UCL), UK.
  N. Bianchi-Berthouze is with the UCL Interaction Centre, UCL, UK.
  M. Pontil is with Fondazione Istituto Italiano di Tecnologia, Genova, Italy, and Department of Computer Science, University College London, UK.
  Corresponding to L. Romeo (E-mail: luca.romeo@unimc.it).

independently without explicitly addressing the consensus among raters. By simultaneously learning multiple related ratings the proposed method provides a strategy to alleviate data insufficiency by aggregating data from multiple raters and improve generalization performance. Thus, it represents a viable solution with, and especially advantageous for, a limited number of raters per data instance compared with the total number of raters, i.e., a sparsely-crossed annotation situation.

Contributions. In summary the main contributions of this work are as follows:

- We propose a Multi-Rater Consensus Learning (MRCL) method which jointly models each rater by leveraging similarities between raters in maximizing a raters consensus loss.
- With multiple experiments, we demonstrate the effectiveness of the proposed approach for sparsely-crossed annotation designs. Our experiments are based on two different naturalistic datasets, and the findings from them show how well the proposed method generalises to unseen data subjects.

Paper Organization. In Section 2, we provide a review of machine learning (ML) approaches used to model multiple ratings (see Section 2.1) and of MTL approaches used in affective computing studies (see Section 2.2). Next, in Section 3 we describe our MRCL approach, while in Section 4 we discuss how the approach is used on the pain behaviour and laughter datasets. The experimental procedure and metrics used are described in Section 5, while the results are reported in Section 6. In Section 7, we discuss the implications of our findings. Finally, in Section 8 we provide a conclusion.

2 Related Work

We review two threads of relevant prior work. One is the state of the art on modelling of ground truth data from multiple raters (see Section 2.1). The second focuses on how MTL, which our proposed method is based on, has typically been used in the context of automatic affect recognition (see Section 2.2).

2.1 Modeling Ratings from Multiple Raters

The modeling of multiple ratings is a topical problem although attention has usually been paid to crowdsourcing scenarios. Beyond the standard approach, e.g. in [5], [6], of combining the multiple ratings into a single label using majority voting, weighted aggregation, or pairwise combinations, there have been other approaches more geared towards filtering out noisy labels, such as the eigenvectors-based matrix completion approach of [7].

There are strategies that have aimed at estimating the unknown true label, for example, using active learning [8], expectation-maximization [9]. For instance [10], proposed a combinatorial aggregation model that infers the 'hidden' ground truth labels by mapping both the labels provided by raters and the labelled data instances into a low-dimensional space based on maximization of information theory metrics. Focusing on ordinal labels in a fully-crossed annotation design, in [11]'s approach, a unique aggregate label was determined by maximising minmax conditional entropy. [12] jointly modeled the rating of each data instance by providing the most likely single true label based on expectation-maximization. [13] similarly employed a Bayesian generative probabilistic model for learning a single latent rating for multiple available ratings. The common limitation of these approaches is that they require each rater to label every data instance, which is not always a feasible annotation strategy. They (e.g. [7], [10], [11]) are also typically based on the assumption that the central tendency of the collective of the multiple ratings is the most reliable label, to the exclusion of accounting for informative variability that can exist across different raters. Unlike these methods, our MRCL strategy aims to model both the consensus between raters as well as the distinctiveness of each rater. This is particularly important when it is difficult to quantify the a priori reliability of any single rater. In doing so, our method overcomes another limitation of those previous approaches: their inability to retrieve or predict the rating of each individual rater since they only learn a single global rating.

More related to our proposed approach is [14] which leverages the MTL approach to estimate both a global rating and individual ratings from different raters. However, our approach is different from their standard MTL where the loss function does not explicitly address similarities between raters by maximizing a raters consensus loss. Our MRCL allows the prediction of a specific rater's rating while also addressing explicitly the consensus of the given sample of raters. While they do not use the MTL mechanism, [15] also has a similar approach to ours with the modelling of both original multiple ratings and their aggregates. However, unlike our approach where these two processes are used to regularize one another through simultaneous modelling, in [15] the two processes are treated as separate models. They used a Hidden Markov Model (HMM) for the aggregate label and dealt with sparsely-crossed annotation for multiple ratings using imputation. Our proposed approach inherently addresses sparsity without need for imputation which can introduce noise. For the individual ratings, they employed a multi-label scheme based on a Conditional Random Field (CRF) with a multi-label LSTM. On the contrary, in our approach we directly learn the original multiple ratings and a global rating based on these, using a single MTL model where the global consensus among raters is imposed during training.

2.2 Multitask Learning Approach in the Affective Computing Scenario

MTL is a widely used machine learning paradigm within the affective computing literature. Several works adopt MTL for simultaneous learning of different but related phenomena/constructs. For instance, in the neural network model of [16] for continuous pain intensity estimation from video and physiological signals, each cluster of subjects with similar pain response profiles was a separate output of the network. Similarly, [17] modelled multiple emotional state indicators and multiple users jointly via a multiple kernel and multitask approach based on a Support Vector Machine. In [18], a joint representation learning strategy was proposed for modelling emotion and identity. [19] also used a MTL approach for learning both emotion labels and facial action units. Other examples include the models of [20], [21] (based on Recurrent Neural Networks) and of [22] (a Deep Belief Network) for jointly learning activation, valence, and acoustic emotion labels. These studies highlight the relevance of MTL for the problem of modelling different but related affective tasks.

The only study we found to have used MTL for modelling affective labels from multiple raters is the work of [23]. Our

work in this paper goes beyond their approach by additionally maximising a consensus among raters. In addition, instead of a neural network implementation, we use a linear model to ensure high interpretability of the model. In particular, our MRCL ensures more direct inferences at feature and rater levels as all raters share the same feature relevance space and feature relevance characterizes both individual ratings and the global rating simultaneously. Further, we evaluate and demonstrate the effectiveness of our approach on datasets with sparsely-crossed annotation design.

3 The Multi-Rater Consensus Learning (MRCL) Methodology

In this section, we formalize our MRCL method for modelling multiple labels from different raters. The method can also be used in the fully-crossed annotated design setting, the method is especially well suited to sparsely-crossed annotation designs, in which leveraging rater similarities enables more effective learning of individual raters.

Let N be the number of raters and J the number of instances in the given training dataset. We denote the label given by the n-th rater to the j-th training data instance $\mathbf{x}_j$ by $y_{n,j} \in \{-1,+1\}$. Let m_n be the number of instances labelled by this rater. We extend the binary classification formulation to the case of multiclass classification in which the rater has to choose a label from more than two classes (cf. the laughter dataset) using a one-vs-all paradigm with independent binary classification tasks where each class is discriminated from the rest.

To learn a single model for all raters, we follow a MTL formulation which minimizes the penalized empirical loss

$$\min_{\mathbf{w}_1,\ldots,\mathbf{w}_N} \sum_{n=1}^{N} \mathscr{L}_n(\mathbf{w}_n) + \Omega(\mathbf{w}_1,\ldots,\mathbf{w}_N) \tag{1}$$

where $\mathbf{w}_n$ is the model parameter for the n-th rater and $\mathscr{L}_n(\mathbf{w}_n)$ is the corresponding empirical loss. The regularization term $\Omega(\mathbf{w})$ encourages similarities between raters. In particular, we use the mean regularization approach [24], [25] and adopt the logistic regression function as the empirical loss:

$$\mathscr{L}_n(\mathbf{w}_n) = \sum_{n=1}^{N} \sum_{j=1}^{m_n} \log\left(1+\exp(-y_{n,j}(\mathbf{w}_n^\top \mathbf{x}_j))\right) \tag{2}$$

Note that we may add one additional component to the input which is always equal to one to add a threshold in the logistic regression function.

We follow the regularization formulation proposed by [24] which encourages the task parameters to be close to their mean. That is, we choose $\Omega(\mathbf{w}) = \lambda \sum_{n=1}^{N} \|\mathbf{w}_n - \frac{1}{N}\sum_{s=1}^{N} \mathbf{w}_s\|$, where $\sum_{s=1}^{N} \mathbf{w}_s$ is the average of the model parameters and the regularization parameter $\lambda > 0$ controls the deviation of each rater from the mean across the raters. More generally, we consider the a quadratic regularizer which encourages linear task relationships and penalizes the euclidean norm of each model, like in ridge regression,

$$\Omega(\mathbf{w}) = \rho_0 \|\mathbf{w}R\|_F^2 + \rho_1 \|\mathbf{w}\|_F^2 \tag{3}$$

where $\|\cdot\|_F$ denotes the Frobenious norm (i.e. the Euclidean norm of the vector formed by the matrix elements), the hyper-parameter ρ_1 controls the magnitude of the model parameters and the hyper-parameter ρ_0 encourages structure between the models. The matrix R encodes linear relatioships between the tasks. Notably, in the mean regularization formulation, the matrix encodes a complete graph structure as follows:

$$R = \mathbb{I}_N - \frac{\mathbb{1}_N}{N}$$

where $\mathbb{I}$ and $\mathbb{1}$ are the identity matrix and the matrix with all ones respectively.

In the validation stage, we aggregate the individual predictions of the raters and compute a validation loss that we use to optimize the hyperparameters ρ_0 and ρ_1 in the regularizer. Specifically, we tune the hyperparameters of the MRCL by maximizing prediction of the consensus among the raters (hereafter referred to as *global rating prediction*). We compute the individual rating prediction for the n-th rater for the i-th test instance as

$$\hat{y}_i = \text{sign}(\mathbf{w}_n^\top \mathbf{x}_i). \tag{4}$$

We then compute the global rating prediction as

$$\hat{y}_i^G = \text{sign}\left(\sum_{n=1}^{N} \text{sign}(\mathbf{w}_n^T \mathbf{x}_i) - \frac{N}{2}\right) \tag{5}$$

and finally compare it to the global rating ground truth

$$y_i^G = \text{sign}\left(\sum_{n=1}^{N_i} y_{n,i} - \frac{N_i}{2}\right) \tag{6}$$

where N_i is the subset of the N raters that label the i-th test instance. The global rating prediction is computed as the majority vote over the individual rating predictions for the given instance. In the same way, the global rating ground truth is based on a majority vote of the individual rating ground truth. Where there is a tie (i.e. no majority), the negative class is assumed. Then, the validation loss that we aim to minimize in the validation set reflects the misclassification error between global rating prediction and global rating ground truth, i.e.

$$\min_{\rho_0,\rho_1} \sum_{k=1}^{K} I(\hat{y}_k^G \neq y_k^G) \tag{7}$$

where K is the number of instances in the validation set.

It is worth noting that the proposed MRCL method can easily be extended to modelling of different but related affective labels. The natural extension will be to enforce similar weights for features not only across the different raters (as in our original MRCL) but also across the different labels. To achieve this, the R matrix should be structured as composition of two submatrices that encode the weights across raters and labels respectively. In such extension, the individual and global rating prediction for a given label can be computed by considering the subset of tasks related to the label.

4 Datasets

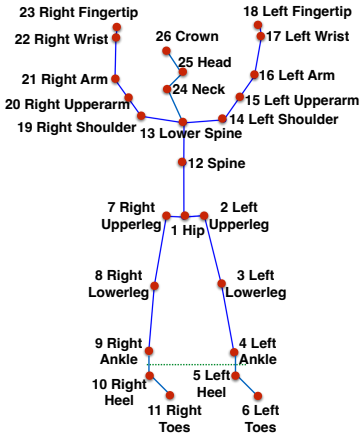
We base our investigation on two body movement datasets: a pain behaviour detection dataset and a laughter type recognition dataset. First, the datasets are particularly interesting because of the different forms of annotation and raters (detailed later in this section) that they offer for evaluating our approach. Second, they comprise naturalistic expressions, which are subject to higher disagreement between raters than acted and stereotypical expressions. Further, they present different levels of agreement between raters possibly because of the knowledge of the raters. The level of disagreement is critical since it is what calls for new methods for

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2023.3297270

addressing the complexity of affect ground truth. Nonetheless, our approach is general and can be applied to any type of multi-rater affect dataset regardless of the affect modality.

4.1 The EmoPain Dataset for Pain Behaviour Detection

The EmoPain dataset [26] is a multimodal dataset captured from people with chronic pain (CP) and healthy control participants carrying out physical exercises (e.g. one leg raised up while standing, sitting down) typical of those performed in clinical settings and representative of everyday movements (e.g. climbing the stairs). The dataset contains full-body kinematic data and rectified surface electromyographic (sEMG) signals from upper and lower back muscles. Both motion capture and sEMG data are sampled at 60Hz (the sEMG data was originally sampled at 1000Hz, but resampled to 60Hz for synchronization with the kinematic data). The kinematic data consists of 3D positions for 26 anatomical joints (see Table 1). We used the 78 (26x3D) joint positions and 4 sEMG signals averaged (over time) for each exercise as predictors in our experiments.



Features Name
X,Y,Z Hip
X,Y,Z Left Upperleg
X,Y,Z Left Lowerleg
X,Y,Z Left Ankle
X,Y,Z Left Heel
X,Y,Z Left Toes
X,Y,Z Right Upperleg
X,Y,Z Right Lowerleg
X,Y,Z Right Ankle
X,Y,Z Right Heel
X,Y,Z Right Toes
X,Y,Z Spine
X,Y,Z Lower Spine
X,Y,Z Left Shoulder
X,Y,Z Left Upperarm
X,Y,Z Left Arm
X,Y,Z Left Wrist
X,Y,Z Left Fingertip
X,Y,Z Right Shoulder
X,Y,Z Right Upperarm
X,Y,Z Right Arm
X,Y,Z Right Wrist
X,Y,Z Right Fingertip
X,Y,Z Neck
X,Y,Z Head
X,Y,Z Crown
EMG Right Lower
EMG Left Lower
EMG Right Upper
EMG Left Upper

TABLE 1: Left - The 26 anatomical joints captured for both the EmoPain and Laughter datasets. Right - List of the features used for our experiments on the EmoPain dataset: 3D joints positions from all 26 joints and sEMG data from the 4 body locations. The features for the Laughter dataset are in Table 2.

The annotation used from this dataset identifies events of one category of protective behaviours, *guarding* [26]. Guarding behaviour is the most prevalent of a larger set of bodily-expressed protective behaviours [27] and is a term that physiotherapists would generally use instead of more specific terms (guarding/stiffness, hesitation, abrupt action). The annotation was done for sit-to-stand, stand-to-sit, reaching forward exercise types and by 30 physiotherapists based on observations of RGB video recordings of the exercise sessions. The raters provided labels for each discrete exercise instance (e.g. one instance of sit-to-stand movement). This approach was taken rather than time-continuous labelling for exercise sessions for each data subject in order to reduce differences in segmentation across raters. Each exercise instance was labelled by a set of 4 raters randomly selected from the 30 physiotherapists, creating a sparsely-crossed annotation design to minimize the workload for each rater. The annotation procedure is described in more detail in [28] although with a focus there on self-efficacy labels (which are one of the other labels rated by the physiotherapists in that study). The guarding labels captured in that same study have not been published as yet, but will be made available (on request to the last author) together with the already open EmoPain dataset.

Figure 1 shows the distribution of exercise instances annotated across all the 30 raters. Figure 1 further shows the distribution of each label (guarding versus not guarding) for the exercise instances seen by each rater. One challenge with the labels is the presence of very few instances for each rater. As can be seen in Figure 2, 10 of the 30 raters saw (i.e. labelled) less than ten exercise instances. Given the low number of instances for these raters, we exclude them in our modelling and instead focus on the 20 raters with at least ten instances labelled. The rationale behind that is to have a minimum number of samples (i.e. 10) for each rater, for a representative performance evaluation.

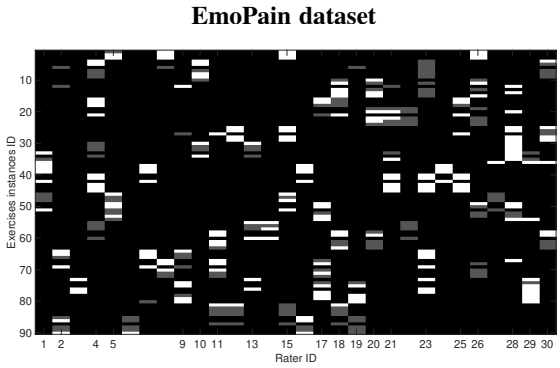


Fig. 1: A plot of the EmoPain dataset exercises instances (y-axis) against the 30 raters (x-axis). Only the selected 20 raters ID are indicated. Black dashes indicate that the given instance was not labelled by the given rater; White dashes indicate that the instance was labelled by the rater as containing guarding behaviour ; Grey dashes indicate that the instance was labelled by the rater as not containing guarding behaviour.

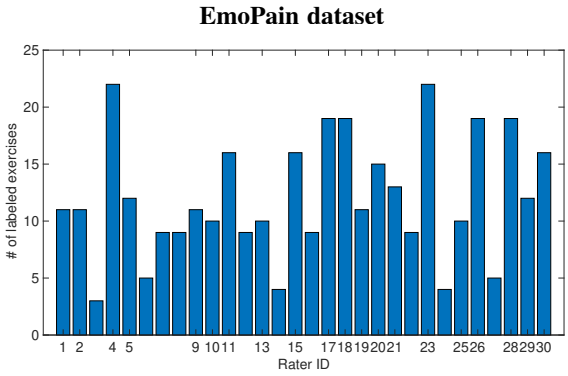


Fig. 2: The number # of labelled exercises (y axis) per rater (x axis) in the EmoPain dataset (only the selected 20 raters ID are indicated).The bar displays the number of white and grey dashes displayed in Figure 1

Figure 3 provides an illustration of our approach for detecting guarding behaviour in a sparsely-crossed annotation design EmoPain dataset.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2023.3297270

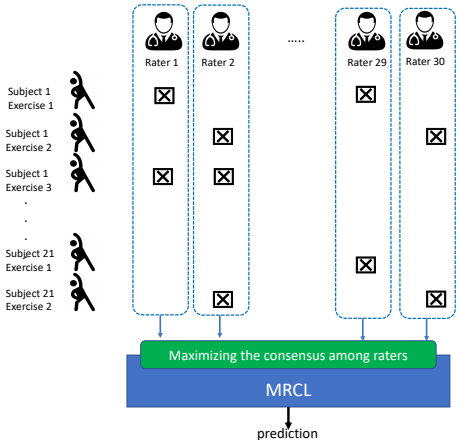


Fig. 3: An illustration of the MRCL approach using the EmoPain dataset. The MRCL jointly learns ratings from multiple raters and the final prediction is achieved by maximising the consensus among raters.

4.2 The Laughter Dataset for Laughter Type Recognition

The Laughter dataset [29], [30] is a dataset captured during dyadic activities inducing laughter and during conversation in between these activities. The dataset contains full-body kinematic data with sampling rate of 60Hz. Similar to the EmoPain dataset, the kinematic data consists of 3D positions for 26 anatomical joints (see Table 1). It further includes 31 features (see Table 2) (e.g. lower and upper-limb features, energy and smoothness measures) extracted from these anatomical joints. As the joint positions data were not available to us, we used the extracted features as predictors in our study. Although there were 18 participants in total, only data for one person from each dyad was captured. So, our investigation was based on data from 9 data subjects. The data collection protocol was designed to elicit different types of laughter: *hilarious*, *social*, *awkward*, *fake* or *no-laugh*. 126 laughter instances were extracted and categorized based on the laughter elicitation used (54, 42, 23 and 7 instances of non-laugh, social, hilarious and awkward respectively).

TABLE 2: List of included features for the Laughter dataset.

Features	Name
Min distance between left hand and head	minLHandHeadDist
Max distance between left hand and head	maxLHandHeadDist
Range distance between left hand and head	rangeLHandHeadDist
Min distance between right hand and head	minRHandHeadDist
Max distance between right hand and head	maxRHandHeadDist
Range distance between right hand and head	rangeRHandHeadDist
Min distance between left hand and hip	minLHandHipDist
Max distance between left hand and hip	maxLHandHipDist
Range distance between left hand and hip	rangeLHandHipDist
Min distance between right hand and hip	minRHandHipDist
Max distance between right hand and hip	maxRHandHipDist
Range distance between right hand and hip	rangeRHandHipDist
Min distance between hands	minHandDist
Max distance between hands	maxHandDist
Range distance between hands	rangeHandDist
Frequency Power (4-6Hz) shoulder displacement: mean of left	Lfour2sixHzPowerWelch
Frequency Power (4-6Hz) shoulder displacement: mean of right	Rfour2sixHzPowerWelch
Azimuthal rotation of shoulders in global space	rangeAzShouldLine
Min angle at upper spine joint	minSpineCurl
Max angle at upper spine joint	maxSpineCurl
Range angle at upper spine joint	rangeSpineCurl
Min angle at lower spine joint	minSpine1Curl
Max angle at lower spine joint	maxSpine1Curl
Range angle at lower spine joint	rangeSpine1Curl
Min angle at neck joint	minNeckCurl
Max angle at neck joint	maxNeckCurl
Range angle at neck joint	rangeNeckCurl
Min anterior-posterior component lower spine-upper spine	minCompSpineAng
Max anterior-posterior component lower spine-upper spine	maxCompSpineAng
Range anterior-posterior component lower spine-upper spine	rangeCompSpineAng
Correlation left and right shoulders superior-inferior displacement	shouldCorr

In our experiments, we used a second set of labels in the dataset. These labels were gathered from multiple observers recruited through a crowdsourcing engine. During the crowdsourced labelling process, each rater was asked to assign one of the 5 laughter types labels to each of the 126 animated avatars generated from the kinematic data. This dataset is interesting as the fully-crossed annotation design based on a large number of raters enabled a more thorough evaluation of our approach through controlled simulations of sparsely-crossed annotation. To simulate a sparsely-crossed annotation design from the original ratings, we randomly selected 4 raters for each data instance in the Laughter dataset, maintaining the same proportion of labels (i.e. same ratio of no-laugh, social and hilarious classes) with respect to the original dataset. The random selection was done 20 times over all data instances. The choice of 4 raters was made to be similar to the number of raters per instance for the EmoPain data. We hereafter refer to simulated sparsely-crossed annotation of the Laughter dataset as *Simulated-Sparse Laughter Labelling*. Similar to our approach with the EmoPain labels, we also applied to the Simulated-Sparse Laughter Labelling dataset the constraint of retaining only raters with at least 10 instances seen by each rater. Figure 4 shows the distribution of data instances across all the 39 raters, for one out of 20 repetitions. Figure 4 further shows the distribution of each labels (*hilarious*, *social*, *awkward*, *fake* or *no-laugh*) for the data instances seen by each rater. To test our MRCL approach with the Simulated-Sparse Laughter Labelling dataset, we exclude awkward and fake labels as those labels were rarely present in the labels (only 7 and 1 instances were labelled as awkward and fake respectively). Thus, we solved three independent binary laughter classification tasks: laugh versus no-laugh, social versus non-social, and hilarious versus non-hilarious.

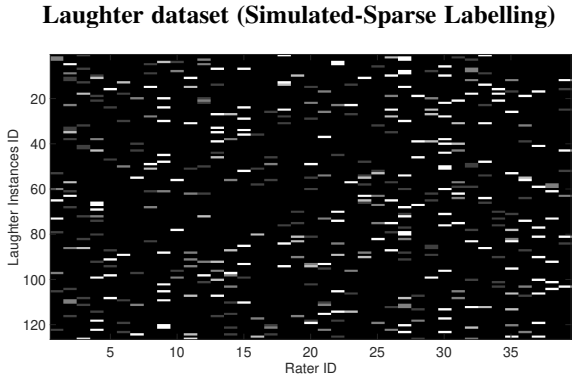


Fig. 4: A plot of instances in one Simulated-Sparse Laughter Labelling data subset (Simulation 1) against the 39 raters. Black dashes indicate that the given instance was not labelled by the given rater; Grey shade dashes indicate that the instance was labelled by the rater as either *hilarious*, *social*, *awkward*, or *fake*; White dashes indicate that the instance was labelled by the rater as *no-laugh*.

5 Experimental Procedure and Metrics

We ran two sets of performance evaluation [1]:

1. Python code used in the experiments will be made available at the following Github repository https://github.com/whylearning22/Multi-Rater-Consensus-Learning

- Global performance evaluation (GPE) - To understand how well the MRCL model represents consensus among raters, we evaluated the global rating prediction (see Equation 5) against the global rating ground truth (see Equation 6).
- Individual performance evaluation (IPE) - To understand how well the same model characterizes the different raters, we evaluated the individual rating predictions (see Equation 4) against the individual rating ground truth ($Y_{n,i}$).

For both GPE and IPE, the following metrics were used:

- *accuracy*: the percentage of correct predictions;
- *macro-F1 (F1)*: the harmonic mean of precision and recall averaged across all output classes;
- *macro-recall (recall)*: the recall is calculated for each output classes and the unweighted mean is then taken;
- *macro-precision (precision)*: the precision is calculated for each output classes and the unweighted mean is then taken;

For both the GPE and IPE, we performed leave-one-subject-out cross-validation (LOSO-CV). In LOSO-CV, an iterative selection of one subject for testing and the other subjects for training is done. The rationale behind the LOSO-CV is the need to account for individual differences and assess the generalizability of the model to an unseen subject. A nested LOSO-CV was implemented to select the best hyperparameters for each classification task by maximizing the F1 score of the GPE prediction. A list of all explored hyperparameters is reported in Table 3. For the EmoPain dataset, we computed the GPE and IPE metrics based on the sum of the confusion matrices for each cross-validation fold and the average of the individual F1 scores computed for each rater respectively. For the Laughter dataset, the metrics were further averaged across 20 repetitions of the randomly selected Simulated-Sparse Labelling data subsets from the Laughter dataset.

TABLE 3: Range of hyperparameters for each algorithm.

Model	Hyperparameters	Range
State-of-the-art		
RF	number of trees	$\{10, 20, 50, 100, 200\}$
	number of predictors to select	$\{20, 30, 40, 60, all\}$
Boosting	learning rate	$\{10^{-3}, 10^{-2}, 10^{-1}, 0.5, 1\}$
	number of learning cycles	$\{20, 50, 100, 200, 300\}$
Logistic	L2-norm regularizer (ρ)	$\{10^{-8}, 5 \cdot 10^{-8}, 10^{-7}, \ldots, 5\}$
MRCL (ours)		
MRCL	structure regularizer (ρ_0)	$\{10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10^1, 2 \cdot 10^1, \ldots, 10^2, 2 \cdot 10^2, \ldots, 10^3, 2 \cdot 10^3, \ldots, 10^4\}$
	L2-norm regularizer (ρ_1)	$\{10^{-8}, 5 \cdot 10^{-8}, 10^{-7}, \ldots, 5\}$

The MRCL was compared with state-of-the-art approaches [5], [6], [31] which only model the global label (hereafter referred to as Single Rater Learning, SRL). As the MRCL's global rating prediction is the most relevant for such comparison, we use the GPE. Although, the ridge logistic regression (Logistic), which is part of the formulation of the MRCL (see Section 3) is the most comparable SRL method, we additionally compare the MRCL with two other methods, Random Forest (RF) and Boosting. These comparisons were also motivated by preliminary tests that showed better performance for both RF and Boosting algorithms than Naïve Bayes, k-Nearest Neighbors, Gaussian Support Vector Machine, and Decision Tree.

We take two different approaches for learning the global label using the SRL methods:

1) Majority voting of individual ratings predictions (Multiple Models SRL) - Here, each individual rater's ratings are modelled separately and the majority vote of the predictions from the different models is then taken as the SRL's global rating prediction.
2) Prediction of majority vote ratings (Single Model SRL) - In this case, the global rating across the individual raters is modelled. The global rating is specified as the majority vote across the ground truth labels. The SRL global rating prediction is taken to be the predicted majority vote. Unlike with the Multiple Models SRL where aggregation is done post-modelling, here it is done pre-modelling.

Although the Single Model SRL is the traditional approach (e.g. [5], [6], [31]), we further performed the 'Multiple Models SRL' comparison for the sake of completeness. In the same vein, we further compared the MRCL's IPE with the individual rating predictions of the Multiple Models SRL. These comparisons aim to demonstrate that the proposed approach is able to model both the commonalities and distinctions among raters simultaneously as opposed to the Multiple and Single Models SRL which are limited in that they only model individual differences or a dominant consensus respectively.

6 Results

In this section, we first describe the selection of statistical tests used in evaluating the results (see Section 6.1) and the agreement between raters (see Section 6.2). Afterwards, we report the classification results in terms of IPE (see Section 6.3) and GPE (see Section 6.4) for the two sparsely-crossed annotation datasets (EmoPain and Simulated-Sparse Laughter Labelling). Table 4 shows the IPE and GPE for both EmoPain and Simulated-Sparse Laughter Labelling datasets in terms of F1 score. Comparisons of the MRCL IPE and MRCL GPE were reported with respect to the Multiple Models SRL and Single Model SRL based on RF, Boosting, and Logistic. Finally, we discuss the potential of this approach for creating a unique space to consider the relevance of features across the individual raters and with respect to the global rating (see Section 6.5).

6.1 Statistical Tests

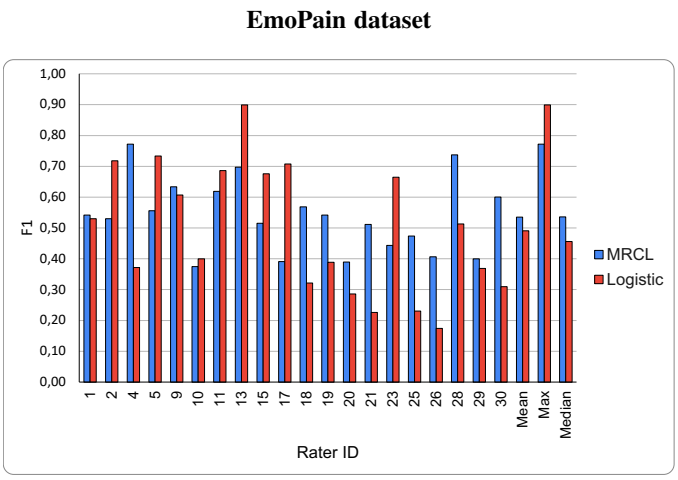
We used statistical tests to rigorously compare the performances of our MRCL approach with the SRL methods. For the EmoPain dataset, we compared F1 scores over: (i) LOSO-CV folds (for GPE), and (ii) raters (for IPE). Accordingly, for the Simulated-Sparse Laughter Labelling dataset, we compared F1 scores over 20 repetitions (i.e. different LOSO-CV experiments with raters randomly selected from the original Laughter dataset) averaged across: (i) LOSO-CV folds for GPE, and (i) raters for IPE.

The F1 scores for the IPE of the MRCL on the EmoPain dataset were found to follow a normal distribution according to the Anderson-Darling test (see Table 5). Similarly, the F1 scores for the IPE and GPE of the MRCL on the Laughter dataset were also found to follow a normal distribution. There was deviation from normality for the GPE of the MRCL on the EmoPain dataset, and so we used the non-parametric, one-sided Wilcoxon signed-rank test ($\alpha = 0.05$) there.

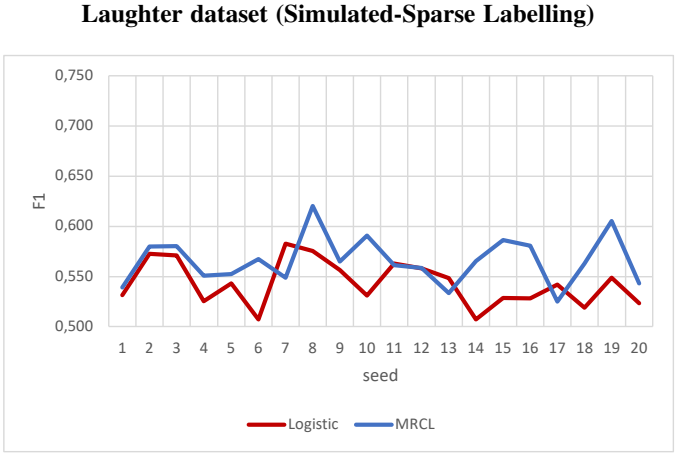
This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2023.3297270

TABLE 4: Sparsely-crossed annotation dataset evaluation: Individual performance evaluation (IPE) and Global Performance Evaluation (GPE) for both EmoPain and Simulated-Sparse Labeling datasets in terms of F1, and comparison with the state-of-the-art Multiple Models SRL and Single Model SRL: RF, Boosting, and Logistic. In bold with ↑ is the best performing algorithm for each column. Stars indicate whether the F1 of MRCL is significantly higher than state-of-the-art approaches ($** = p < .01, * = p < .05$).

	EmoPain		Laughter (Simulated-Sparse)					
	IPE	GPE	IPE			GPE		
Method			**No-laugh**	**Social**	**Hilarious**	**No-laugh**	**Social**	**Hilarious**
Multiple Models SRL								
RF	0.447∗	0.688	0.505∗∗	0.476	0.533 ∗∗	0.589	0.442	0.519∗∗
Boosting	0.416∗	0.636∗	0.485∗∗	0.498	0.490∗∗	0.585	0.470	0.569 ∗∗
Logistic	0.491	0.700	0.543∗∗	0.510	0.571∗∗	0.569∗∗	0.468	0.594∗∗
Single Model SRL								
RF	NA	0.677	NA	NA	NA	**0.691**↑	0.459	**0.639**↑
Boosting	NA	0.542∗	NA	NA	NA	0.672	**0.507**↑	0.601
Logistic	NA	0.689	NA	NA	NA	0.598	0.495	0.612
MRCL (Ours)	**0.540**↑	**0.778**↑	**0.566**↑	**0.525**↑	**0.594**↑	0.605	0.498	0.636



(a) IPE for EmoPain for each rater: MRCL vs Logistic (Multiple Models SRL)



(b) Averaged IPE for Laughter (Simulated-Sparse Labelling) for each seed (i.e. each simulation run): MRCL vs Logistic (Multiple Models SRL)

Fig. 5: Individual Performance Evaluation (IPE) for both EmoPain and Laughter (Simulated-Sparse Labelling).

TABLE 5: Normality Anderson-Darling test ($\alpha = 0.05$) of the F1 scores for MRCL.

	EmoPain	Laughter (Simulated-Sparse)		
		No-laugh	**Social**	**Hilarious**
IPE	$A = 0.320$, $p = 0.522$	$A = 0.196$, $p = 0.885$	$A = 0.508$, $p = 0.180$	$A = 0.348$, $p = 0.451$
GPE	$A = 1.495$, $p < 0.01$	$A = 0.468$, $p = 0.228$	$A = 0.288$, $p = 0.615$	$A = 0.251$, $p = 0.726$

6.2 Interrater Agreement

Interrater agreement of 0.72, 0.71, and 0.63 was found for the sit-to-stand, bend, and forward reach exercises in the EmoPain dataset based on intraclass correlation (ICC) [32]. For the Simulated-Sparse Laughter Labelling, we obtained an average interrater agreement ICC(C,1) [32] over the 20 random selections of 0.20. Although both far from perfect rater agreement (expected in naturalistic datasets), the EmoPain dataset has a higher level of agreement than the Simulated-Sparse Laughter Labelling dataset. The difference in levels of agreement is likely due to difference in annotation settings. For example, in the EmoPain dataset the raters were experts that share specific knowledge of movement behaviour and the observation stimuli were RGB videos, which provides some contextual information to the rater. The Laughter dataset, on the other hand, is based on naive observers, typical for non-clinical or non-expert applications. Further, the rating of the Laughter dataset was based on observation of minimalistic animation videos, i.e., stick figures animated using the original motion capture data from the dataset. Such stimuli is typically used for labelling when the focus of the study is to understand how people perceive affect through a specific modality, as well as when the aim is to ensure that both the human rater and the machine learning model observe and interpret the same stimulus.

6.3 Individual Performance Evaluation (IPE)

We found that for the EmoPain, the MRCL's IPE was greater than the IPE for the Multiple Models SRL. The difference was statistically significant for the RF ($t_{19} = 2.011$; $p < 0.05$) and Boosting ($t_{19} = 2.440$; $p < 0.05$) models. Although no statistically significant difference was found for the Logistic model ($t_{19} = 0.949$; $p = 0.177$), as can be further seen in Figure 5a, the MRCL outperforms it for 12 out of 20 raters (60 %).

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2023.3297270

For the Simulated-Sparse Laughter Labelling dataset, the MRCL's IPE was greater than the performance of the individual rating predictions of the Multiple Models SRL for the no-laugh/laugh (F1 score = 0.566) and hilarious/non-hilarious (F1 score = 0.594) tasks. In the case of the no-laugh/laugh ($t_{19} = 3.433; p < 0.01$) and hilarious/non-hilarious ($t_{19} = 4.231; p < 0.01$) tasks, the differences were statistically significant for the Logistic model. The difference was also significant for the RF ($t_{19} = 8.566; p < 0.01$) and Boosting ($t_{19} = 10.654; p < 0.01$) models in the hilarious/non-hilarious task. This was also true for the RF ($t_{19} = 10.492; p < 0.01$) and Boosting ($t_{19} = 10.705; p < 0.01$) models in the no-laugh/laugh task. As can be further seen in Figure 5b, the MRCL outperforms the Logistic in the no-laugh/laugh task for 14 out of 20 simulations (70%). The MRCL was not better than chance level performance (F1 score = 0.525) for the social/non-social task. The difficulty of solving social/non-social task classification without any contextual information about the performed activity is also highlighted in the original data collection study for this dataset [29].

EmoPain dataset

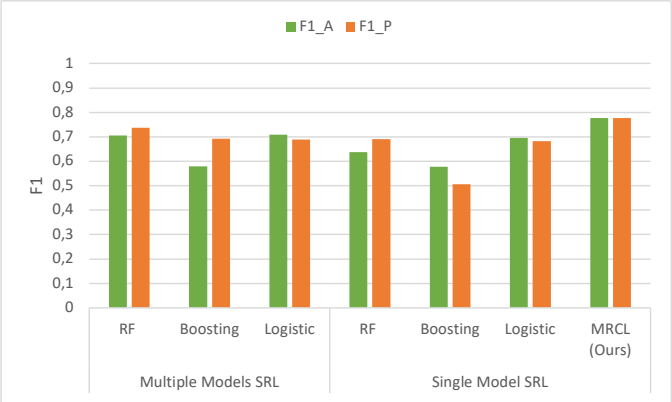


Fig. 6: Global Prediction (GPE) F1 score computed for each class of the EmoPain labels, i.e. the presence (F1_P) and absence (F1_A) of guarding behaviour respectively, showing comparison of the MRCL global prediction to: (i) Multiple Models SRL and (ii) Single Model SRL for the baseline algorithms RF, Boosting, and Logistic.

6.4 Global Performance Evaluation (GPE)

For the EmoPain dataset, with F1 score = 0.778, the MRCL's GPE outperforms the global rating prediction of the Multiple Models SRL. The difference is significant ($W = 40.5; Z = 2.075; p < 0.05$) for the Boosting model, but not for the RF ($W = 18; Z = 1.468; p = 0.071$) or Logistic ($W = 21.5; Z = 1.185; p = 0.118$). Figure 6 shows that this higher performance of the MRCL holds for both guarding behaviour classes (presence (F1_P) and absence (F1_A)). As with the Multiple Models SRL, the Single Model SRL performs worse than the MRCL. The difference is statistically significant for Boosting ($W = 66.5; Z = 2.120; p < 0.05$) but not for RF ($W = 55; Z = 1.217; p = 0.118$) or Logistic ($W = 43; Z = 1.537; p = 0.062$). This performance of the MRCL also holds for both guarding behaviour classes (see Figure 6).

For the Simulated-Sparse Laughter Labelling dataset, with F1 score = 0.605 and F1 score = 0.636 the MRCL approach outperforms the global rating prediction of the Multiple Models SRL in the no-laugh/laugh and hilarious/non-hilarious tasks respectively. This difference is statistically significant for the Logistic model, in both no-laugh/laugh ($t_{19} = 3.814; p < 0.01$) and hilarious/non-hilarious ($t_{19} = 2.969; p < 0.01$) tasks. Similarly, the difference is statistically significant for the RF ($t_{19} = 7.843; p < 0.01$) and Boosting ($t_{19} = 3.545; p < 0.01$) models in the hilarious/non-hilarious task. The difference was not statistically significant for the RF ($t_{19} = 1.011; p = 0.162$) or Boosting ($t_{19} = 1.119; p = 0.139$) models in discriminating between no-laugh and laugh. Neither the MRCL (F1 score = 0.498) nor the Multiple Models SRL were better than chance level performance for detecting social laughter. Although no statistical significant differences were found, the performance of the MCRL was higher than that of the Single Model SRL in distinguishing between no-laugh and laughter ($t_{19} = 0.713; p = 0.242$) and between hilarious and non-hilarious ($t_{19} = 1.488; p = 0.077$) based on the Logistic algorithm. The performance of the MRCL was not higher than the performance of the Single Model SRL in either the no-laugh/laugh or hilarious/non-hilarious tasks based on RF. However, in the no-laugh/laugh task, the MRCL achieved a 22% and 16% increase in terms of true positive rate compared to RF and Boosting.

6.5 Feature Space Analysis

In order to further understand the MRCL approach, we analyzed its behaviour with respect to the feature space. We computed the importance of each input feature by averaging the magnitude of the MRCL model coefficients of each LOSO-CV fold for each individual rating prediction. Typical modelling approaches enforce exploration by congregate level where feature relevance has been learned for a ground-truth aggregate (such as Figure 7a and Figure 7b for the average rater in the EmoPain and Simulated-Sparse Laughter dataset respectively) or at either individual rater level (such as Figure 7c and Figure 7d for independent models per EmoPain and Simulated-Sparse Laughter rater respectively). Such analyses shed light on the importance of features for each of these levels independently, and so the feature representations obtained are limited for reasoning about the consensus in feature relevance both among the raters and between the raters and the average. We can see in Figures 7a and 7b (for the EmoPain and Simulated-Sparse Laughter Labelling datasets respectively) that feature relevance when only the average ground truth is learnt does not allow one to see how the importance of each feature differs between raters. In turn, independent modelling of individual ratings alone, which does not account for commonalities between raters, cannot provide insight into differences and similarities in feature importance that might explain the (dis)agreement between raters. Figures (7c and 7d) show the feature relevance for such modelling based on the EmoPain and Simulated-Sparse Laughter Labelling datasets.

Instead, with feature relevance for the MRCL we can, for example, make more direct inferences at all levels as each rater is individually represented, all raters share the same feature relevance space, and relevance characterizes both individual ratings and the average rating simultaneously. We can notice in Figure 7e (for the EmoPain dataset) that the optimization of individual ratings together with their average has led the MRCL to prioritize the same set of features across raters. This suggests very high agreement in terms of the features used by the raters, which diffuses to relevance for the average rating. For the Simulated-Sparse Laughter Labelling dataset, as can be seen in 7f there is

disagreement in the relevance of features across raters and minimal consensus. While low feature weighting commonly across raters may simply reflect feature redundancy, the shared feature relevance space in the MRCL (shared between raters as well as between individual and average ratings) enables comparison. In addition, the shared space explicitly enhances consensus in feature relevance across raters.

7 Discussion

We set out to explore a new approach (MRCL) for simultaneously modelling aggregate ratings across multiple raters and also representing individual ratings for sparsely-crossed annotation datasets. We have compared our MRCL learning algorithm with state-of-the-art algorithms that learn only aggregate ratings (or one individual rating at a time) on two very different datasets: different with respect to the type and context of the rated behaviours, the type of stimulus labelled, the labels themselves, and the expertise of the raters. Most importantly, the datasets were different in levels of agreement between raters. In this section, we discuss the implication of our findings for interrater variability modelling, together with a high-level analysis of the MRCL.

We found higher performance for the MRCL in predicting individual ratings (IPE) for both the EmoPain and Simulated-Sparse Laughter Labelling datasets compared with the state-of-the-art models. This suggests that as hypothesized in Section 3, the raters are indeed different related 'tasks', i.e. there are commonalities among raters although there are also distinctions between them. The MRCL's use of both individual judgements as well as the dominant consensus (majority vote) enables both commonalities and distinctions to be used in modelling the given construct (e.g. guarding behaviour in the case of the EmoPain dataset). Studies on the sources of interrater variability in fact highlight the importance of strategically integrating multiple ratings, moving away from the view that variability is only due to rater errors or unwanted biases [33]. Findings in [33], [34] point to four possible categories of variability in labelling that affective computing researchers may find valuable to incorporate in their recognition models, rather than simply discarding them as aggregation methods [7], [10], [11] do. We discuss below how a learning algorithm that takes the approach of the MRCL addresses each source of variability.

7.1 Representing Multiple Truths

Variability may be the result of the existence of more than one point of view [34]. This is pertinent to affect where there may be more than one 'truth' [35]. Aggregation approaches [8], [9], [12], [13] give voice to the dominant version of truth alone, without accounting for minority groups within the ratings. As findings in [34] show, differences in points of view are not merely idiosyncratic, but rather there is evidence that there may be clusters of opinions where multiple raters belong to each cluster. This is particularly true when the raters have limited access to the data collection context (e.g., knowledge of the situation, of the person). Although the MRCL does not explicitly model clusters in ratings, in its approach multiple points of view are represented and learned. Further, our findings of better IPE performance of the MRCL show that the MRCL approach has particular advantage for individual raters represented by only a small set (about 10%) of data instances. This is likely because the MRCL brings together data instances for the different raters thereby increasing the effective sample size for each rater culminating in improved individual rating predictions. The performance gain of the MRCL with respect to state-of-the-art methods is also correlated with interrater agreement. Simultaneous learning of multiple correlated tasks can effectively improve generalization performance. The first evidence of that was found in the high MRCL-GPE for the EmoPain dataset that mirrors the higher levels of interrater agreement, higher than for the Simulated-Sparse Laughter Labelling dataset (ICC=0.63 and 0.20 respectively). Moreover, this outcome is also confirmed by the high correlation that we found between the ICC and MRCL'IPE (Pearson correlation = 0.432) and between the ICC and MRCL-GPE (Pearson correlation = 0.363) for the 20 simulated sparsely-crossed annotation settings for the Laughter dataset (no-laugh/laugh task).

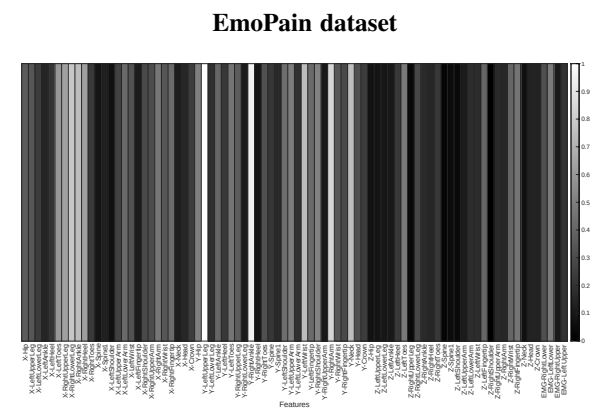
Beyond the MRCL method, there is also multilabel learning [15], [29], [36] which preserves multiple truths. However, this requires fully-crossed annotation while our MRCL method works for sparsely-crossed designs. Another related approach is uncertainty learning. For example, in [37], the agreement between raters is predicted and used to regularize the prediction of the consensus. While such methods model interrater variability, they do not actually represent the multiple versions of truth, which the MRCL does. The model of [38] like the MRCL optimizes both individual ratings and the consensus simultaneously. However, in their approach, for data instances with lower agreement between the ratings, higher priority is given to the individual rating prediction. Meanwhile, for data instances with higher agreement between the ratings, higher priority is given to the consensus prediction. While interesting, the approach penalizes individual ratings more when they deviate from high consensus. This aligns with the view of noise and bias being the primary (and unwanted) source of variability in multiple ratings. The MRCL differently does not bias the characterization of any rater beyond regularization of the individual rating predictions using the consensus loss. The method of [38] further assumes same sets of raters across instances.

7.2 Capturing Differences in Focus of Attention & Realising A Consensus

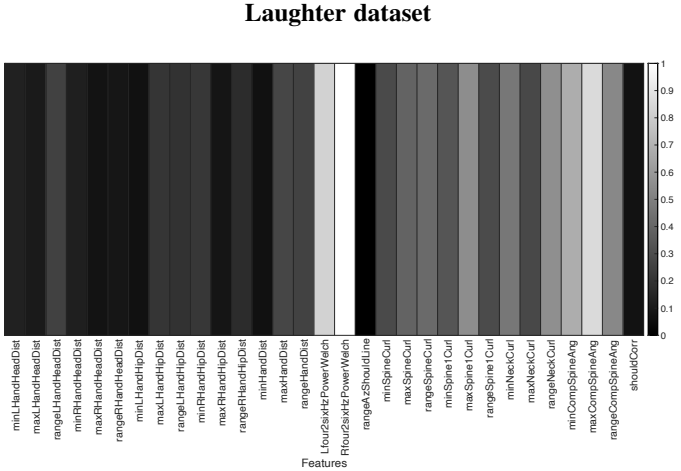
Differences in rater attention focus - Multiple points of view across raters seems to, at least partially, be a result of different foci or different weights placed on the multiple aspects of the observed behaviour [33]. This is true not only for naive observers but to a certain extent also for experts as in the case of chronic pain management teams (physiotherapists vs psychologists), or physiotherapists with different background (e.g., biomechanics vs neurology) [33]. The MRCL addresses this in its approach in which individual ratings are not merely represented but rather each rater is even characterised in terms of common patterns with others. Indeed, in the EmoPain a space of common features is identified based on the MRCL, whereas with the Multiple Models SRL, only individual differences are brought up. Findings of strong feature relevance similarities for the EmoPain dataset and weaker similarities for the Simulated-Sparse Laughter Labelling dataset reflect their levels of interrater agreement (ICC=0.63 and 0.20 respectively) suggesting that the MRCL itself finds different foci of attention to explain rating differences.

Realising a label by integrating multiple inferences - Raters seem to make multidimensional inferences that are then integrated to form their overall impression, i.e. the rating provided [33]. This rating may itself be formed in parallel with the 'multidimensional

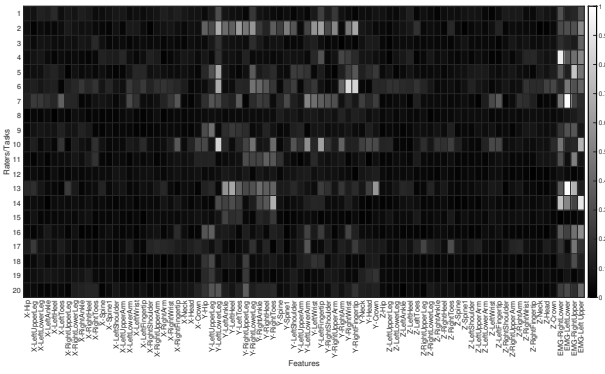
This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2023.3297270



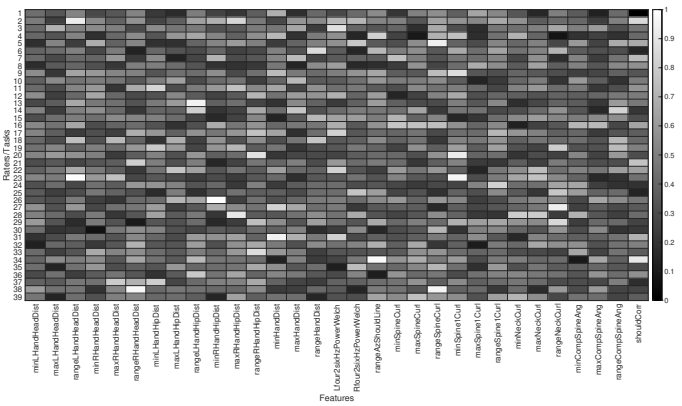
(a) Single Model SRL



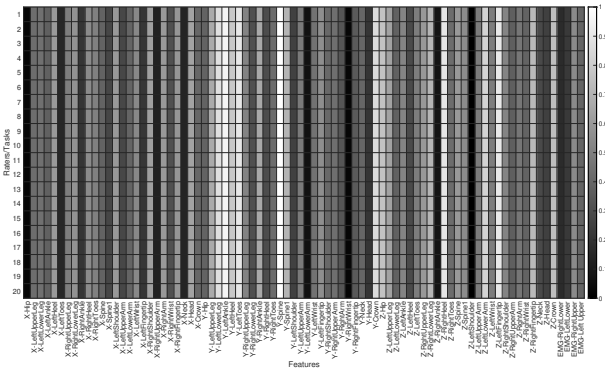
(b) Single Model SRL



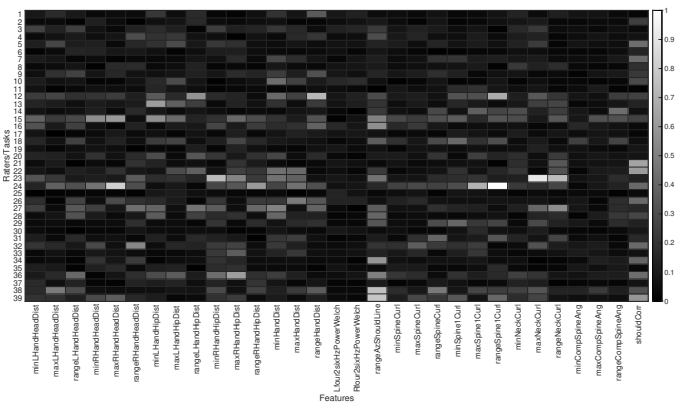
(c) Logistic (Multiple Models SRL)



(d) Logistic (Multiple Models SRL)



(e) MRCL (Ours) model



(f) MRCL (Ours) model

Fig. 7: Visualization of guarding and no-laugh behaviour localization in the feature dimension and for each rater for the Single Model SRL, the Logistic Multiple Models SRL, and MRCL (ours). The feature importance was extracted by averaging the magnitude weights coefficients of each LOSO-CV and LOLO-CV folds for EmoPain and Simulated-Sparse Laughter datasets respectively.

inferences' (low-level judgements). The process of integration may be an additional source of variability between raters. While the MRCL does not address this, it mimics this process of judgement making by learning multiple low-level inferences (individual ratings from multiple raters) in parallel with a single high-level score (qualitative or quantitative) that summarizes them. This highlights the naturalness of the MRCL approach. We found that this approach led to higher performance in predicting the consensus (GPE) compared with standard algorithms, although this was only true for the EmoPain and not the Simulated-Sparse Laughter Labelling dataset. This may be a result of the lower interrater agreement in the Simulated-Sparse Laughter Labelling dataset, which could have caused the MRCL higher confusion in deducing the consensus. Further, assessment noise (an unwanted source of variability) may be of significance here due to the type of visual stimuli (stick figure animations) used for observation in the Laughter dataset.

Dealing with Ambiguity - Indeed, a fourth source of interrater variability is ambiguity, e.g., in the assessment criteria [33]. Findings in emotion studies on interactions between language (or emotion labels) and emotion perception (see [39] for a review) highlight the pertinence of ambiguity in observer annotation of affective behaviour. It is also known that observer and subject characteristics, e.g. culture and context, influence emotion perception [40]. The MRCL does not address this variability source, but the findings discussed above do emphasize advantage in preserving part of the variability in observer ratings as opposed to discarding it completely, e.g. with majority voting. Our findings suggest that when the prediction of the consensus is of more critical importance than prediction of the individual ratings themselves, the MRCL is best suited to ratings with medium to high interrater agreement. Nevertheless, when prediction of individual ratings is at least as important as prediction of the consensus, the MRCL outperforms SRL methods regardless of the agreement between the ratings.

7.3 Limitations and Future Work

While the results shed light on the value of our approach for dealing with sparsely-crossed annotation designs, the optimum level of sparsity that enables the MRCL achieve competitive performance need to be further investigated. Future directions may include active learning methodologies that inform the sampling strategy to set the optimal level of sparsity for annotation. This could provided the proposed approach with the optimal annotation subset. Finally, the proposed approach could be generalized to other datasets and applications where multiple affect labels are available, in a sparse setting with different levels of interrater agreement. It will be valuable for the research community to contribute naturalistic datasets that reflect real life applications and enable such development of our approach. Naturalistic datasets addressing the complex nature of real life labelling are currently lacking [41].

One significant limitation of the current study is that we do not model the temporal information within the affective expressions. Future work could address this with a spatio-temporal MRCL that models non-linear temporal relationships of the input features while promoting consensus among raters. This limitation could be also addressed by representing the input features that belong to the same subject as a bag of temporal instances and encouraging an ordinal structure of the instance within each bag (i.e. modeling the temporal evolution of the trajectory) [42]).

8 Conclusions

Our work proposes a MRCL for modelling affect in sparsely-crossed annotation designs. We tested our approach across two naturalistic datasets with a high total number of raters and with different levels of agreement between raters. The significant improvement obtained with the MRCL suggests that the proposed approach is a viable solution for increasing performance in the context of sparse ratings. Although the increase in performance is more evident at global rating level when the agreement among raters is high, at the level of individual ratings the MRCL nevertheless outperforms SRL methods regardless of the agreement between the ratings. Findings of feature space analysis suggests how MRCL itself finds different foci of attention to explain rating differences.

References

[1] 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, 2012.

[2] J. Zhang, X. Wu, and V. S. Sheng, "Learning from crowdsourced labeled data: a survey," *Artificial Intelligence Review*, vol. 46, no. 4, pp. 543–576, 2016.

[3] L. Wang, Y. Li, J. Zhou, D. Zhu, and J. Ye, "Multi-task survival analysis," in *2017 IEEE International Conference on Data Mining (ICDM)*, 2017, pp. 485–494.

[4] H. Meng, A. Kleinsmith, and N. Bianchi-Berthouze, "Multi-score learning for affect recognition: the case of body postures," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 225–234.

[5] V. S. Sheng, "Simple multiple noisy label utilization strategies," in *2011 IEEE 11th International Conference on Data Mining*, 2011, pp. 635–644.

[6] N. B. Shah, S. Balakrishnan, and M. J. Wainwright, "A permutation-based model for crowd labeling: Optimal estimation and robustness," *IEEE Transactions on Information Theory*, vol. 67, no. 6, pp. 4162–4184, 2020.

[7] N. Dalvi, A. Dasgupta, R. Kumar, and V. Rastogi, "Aggregating crowdsourced binary ratings," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 285–294.

[8] Y. Yan, R. Rosales, G. Fung, and J. G. Dy, "Active learning from crowds," in *ICML*, 2011.

[9] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds." *Journal of Machine Learning Research*, vol. 11, no. 4, 2010.

[10] X. Liu, L. Li, and N. Memon, "A lightweight combinatorial approach for inferring the ground truth from multiple annotators," in *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, 2013, pp. 616–628.

[11] D. Zhou, Q. Liu, J. Platt, and C. Meek, "Aggregating ordinal labels from crowds by minimax conditional entropy," in *International conference on machine learning*. PMLR, 2014, pp. 262–270.

[12] J. Whitehill, T.-f. Wu, J. Bergsma, J. Movellan, and P. Ruvolo, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," *Advances in neural information processing systems*, vol. 22, pp. 2035–2043, 2009.

[13] P. Welinder, S. Branson, P. Perona, and S. Belongie, "The multidimensional wisdom of crowds," *Advances in neural information processing systems*, vol. 23, pp. 2424–2432, 2010.

[14] H. Kajino, Y. Tsuboi, and H. Kashima, "A convex formulation for learning from crowds," in *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[15] A. T. Nguyen, B. C. Wallace, J. J. Li, A. Nenkova, and M. Lease, "Aggregating and predicting sequence labels from crowd annotations," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2017. NIH Public Access, 2017, p. 299.

[16] D. Lopez-Martinez, O. Rudovic, and R. Picard, "Physiological and behavioral profiling for nociceptive pain estimation using personalized multitask learning," *arXiv preprint arXiv:1711.04036*, 2017.

[17] 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.

[18] B. Romera-Paredes, A. Argyriou, N. Berthouze, and M. Pontil, "Exploiting unrelated tasks in multi-task learning," in *International conference on artificial intelligence and statistics*, 2012, pp. 951–959.

[19] G. Pons and D. Masip, "Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition," *arXiv preprint arXiv:1802.06664*, 2018.

[20] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 19–26. [Online]. Available: https://doi.org/10.1145/3133944.3133949

[21] S. Chen and Q. Jin, "Multi-modal dimensional emotion recognition using recurrent neural networks," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 49–56. [Online]. Available: https://doi.org/10.1145/2808196.2811638

[22] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2d continuous space," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 3–14, 2017.

[23] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters*, vol. 66, pp. 22–30, 2015, pattern Recognition in Human Computer Interaction. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865514003572

[24] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 109–117.

[25] G. Denevi, C. Ciliberto, D. Stamos, and M. Pontil, "Learning to learn around a common mean," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[26] M. S. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng, A. Kemp, M. Shafizadeh *et al.*, "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.

[27] F. J. Keefe and A. R. Block, "Development of an observation method for assessing pain behavior in chronic low back pain patients." *Behavior therapy*, 1982.

[28] 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*, 2018.

[29] 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.

[30] G. McKeown, W. Curran, D. Kane, R. Mccahon, H. J. Griffin, C. McLoughlin, and N. Bianchi-Berthouze, "Human perception of laughter from context-free whole body motion dynamic stimuli," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 306–311.

[31] C. Wang, M. Peng, T. A. Olugbade, N. D. Lane, A. C. D. C. Williams, and N. Bianchi-Berthouze, "Learning bodily and temporal attention in protective movement behavior detection," *arXiv preprint arXiv:1904.10824*, 2019.

[32] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients." *Psychological methods*, vol. 1, no. 1, p. 30, 1996.

[33] P. Yeates, P. O'Neill, K. Mann, and K. Eva, "Seeing the same thing differently," *Advances in Health Sciences Education*, vol. 18, no. 3, pp. 325–341, 2013.

[34] A. Gingerich, C. P. van der Vleuten, K. W. Eva, and G. Regehr, "More consensus than idiosyncrasy: Categorizing social judgments to examine variability in mini-cex ratings," *Academic Medicine*, vol. 89, no. 11, pp. 1510–1519, 2014.

[35] F. Cavicchio, S. Dachkovsky, L. Leemor, S. Shamay-Tsoory, and W. Sandler, "Compositionality in the language of emotion," *PloS one*, vol. 13, no. 8, p. e0201970, 2018.

[36] M. H. Jensen, D. R. Jørgensen, R. Jalaboi, M. E. Hansen, and M. A. Olsen, "Improving uncertainty estimation in convolutional neural networks using inter-rater agreement," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 540–548.

[37] C. Wang, Y. Gao, C. Fan, J. Hu, T. L. Lam, N. D. Lane, and N. Bianchi-Berthouze, "Agreementlearning: An end-to-end framework for learning with multiple annotators without groundtruth," *arXiv preprint arXiv:2109.03596*, 2021.

[38] C. H. Sudre, B. G. Anson, S. Ingala, C. D. Lane, D. Jimenez, L. Haider, T. Varsavsky, R. Tanno, L. Smith, S. Ourselin *et al.*, "Let's agree to disagree: Learning highly debatable multirater labelling," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 665–673.

[39] K. A. Lindquist and M. Gendron, "What's in a word? language constructs emotion perception," *Emotion Review*, vol. 5, no. 1, pp. 66–71, 2013.

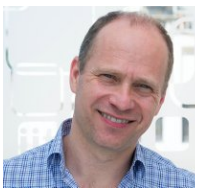
[40] H. A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: a meta-analysis." *Psychological bulletin*, vol. 128, no. 2, p. 203, 2002.

[41] T. Olugbade, M. Bieńkiewicz, G. Barbareschi, V. D'Amato, L. Oneto, A. Camurri, C. Holloway, M. Björkman, P. Keller, M. Clayton, A. C. d. C. Williams, N. Gold, C. Becchio, B. Bardy, and N. Bianchi-Berthouze, "Human movement datasets: An interdisciplinary scoping review," *ACM Computing Surveys*, 2022.

[42] D. Dennis, C. Pabbaraju, H. V. Simhadri, and P. Jain, "Multiple instance learning for efficient sequential data classification on resource-constrained devices," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 953–10 964.

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 Tenure Track Assistant Professor on Computer Science with the Department Economics and Law, University of Macerata, Macerata, Italy. He is also affiliated with the Unit of 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.

Temitayo Olugbade is a Lecturer in Computer Science and AI at University of Sussex and Honorary Fellow at University College London. Her research pursues development and application of state-of-the-art machine learning methods to new and challenging affective computing 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.

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 300 papers in affective computing, human-computer interaction and pattern recognition. She has investigated affect-aware technology in real-life contexts: e.g. Physical rehabilitation: EPSRC Emo&Pain, H2020 EnTimeMent; Sustainable fashion industry: EPSRCTextile Circulary Centre, EPSRC CX & Dematerialization Tools; Physiological sensing in real life context: EPSRC Digital Sensoria, H2020 HU-MAN Manufacturing, Bentley-funded comfort level detection.