Science
Robotics

robotics.sciencemag.org/cgi/content/full/3/19/eaao6760/DC1

Supplementary Materials for

Personalized machine learning for robot perception of affect and engagement in autism therapy

Ognjen Rudovic*, Jaeryoung Lee, Miles Dai, Björn Schuller, Rosalind W. Picard

*Corresponding author. Email: orudovic@mit.edu

Published 27 June 2018, *Sci. Robot.* **3**, eaao6760 (2018)
DOI: 10.1126/scirobotics.aao6760

This PDF file includes:

Note S1. Details on model training and alternative approaches.

In table S1, we compare different methods described in the Alternative approaches section of the paper and further detailed below. In fig. S1, we depict the error distributions of the top performing methods, highlighting the regions in the error space, where the proposed PPA-net was most effective (and otherwise). Fig. S2 shows the convergence rates of the evaluated deep models, in terms of the learning steps and the loss minimization (MSE). Note that the proposed PPA-net was able to fit the target children significantly better, while still outperforming the compared methods on the previously unseen data of those children (table S1). In traditional ML, where the goal is to be able to generalize to previously unseen subjects, this could be considered as algorithmic bias (the model overfitting). By contrast, in personalized ML, as proposed here, it is beneficial as it allows the model to perform best on unseen data of the target subject on whom we aim to personalize the model. Fig. S3 depicts the contribution of each modality to the estimation performance (Effects of different modalities). The bars in the graph show ICC (average±SD) for each modality obtained across the children. The PPA-net configuration used to make predictions from each modality was the same as that used in the multi-modal scenario. However, the size of the auto-encoded space varied in order to accomodate the size of the input features. Specifically, the optimal size of the encoded features per modality was: 150 (face), 50 (body), and original feature size: 24 for the audio, and 30 for the physiology modality, were used. In what follows, we provide additional details on the training procedures for the alternative methods used in our experiments.

- **MLP:** We used the standard multi-layer perceptron to train the child-dependent MLP (CD-MLP) deep network. We used the same architecture/layer types as in our GPA-net; however, the training of its layers was done in traditional manner ('at once') using the Adadelta algorithm. The number of epochs was set to 200 with early stopping on the validation set. The personalization of this network (P-MLP) was accomplished by cloning the last layer in the network (inference layer) and fine-tuning it to each child using SGD ($lr = 0.03$). We also included a leave-one-child-out experiment in which we trained the MLP deep model using the data of all children but the target child (CI-MLP). This was repeated for all the children. We used this model as it acts as the baseline for our personalized approach. We evaluated this model in the same way as the reported methods (i.e., on the 40% of the left-out data of the target child).

- **LR:** Lasso Regression (LR) (*5*) is the standard linear regression with L-1 regularization on the design matrix. The regularization parameters were set on the validation data used in our experiments to obtain the best performance.

- **SVR:** Support Vector Regression (SVR) (*5*) is the standard kernel-based method used for non-linear regression. It defines a kernel matrix computed by applying a pre-defined kernel function to data pairs. We used the standard isotropic Radial Basis Function (RBF). Due to the non-parametric nature of this method, it was computationally infeasible to use all training data to form the kernel matrix. To circumvent this, we trained one SVR per child (using all training data). Then, we selected support vectors (SV) – the most discriminative examples – from each child (SVs=1000) and re-trained the model using these SVs (35k data points in total). To prevent the model from overfitting, the penalty parameter C was selected on the validation data.

- GBRT: Gradient Boosted Regression Trees (GBRT) (*39*) is a generalization of boosting to arbitrary differentiable loss functions. We set the number of basis functions (weak learners) to 35, corresponding to the number of children. The trees' depths were varied from d=3-10, and we found that d=5 performed the best on the validation set. This configuration was used to produce the results reported.

For these methods, we used the publicly available implementations. Specifically, for MLP, we used the Keras API (*51*), and for the rest, we used *sklearn* (*53*), a Python toolbox for ML.

Note S2. Data set.

We presented the results obtained on the multi-modal dataset of children with autism (MDCA) (*27*), containing recordings of the children undergoing occupational therapy for autism. The therapy was led by experienced child therapists and assisted by a humanoid robot NAO. The goal of the therapy was to teach the children to recognize and imitate emotive behaviors (using the Theory of Mind concept (*54*)) as expressed by the NAO robot. During the therapy, the robot was operated by the therapist, but the data were collected to enable future autonomous perception of the affective states of a child learner by the robot. The data include: (i) video recordings of facial expressions, head, and body movements, pose, and gestures, (ii) autonomic physiology (heart rate (HR), electrodermal activity (EDA), and body temperature (T)) from the children, as measured on their non-dominant wrist, and (iii) audio-recordings (Fig. 1). The data come from 35 children with different cultural backgrounds. Namely, 17 children (15 males / 2 female) are from Japan (C1), and 19 children (15 males / 4 females) are from Serbia (C2) (*27*). Note that in this paper we excluded the data of one male child from C2 due to the low-quality recording. Each child participated in a 25-minute-long child-robot interaction; however, on average, only 60% of these data (containing the four key therapy steps: pairing, recognition, imitation and story-telling) were coded by human experts, and used to train/evaluate the models. The remaining data contained the introduction and therapy closing stages, considered less relevant for target analysis. Children's ages varied from 3-13, and all the children have a prior diagnosis of autism (see table S2). More details about the data, recording setting, and therapy stages can be found in (*27*).

Note S3. Feature processing.

The raw data of synchronized video, audio, and autonomic physiology recordings were processed using state-of-the-art, open-source tools. For analysis of facial behavior, we used the OpenFace toolkit (*29*). This toolkit is based on Conditional Local Neural Fields, a ML model for detection and tracking of 68 fiducial facial points, described as 2D coordinates (x, y) in face images (Fig. 1). It also provides 3D estimates of head-pose and eye-gaze direction (one for each eye), as well as the presence and intensity (on a 6 level Likert scale) of 18 facial action units (AUs) (*55*). The latter are usually referred to as the judgment level descriptors of facial activity in terms of activations of facial muscles. Most human facial expressions can be described as a combination of these AUs and their intensities, and they have been the focus of research on the automated analysis of facial expressions (*56*). For capturing body movements, we used the OpenPose toolkit (*30*) for automated detection of 18-keypoint body pose locations, 21-keypoint hand estimation, and 70 fiducial facial landmarks (all in 2D), along with their detection confidence (0-1). From this set, we used only the body pose and facial landmarks and disregarded the hand tracking (due to frequent occlusions of the children's hands). OpenPose is built on recent advances in convolutional neural networks (*6*), and the part affinity fields for part association (*30*).

For audio signal processing, we used the openSMILE toolkit (*31*) to automatically extract acoustic low-level descriptors (LLDs) from the speech waveform on the frame level. Specifically, we used 24 LLDs (pitch, MFCC, LSP, etc.) provided by openSMILE. These features were computed over sliding windows of length 100 ms with a 10 ms shift and then aligned with the visual features using time-stamps stored during the data recording.

To measure the biosignals based on autonomic physiology (HR, EDA and T), we used the commercially available E4 wrist-worn sensor (*25*). This wristband provides real-time readings of blood volume pulse (BVP) and HR (64Hz), EDA via the measurement of skin conductance (4Hz), skin temperature T (4Hz), and 3-axis accelerometer (ACC) data (32Hz). From these signals, we also extracted additional commonly used hand-crafted features (*25*), as listed in table S3. Since HR is obtained from BVP, we used only the raw BVP. Again, these were temporally aligned with the visual features using time-stamps stored during the data recording.

The multi-modal learning has been achieved by consolidating these features to act as predictors of target affective states and engagement in the PPA-net. From the OpenPose output, we used the face and body features with the detection confidence over each feature set (face and body) above 30%, which we found to be a good threshold by visually inspecting the detection results. The final feature set was formed as follows. (i) *Visual*: we used the facial landmarks from OpenPose, enhanced with the head-pose, eye-gaze, and AUs, as provided by OpenFace. (ii) *Body*: we merged the OpenPose body-pose features, and E4 ACC features encoding the hand movements. (iii) *Audio*: the original feature set was kept. (iv) *Physiology*: containing the features derived from the E4 sensor, without the ACC features.

We also included an auxiliary feature set provided by the expert knowledge. Namely, the children's behavioral severity at the time of the interaction (after the recordings) was scored on the CARS (*34*) by the therapists (table S2). The CARS form is typically completed in less than 30 minutes, and it asks about 15 areas of behavior defined by a unique rating system (0-4) developed to assist in identifying individuals with ASC. The rating values given for the 15 areas are summed to produce a total score for each child. CARS covers the three key behavioral dimensions pertinent to autism: social-emotional, cognitive, and sensory, and based on the total scores, the children fall into one of the following categories: (i) no autism (score below 30), (ii) mild-to-moderate autism (score: 30–36.5), and (iii) moderate-to-severe autism (37–60). We used this 15-D feature set (the CARS scores for each of the 15 areas) as a unique descriptor for each child - encoding the expert knowledge about the children's behavioral traits. Table S3 summarizes all the features we used to train/evaluate our models.

Note S4. Data coding.

The dataset was labeled by human experts along two commonly used affective dimensions (valence and arousal) and engagement. Each dimension was rated on a continuous scale from -1 to $+1$. Five expert therapists (two from C1 and three from C2) coded independently while watching the audio-visual recordings of target interactions. To measure coders' agreement, we used the intra-class correlation (ICC) score, type (3,1) (*28*). This score measures the proportion of the variance that is attributable to objects of measurement compared to the overall variance of the coders. The ICC is commonly used in behavioral sciences to assess the agreement of judges. Unlike the well-known Pearson Correlation, this ICC penalizes the scale differences and offset between the coders, which makes it a more robust measure of coders' agreement. The codings were aligned using standard alignment techniques: we applied time-shifting of ± 2 seconds to each coder and selected the shift which produced the highest average inter-coder agreement. The ground-truth labels used to evaluate the ML models were obtained by averaging the codings of the three coders who had the highest agreement (based on pair-wise ICC scores). We empirically found that this significantly reduced outlying codings. The obtained coding ("the gold standard") was then used as the ground truth for training ML models for estimation of valence, arousal, and engagement levels during the child-robot interactions. Note that in our previous work (*27*), we used discrete annotations for the three target dimensions. Since those were coded per manually selected engagement episodes, for this work we re-annotated the data to obtain more fine-grained (continuous) estimates of the affect and engagement from the full dataset. The description of the exemplary behavioral cues used during the coding process is given in table S4.

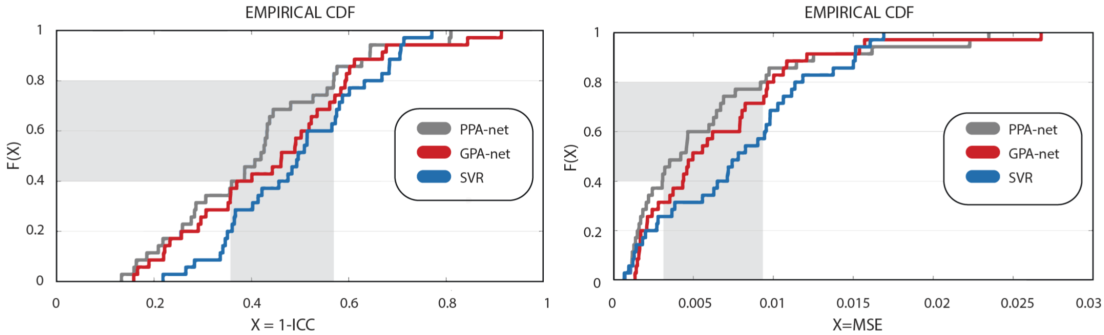

Figure S1. Empirical cumulative distribution function of ICC and MSE. The average scores were computed between the estimated valence, arousal, and engagement levels, and the gold-standard codings. We show the performance by three top ranked models (based on $TaskRank$ in table S1). The individual performance scores for 35 children were used to compute the CDFs in the plots. Note that the improvements due to the network personalization are most pronounced for $40\% < F(X) < 75\%$ of the children. On the other hand, the model personalization exhibits similar performance on the children for whom the group-level models perform very well ($0\% < F(X) < 40\%$), or largely underperform ($75\% < F(X) < 100\%$). This indicates that for the underperforming children, the individual expressions of affect and engagement vary largely across the children. Thus, more data from those children are needed to achieve a more effective model personalization.

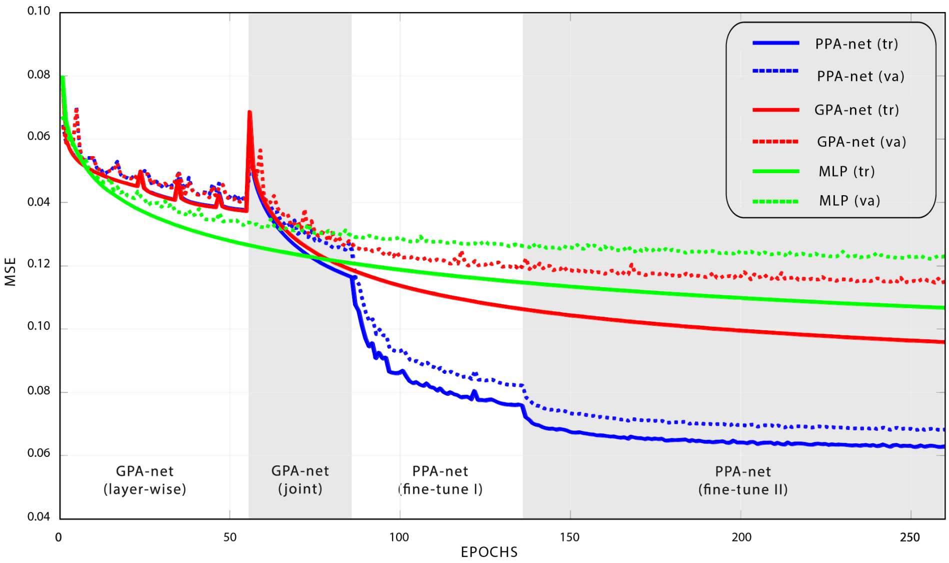

Fig. S2. The learning of the networks. MSE during each epoch in the network optimization is shown for the personalized (PPA-net and CD-MLP) and group-level (GPA-net) models, and for training (tr) and validation (va) data. The GPA-net learns faster and with a better local minimum compared to CD-MLP. This is due to the former using layer-wise supervised learning strategy. This is further enhanced by fine-tunning steps in PPA-net, achieving the lowest MSE during the model learning, which is due to its ability to adapt its parameters to each culture, gender and individual.

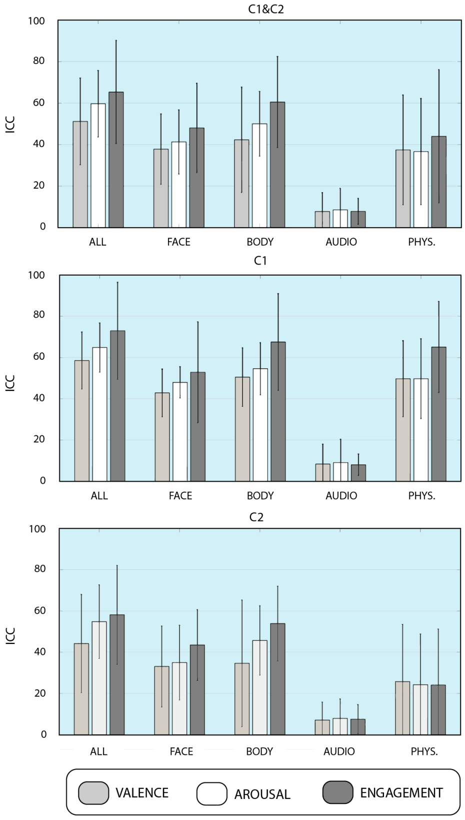

Fig. S3. PPA-net: The performance of the visual (face and body), audio, and physiology features. The fusion approach ('ALL') outperforms the individual modalities, evidencing the additive contribution of each modality to predicting the target outputs. The large error bars reflect the high level of heterogeneity in the individual performance of the network on each child, as expected for many children with ASC.

Table S1. Comparisons with alternative approaches. The average±SD of ICC (%) for estimation of the children's valence, arousal, and engagement levels. *TaskRank* quantifies the portion of tasks (35 children × 3 outputs = 105 total) on which the target model outperformed the compared models, including the standard MLP network with last layers personalized for each child (P-MLP), joint learning of a child-dependent MLP (CD-MLP), Lasso (linear) regression (LR), support-vector regression (SVR) with a Gaussian kernel, and gradient-boosted regression trees (GBRT). We also include the results obtained with the child-independent MLP (CI-MLP) - evaluated on previously unseen children.

Models	Valence	Arousal	Engagement	Average	*TaskRank (in %)*
PPA-net	**52±21**	**60±16**	**65±24**	**59±21**	**46.5 (1)**
GPA-net	47±28	57±15	60±25	54±24	10.2 (4)
P-MLP	47±18	54±15	59±22	53±20	3.29 (5)
CD-MLP	43±22	52±15	57±23	51±20	2.81 (6)
LR	28±22	35±19	37±21	34±21	2.52 (7)
SVR	45±26	56±14	49±22	50±21	21.2 (2)
GBRT	47±21	51±15	49±22	49±20	12.0 (3)
CI-MLP	16±17	23±16	21±23	20±19	1.45 (8)

Table S2. Summary of the child participants [taken from (*27*)]. The average CARS scores of the two groups are statistically different ($t(34) = -3.35, p = 0.002$). In our study, the data of one male child (C2) were excluded due to the low-quality of his recordings.

	C1 (Japan)	**C2 (Serbia)**
Age	7.59± 2.43	9.41± 2.46
Age range	3–12	5–13
Gender (male:female)	15:2	15:4
CARS	31.8± 7.1	40.3± 8.2

Table S3. Summary of the features.

Feature ID	Modality	Description
1-209	FACE (OpenPose)	Facial landmarks: 70x2 (x,y) and their confidence level (c)
210-215	FACE (OpenFace)	Head pose: 3D location and 3D rotation
216-223	FACE (OpenFace)	Eye gaze: 2x3 - 3D eye gaze direction vector in world coordinates for left and right eye + 2D eye gaze direction in radians in world coordinates for both eyes
224-240	FACE (OpenFace)	Binary detection of 18 AUs: AU01, AU02, AU04, AU05, AU06, AU07, AU09, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, AU28, AU45
241-258	FACE (OpenFace)	Intensity estimation (0-5) of 17 AUs: AU01, AU02, AU04, AU05, AU06, AU07, AU09, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, AU45
259-312	BODY (OpenPose)	Body pose: 18x3 - the pose keypoints containing the body part locations (x,y) and detection confidence (c)
313-327	BODY (E4 ACC)	Accelerometer data: 3D raw signal (x, y, z), z-normalized vector $\sqrt{x^2 + y^2 + z^2}$ (mean±SD within 5 sec window), mean, SD, max, min, 1st diff, abs value of 1st diff, abs value of normalized 1st diff, 2nd diff, abs value of 2nd diff, abs value of normalized 2nd diff, mean amplitude deviation (10 sec window)
328-338	PHYSIOLOGY (EDA)	Raw EDA and its: z-normalized value (30 sec window), mean, SD, max, min, integration, slope, number of peaks, amplitude, number of zero-crossings
339-347	PHYSIOLOGY (HR)	Raw HR signal and its: z-normalized value (4s window size), mean, 1st diff, abs value of 1st diff, abs value of the normalized 1st diff, 2nd diff, abs value of 2nd diff, abs value of normalized 2nd diff
348-357	PHYSIOLOGY (T)	Raw T signal and its: z-normalized value (4s window size), mean, 1st diff, abs value of 1st diff, abs value of the normalized 1st diff, 2nd diff, abs value of 2nd diff, abs value of normalized 2nd diff
358-381	AUDIO (openSMILE)	LLDs: RMS energy, Spectral flux, Spectral entropy, Spectral variance, Spectral skewness, Spectral kurtosis, Spectral slope, Harmonicity, MFCC 0, MFCC 1–10, MFCC 11–12, MFCC 13–14, Log Mel frequency band 0–7, LSP frequency 0–7, F0 (ACF based), F0 (SHS based), F0 envelope, Probability of voicing, Jitter, Shimmer, Logarithmic HNR, Sum of auditory spectrum (loudness), ZCR, Logarithmic HNR
382-396	CARS	15 ratings (0-4): relating to people, emotional response, imitation, body use, object use, adaptation to change, listening response, taste-smell-touch, visual response, fear or nervous, verbal communication, activity level, nonverbal communication, level and consistency of intellectual response, general impression

Table S4. The coding criteria. All three dimensions are coded on a continuous scale by human experts, using the discrete levels (the first column) as the reference points.

Level	Dimension	Description
negative (-1)	Valence	The child shows clear signs of experiencing unpleasant feelings (being unhappy, angry, visibly upset, showing dissatisfaction, frightened), dissatisfaction, and disappointment (e.g. when NAO showed an expression that the child did not anticipate)
neutral (0)	Valence	The child seems alert and/or attentive with no obvious signs of any emotion, pleasure, or displeasure
positive(+1)	Valence	The child shows signs of intense happiness (e.g. clapping hands), joy (in most cases followed with episodes of laughter), and delight (e.g. when NAO performed)
negative (-1)	Arousal	The child seems very bored or uninterested (e.g. looking away, not showing interest in the interaction, sleepy, passively observing)
neutral (0)	Arousal	The child shows no obvious signs of physical activity (face, head, hand, and/or bodily movements); seems calm, thinking, air-drawn
positive(+1)	Arousal	The child performs an intense and/or sudden physical activity followed by constant (sometimes rapid) movements like hand clapping, touching face/head/knees, actively playing with the robot, wiggling in the chair (C2) or on the floor (C1), jumping, walking around the room, being highly excited, shouting
negative (-1)	Engagement	The child is not responding to the therapist and/or NAO's prompts at all, walking away from NAO, looking for other objects in the room, and/or ignoring the robot and the therapist
neutral (0)	Engagement	The child seems indifferent to the interaction, is looking somewhere else, not paying full attention to the interaction; the therapist repeats the question and/or attempts the task a few times, until the child complies with the instructions
positive(+1)	Engagement	The child is paying full attention to the interaction, immediately responds to the questions of the therapist, reacting to NAO spontaneously and executing the tasks; the child seems very interested with minimal or no incentive from the therapist to participate in the interaction