2024 19th Annual System of Systems Engineering Conference (SoSE) | 979-8-3503-6591-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/SOSE62659.2024.10620945

Human Emotion Estimation through Physiological Data with Neural Networks

1st Jhair Gallardo
Department of Imaging Science
Rochester Institute of Technology
Rochester, United States
gg4099@rit.edu

2nd Celal Savur
Dept. Elect. & Microelectronic Eng.
Rochester Institute of Technology
Rochester, United States
cs1323@rit.edu

3rd Ferat Sahin
Dept. Elect. & Microelectronic Eng.
Rochester Institute of Technology
Rochester, United States
feseee@rit.edu

4th Christopher Kanan
Department of Computer Science
University of Rochester
Rochester, United States
ckanan@cs.rochester.edu

***Abstract*—Effective collaboration between humans and robots necessitates that the robotic partner can perceive, learn from, and respond to the human's psycho-physiological conditions. This involves understanding the emotional states of the human collaborator. To explore this, we collected subjective assessments - specifically, feelings of *surprise*, *anxiety*, *boredom*, *calmness*, and *comfort* — as well as physiological signals during a dynamic human-robot interaction experiment. The experiment manipulated the robot's behavior to observe these responses. We gathered data from this non-stationary setting and trained an artificial neural network model to predict human emotion from physiological data. We found that using several subjects' data to train a general model and then fine-tuning it on the subject of interest performs better than training a model only using the subject of interest data.**

***Index Terms*—Human-robot collaboration, physiological signals, machine learning**

I. Introduction

In a human-robot collaborative environment, it is important that the robot has access to information about the current emotional state of the human subject it is working with. A human experiencing anxiety may behave differently compared to one who is calm while collaborating with a robot. Having access to such information can help the robot adapt its behavior, thereby increasing its efficiency. This requires integrating robotics with physiological computing, an interdisciplinary field that focuses on inferring a person's psycho-physiological state. Human-computer interaction, brain-computer interaction, and affective computing are part of physiological computing [1]. It can enable robots to recognize, interpret, and dynamically change their behavior based on a person's psycho-physiological state.

According to NSF Research Statement for Cyber Human Systems (2018-2019), "*improve the intelligence of increasingly autonomous systems that require varying levels of supervisory control by the human; this includes a more symbiotic relationship between human and machine through the development of systems that can sense and learn the human's cognitive and physical states while possessing the ability to sense, learn, and adapt in their environments*" [2]. Thus, to interactively ensure trust and safety between the human and the robot, the robot should sense a human's cognitive and physical state, which will help build trust.

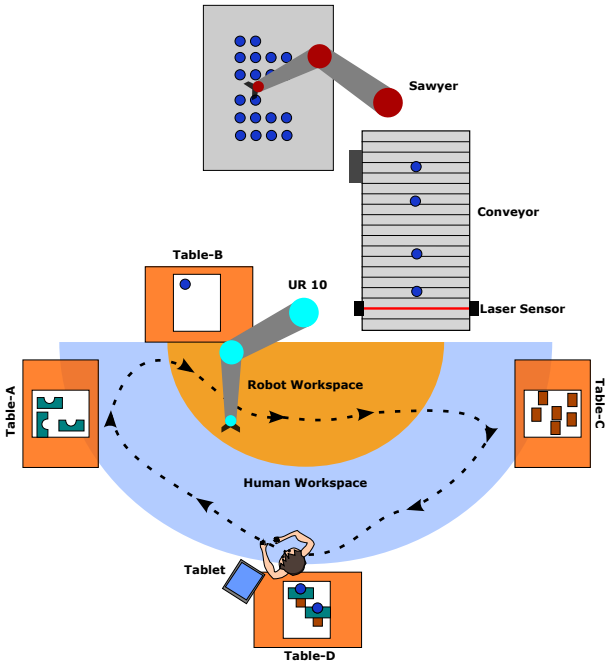


Fig. 1: Overview of HRC experiment. A participant collects parts from Table-A, Table-B, and Table-C, assembles them, and then drops it onto Table-D.

In a human-robot interaction setup, a change in a robot's motion can affect the human physiological state. Experiments such as [3] and [4] revealed that the robot's trajectory affects human skin conductivity. The literature review in [5] highlights using the 'psycho-physiological' method to evaluate human response and behavior during human-robot interaction.

In this research, we aim to evaluate different ways to train an artificial neural network model to predict the human emotional

979-8-3503-6591-7/24/$31.00 ©2024 IEEE 153

state and compare their performance. We compare models that are trained using several subjects, e.g., a general dataset, models that are trained only on the data of the subject of interest, and fine-tuned models that are trained on the general dataset and then fine-tuned on the subject of interest.

II. Related Work

We briefly review methods for estimating the psycho-physiological state from physiological signals, ranging from Electrocardiogram (ECG), Galvanic Skin Response (GSR), Electroencephalography (EEG), Electromyography (EMG), respiration rate (RSP), and Pupil dilation.

In [3], a method was presented to calculate the danger index by using distance and relative velocity between a human and a robot, and the inertia of the closest point to the human as suggested in [6]. Then the real-time calculated danger index is used to control the robot's trajectory on a real robot. Similarly, the same authors [7] tried to detect anxiety triggered by two trajectory planners. Biological signals and subjective responses were collected from subjects during the experiment. The result of the subjective responses from the experiment showed that the subjects felt less anxiety during a safe planner than the classical planner. Moreover, the researcher found that the corrugator EMG signal did not help to estimate arousal and valence. However, they have found a strong positive correlation between anxiety and speed, surprise and speed, and a negative correlation between calm and speed. In subsequent extensions of this work [8], [9], they showed that a Hidden Markov Model (HMM) outperforms Fuzzy inference for estimating arousal and valence from physiological signals.

In [10], ECG, GSR, RPS, Skin Temperature, EEG, and Eye tracking signals were used to estimate a human's workload and effort. During trials, subjects were asked to fill out the NASA-TLX questionnaire. In [11], physiological signals were used to detect ErrP from EEG signals, using both simulation and actual robots. Their result suggests that brain-computer interfaces could be used for continuous adaptation when no explicit information about the goal exists. In [12], robot actions were implicitly validated by using ErrPs derived from EEG signals, resulting in a classification accuracy of 69.0% for detecting an incorrect robot's actions. Similarly, [13] used EEG signals with error-related potential to fix the robot's mistake during the task. In their experiment, they used a Baxter robot to make decisions based on EEG signals in real time. In [14], they focused on hand-over tasks where a robot handed an item to a human. They collected physiological signals and showed that physiological responses vary between the robot's motions. In [15], ECG, EDA, and pupillometry signals, along with subjective responses, were used to estimate the human comfort index by using the circumplex model. They collected data from multiple experiments, and the proposed model was validated.

III. Metodology

This work aims to determine a human's emotional state from their physiological data while working with a robot in a collaborative environment. To do this, we define a human-robot collaborative setting where the human and the robot must complete a task together. On each trial, we collect physiological signals from the human and their corresponding feedback about their current emotions. Using this approach, we gathered a dataset with physiological signals and their corresponding emotion scores. Then, we train a model that uses the physiological signals as input and predicts the corresponding emotion scores of a subject.

TABLE I: List of Robot's behaviors

Robot Algorithm	Description
TriSSM-Vo	A DSS algorithm taking into account both the robot's velocity and the human's velocity.
TriSSM-Vo-with-UnCI	UnCI metric provided to the TriSSM-Vo algorithm to adjust cushioning distance and directed speed.
TriSSM-Vr	A DSS algorithm taking into account only the robot's velocity, not the human's velocity.
TriSSM-Vr-with-UnCI	UnCI metric provided to the TriSSM-Vr algorithm to adjust cushioning distance and directed speed

A. Dataset

The dataset used in this research was collected in a human-robot collaboration task where the participant did an assembly task and the robot provided a part for the assembly [16]. The data was collected from healthy college students (N=36). The participants consisted of 24 male and 12 female subjects (*Mean Age*= 28.28, *SD*= 7.93). The experiment consisted of 8 trials, and each trial consisted of multiple cycles. A cycle is defined as a sequence of small tasks where a participant collects parts from Table-A, Table-B, and Table-C, assembles them, and then drops them onto Table-D, as shown in Fig. 1. Specifically, the subject starts at Table-D, picks a component from Table-A, then moves to Table-B where they pick up the second component, and then crosses the robot workspace to obtain the last component from Table-C. The subject then assembles the three components and places the part on Table-D. Finally, the subject answers the questionnaire on the Tablet. The subject repeats this cycle for approximately six minutes. The trials were randomized to reduce the order effect. In some cases, the experiment took less than 6 minutes because the UR10 robot was completing all the parts. Hence, the trial is terminated either at the 6-minute limit or when no part is left for the robot to pick up. Subjective responses of *comfortability*, *safety*, *surprise*, *anxiety*, *calmness*, and *boredom* levels were collected after each part was delivered to *Table-D* during each trial and after each trial.

This experiment was designed to test the effect of the dynamic speed and separation monitoring (DSS) [17] algorithm and the Comfortability Index Estimation System (CIES) model [16] on human emotion. Table I lists the UR-10 robot behaviors. The TriSSM-Vo and TriSSM-Vr algorithms are detailed in [18].

A custom Android app was developed for this experiment to have better signal labeling. The app lets participants enter their

TABLE II: Physiological metrics extracted from the ECG, GSR, and Pupillometry signals

Type	Metric	Unit	Description
ECG	*Mean HR*	bpm/min	Mean of Heart rate
	Mean RR	ms	Mean of RR/IBI intervals
	SDNN	ms	Standard deviation of RR/IBI intervals
	RMSSD	ms	The root-mean-square of the difference of consecutive RR/IBI intervals
	pNN50	%	Percentage of successive RR/IBI intervals that differ by more than 50 ms
GSR	*Tonic Mean*	Micro-siemens	Mean of tonic component of GSR signal
	Tonic Std.	Micro-siemens	Standard deviation of tonic component of GSR
	Phasic Mean	Micro-siemens	Mean of phasic component of GSR signal
	Phasic Std.	Micro-siemens	Standard deviation of tonic component of GSR
	Onset Rate	onset/sec	SCR onset rate per second
	Peak Amp. Mean	Micro-siemens	Mean of Peak amplitude (SCR)
	Rise Time Mean	ms	Mean of rise time (SCR)
	Recovery Time Mean	ms	Mean of recovery time (SCR)
Pupil	*Pupil Mean*	pixel	Mean of pupil size
	Pupil Std.	pixel	Standard deviation of pupil size

subjective responses immediately after each part assembly (cycle). Hence, this approach produced more subjective data and a better idea of how the dependent variables changed during the trials. The experiment was approved by the Human Subject Research office at the Rochester Institute of Technology. Informed consent was obtained from each participant before the experiment. Three sensors were used:

1) *BioHarness* is a wireless chest strap that allows the recording of an ECG signal. In addition to the ECG, the device provides respiration rate, heart rate, RR intervals, acceleration (3 axes), and device information.
2) *Shimmer3 GSR+* is a widely used device in research due to its Bluetooth connectivity. The device provides one GSR channel that measures the conductance of the skin and one PPG channel that measures the amount of reflected light (volumetric variations of blood circulation) from the vein [19]. The sensor sampling rate was set to 128 Hertz (Hz). During the experiment, we asked participants to minimize their motion when using the hand on which the sensor was placed. This was critical since the GSR signal is sensitive to motion artifacts and cannot be removed from the signal.
3) *Pupil labs* headset is open-source hardware (eyeglasses) that has three cameras, two of which look at the eyes and one point to the subject's perspective [20]. The eye cameras operate at 120 frames per second (fps), and the world camera records at 30 fps. This device is widely used in research and provides various signals such as pupil diameters, gaze location, and a real-time camera stream. This research used the headset to collect pupil dilation and gaze location.

Sensors were calibrated for each subject independently. The subjects started with a baseline recording where they sat in front of the robot and were asked to relax for five minutes. Those wearing glasses were asked to remove them for better pupil signal quality. The pupil lab headset was connected to a Samsung S8 smartphone with a Pupil Mobile app that transmits data to a local machine. The Shimmer3 GSR+ connects via Bluetooth to a computer, and a custom application [21] developed to acquire these signals was used to send data using the Lab Stream Layer (LSL) protocol. The BioHarness was connected over Bluetooth to a computer as well. All

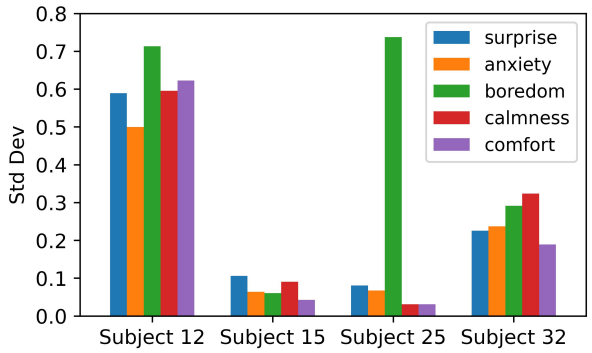


Fig. 2: Emotion scores variability for each testing subject shown as standard deviation (Std. Dev.) measured on the recorded emotion scores that have a range of $[0, 2]$.

devices were synchronized using the LSL stream library [22]. A modified version of the custom data collection app that generates automated event markers (*trial start/stop*) and manual event markers during data collection was used [23].

There are many ways to extract features from physiological signals, such as time and frequency domain, power density, morphological features, and so on. In this research, we extracted the most commonly used time and frequency features from each physiological signal. The extracted features are listed in Table II.

B. Data Split

We use the dataset described in Sec. III-A to perform emotion estimation, which consists of training a model to predict subject emotions from physiological inputs.

After cleaning up corrupted experiments (sensors failure and large noise in the signals), the dataset has a total of 32 subjects. We use 28 subjects as our *general* set, where we further split it into training and validation sets with a random split of 80% and 20%. This general set is used to train a general model on a rich amount of data, where different subjects have different responses to the human-robot collaboration experiment. In this way, the model has a general knowledge of the emotion estimation task. We use the remaining 4 subjects (12, 15, 25, and 32) to test our models on each separately. We chose these subjects because they represent a broad spectrum of variability in their emotion score responses, ranging from high (subject 12), low (subject 15), mixed (subject 25), and medium (subject 32) variability, as shown in Fig 2. Subject 25 was the only one who had coffee before the data collection experiment, possibly explaining their mixed standard deviation distribution in their emotions. We split each testing subject data into a training and test set, where the training set corresponds to the first 5 trials, and the testing set corresponds to the remaining 3 trials. We use the training set of the testing subject to fine-tune the general model or train models specific to the testing subject.

C. Machine Learning Models

We use a Multi-Layer Perceptron (MLP), a type of neural network, as our main model. We train the MLP to perform

regression over the emotion scores using the features obtained from physiological data as inputs (Table II). We followed [15] where an MLP model was used for emotion estimation on a different dataset using single-task learning. Unlike [15], we use a multi-task learning approach where regression is done on all emotions with a shared network and a multi-node output layer. Multi-task learning often performs better than single-task learning in neural networks [24].

We explore 6 different methods to perform emotion estimation, described as follows:

- **General model**: We train the MLP network using the general set. Then, we freeze the model. We generate predictions for the testing subjects using the frozen model. No fine-tuning is done on the testing subjects.
- **Fine-tuned model**: Similar to the general model, we train the MLP network using the general set. Then, we fine-tune the model using the available training set of the testing subject.
- **Subject-Specific model (S-specific)**: We train the MLP network on the training set of the testing subject. The general set is not used here.
- **General Mean (Mean-G)**: It uses the target mean value of the general set as its prediction. The MLP network is not used here.
- **Subject Mean (Mean-S)**: It uses the target mean value of the available training set of the testing subject as its prediction. The MLP network is not used here.

D. Training Details

We train a 3-layer MLP network, with 128 neurons in the first layer, 64 neurons in the second, and 5 output neurons corresponding to Surprise, Anxiety, Boredom, Calmness, and Comfort. ReLU activations are used in the hidden layers and dropout is used in them for regularization. The input to the MLP is normalized using quantile information[1] to reduce the impact of outliers. The 5 targets are normalized by making their range $[0, 1]$. For the general set, the MLP is trained for 130 epochs using AdamW with a learning rate of 2e-3, weight decay 1e-2, batch size 32, and Huber Loss. During fine-tuning, we use a learning rate of 1e-3 and a weight decay of 3e-2. Other hyperparameters are identical to general set training.

E. Evaluation

To evaluate model performance, we calculate the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) obtained by predicting the emotion scores on the test set of each subject (the last 3 trials). If a model uses the general set, we calculate the average performance over different training-validation splits of the general set using 5-fold cross-validation.

IV. Results

We show the results obtained for subjects 12, 15, 25, and 32, where we compare the performance of all the 5 different models mentioned in Section III-C. Tables III, IV, V, and VI show RMSE and MAE values as mean$\pm$standard deviation across the 5 folds of the general set, except for models that do not use the general set.

Table III shows the results obtained for subject 12. For RMSE, the *Fine-tuned* model performs best on anxiety and comfort while being the second best for calmness. The *General* model only achieves the best performance on surprise. This shows that per-subject fine-tuning improves overall performance. The *S-specific* model performs the best for Boredom, while the *Mean-S* model has the best performance for Calmness. The *Mean-G* is the second best for 3 emotions (surprise, anxiety, and boredom) For MAE, the *Mean-G* model performs the best overall.

Table IV shows the results obtained for subject 15. For RMSE, the *Mean-S* model is the best, achieving the lowest error for surprise, anxiety, boredom, and comfort. This is expected since subject 15 has the lowest standard deviation in their responses among the testing subjects, meaning that using the mean of their training set has a low error. The *Fine-tuned* model achieves the second-best performance by matching *Mean-S* on surprise and comfort while achieving the best performance on calmness. This suggests that even with a low standard deviation, the *Fine-tuned* model can still achieve competitive performance on subject 15. The *General* model and the *Mean-G* model perform poorly on this subject, showing that models using only the general set knowledge won't perform well on subjects with low standard deviation in their targets, hence, fine-tuning is needed. A similar trend to RSME is seen for MAE, where the *Mean-S* is the best model while the *Fine-tuned* is the second best.

Table V shows the results obtained for subject 25. This subject has a low standard deviation for surprise, anxiety, calmness, and comfort. Incidentally, for RMSE, we notice that *Mean-S* is the best model for those same emotions with low standard deviation. This is the same phenomenon happening in subject 15, where the mean of the testing subject yields good prediction performance. However, for boredom, which has a high standard deviation on subject 25, the model with the lowest error is *General*. This is expected since using only the mean of the training set of the testing subject is not a good prediction value when the standard deviation is high for a specific emotion. A similar trend to RSME is seen for MAE, where the *Mean-S* is the best model.

Finally, Table VI shows the results obtained for subject 32, which has medium variability in their emotion target values. For RSME, we see that *Mean-S* is the best model, achieving the lowest error for boredom, calmness, and comfort. Meanwhile, the *S-specific* model is the second-best, achieving the lowest error for surprise and anxiety. This suggests that, for subject 32, using the information of the subject itself suffices to be able to predict their emotion scores. We argue this happens because the variability of the emotion responses (standard deviation of the targets) in subject 32 is not high enough, making it easier to predict their targets using something as simple as the mean of their responses. A similar trend to

[1]We use the QuantileTransform function from scikit-learn with a uniform distribution and 100 quantiles.

TABLE III: **Results for subject 12.** Bold and underlined numbers indicate the best and second best performance respectively.

Models	RMSE					MAE				
	Surprise	Anxiety	Boredom	Calmness	Comfort	Surprise	Anxiety	Boredom	Calmness	Comfort
General	**0.250**$\pm$0.007	0.280$\pm$0.005	0.345$\pm$0.017	0.286$\pm$0.011	0.297$\pm$0.008	**0.197**$\pm$0.004	0.214$\pm$0.004	0.283$\pm$0.017	0.230$\pm$0.003	0.218$\pm$0.005
Fine-tuned	0.274$\pm$0.026	**0.258**$\pm$0.010	0.411$\pm$0.006	0.281$\pm$0.010	**0.279**$\pm$0.019	0.225$\pm$0.018	0.211$\pm$0.010	0.358$\pm$0.008	0.227$\pm$0.008	0.224$\pm$0.015
S-specific	0.272	0.269	**0.295**	0.346	0.352	0.198	0.218	**0.237**	0.284	0.292
Mean-G	0.271$\pm$0.001	0.266$\pm$0.001	0.323$\pm$0.002	0.282$\pm$0.001	0.294$\pm$0.001	0.223$\pm$0.001	**0.198**$\pm$0.001	0.256$\pm$0.002	**0.212**$\pm$0.000	**0.214**$\pm$0.000
Mean-S	0.277	0.276	0.353	**0.277**	0.287	0.233	0.208	0.292	0.215	0.229

TABLE IV: **Results for subject 15.** Bold and underlined numbers indicate the best and second best performance respectively.

Models	RMSE					MAE				
	Surprise	Anxiety	Boredom	Calmness	Comfort	Surprise	Anxiety	Boredom	Calmness	Comfort
General	0.118$\pm$0.012	0.115$\pm$0.002	0.437$\pm$0.041	0.286$\pm$0.044	0.108$\pm$0.012	0.113$\pm$0.013	0.110$\pm$0.003	0.429$\pm$0.042	0.281$\pm$0.045	0.105$\pm$0.012
Fine-tuned	**0.057**$\pm$0.001	0.028$\pm$0.001	0.039$\pm$0.008	**0.045**$\pm$0.001	**0.023**$\pm$0.001	0.035$\pm$0.001	0.021$\pm$0.001	0.032$\pm$0.009	**0.038**$\pm$0.001	0.009$\pm$0.001
S-specific	0.061	0.029	0.026	0.082	0.076	0.039	0.022	0.020	0.068	0.062
Mean-G	0.211$\pm$0.002	0.173$\pm$0.002	0.432$\pm$0.002	0.208$\pm$0.002	0.158$\pm$0.003	0.206$\pm$0.002	0.171$\pm$0.002	0.431$\pm$0.002	0.203$\pm$0.002	0.156$\pm$0.003
Mean-S	**0.057**	**0.027**	**0.024**	0.046	**0.023**	**0.034**	**0.020**	**0.019**	**0.038**	**0.008**

TABLE V: **Results for subject 25.** Bold and underlined numbers indicate the best and second best performance respectively.

Models	RMSE					MAE				
	Surprise	Anxiety	Boredom	Calmness	Comfort	Surprise	Anxiety	Boredom	Calmness	Comfort
General	0.134$\pm$0.016	0.149$\pm$0.007	**0.502**$\pm$0.030	0.319$\pm$0.044	0.137$\pm$0.006	0.129$\pm$0.017	0.145$\pm$0.008	**0.499**$\pm$0.031	0.316$\pm$0.043	0.134$\pm$0.006
Fine-tuned	0.031$\pm$0.002	0.029$\pm$0.001	0.562$\pm$0.016	0.033$\pm$0.002	0.019$\pm$0.000	0.024$\pm$0.001	0.024$\pm$0.001	0.538$\pm$0.016	0.028$\pm$0.002	0.015$\pm$0.001
S-specific	0.027	0.025	0.537	0.065	0.080	0.021	0.021	0.502	0.052	0.063
Mean-G	0.226$\pm$0.002	0.199$\pm$0.002	0.512$\pm$0.002	0.247$\pm$0.002	0.151$\pm$0.003	0.224$\pm$0.002	0.198$\pm$0.002	0.511$\pm$0.002	0.247$\pm$0.002	0.150$\pm$0.003
Mean-S	**0.026**	**0.022**	0.581	**0.017**	**0.018**	**0.020**	**0.019**	0.580	**0.012**	**0.013**

TABLE VI: **Results for subject 32.** Bold and underlined numbers indicate the best and second best performance respectively.

Models	RMSE					MAE				
	Surprise	Anxiety	Boredom	Calmness	Comfort	Surprise	Anxiety	Boredom	Calmness	Comfort
General	0.184$\pm$0.019	0.226$\pm$0.022	0.235$\pm$0.029	0.233$\pm$0.015	0.141$\pm$0.009	0.172$\pm$0.021	0.212$\pm$0.021	0.206$\pm$0.030	0.223$\pm$0.014	0.136$\pm$0.008
Fine-tuned	0.141$\pm$0.005	0.143$\pm$0.004	0.153$\pm$0.007	0.197$\pm$0.013	0.108$\pm$0.005	0.115$\pm$0.005	0.121$\pm$0.005	0.121$\pm$0.006	0.170$\pm$0.010	**0.096**$\pm$0.005
S-specific	**0.110**	**0.122**	0.150	0.214	0.163	**0.082**	**0.095**	0.116	0.184	0.141
Mean-G	0.180$\pm$0.002	0.176$\pm$0.002	0.307$\pm$0.002	0.211$\pm$0.002	0.131$\pm$0.003	0.175$\pm$0.002	0.165$\pm$0.002	0.288$\pm$0.002	0.203$\pm$0.002	0.128$\pm$0.003
Mean-S	0.121	0.139	**0.132**	**0.163**	**0.102**	0.107	0.123	**0.100**	**0.155**	0.098

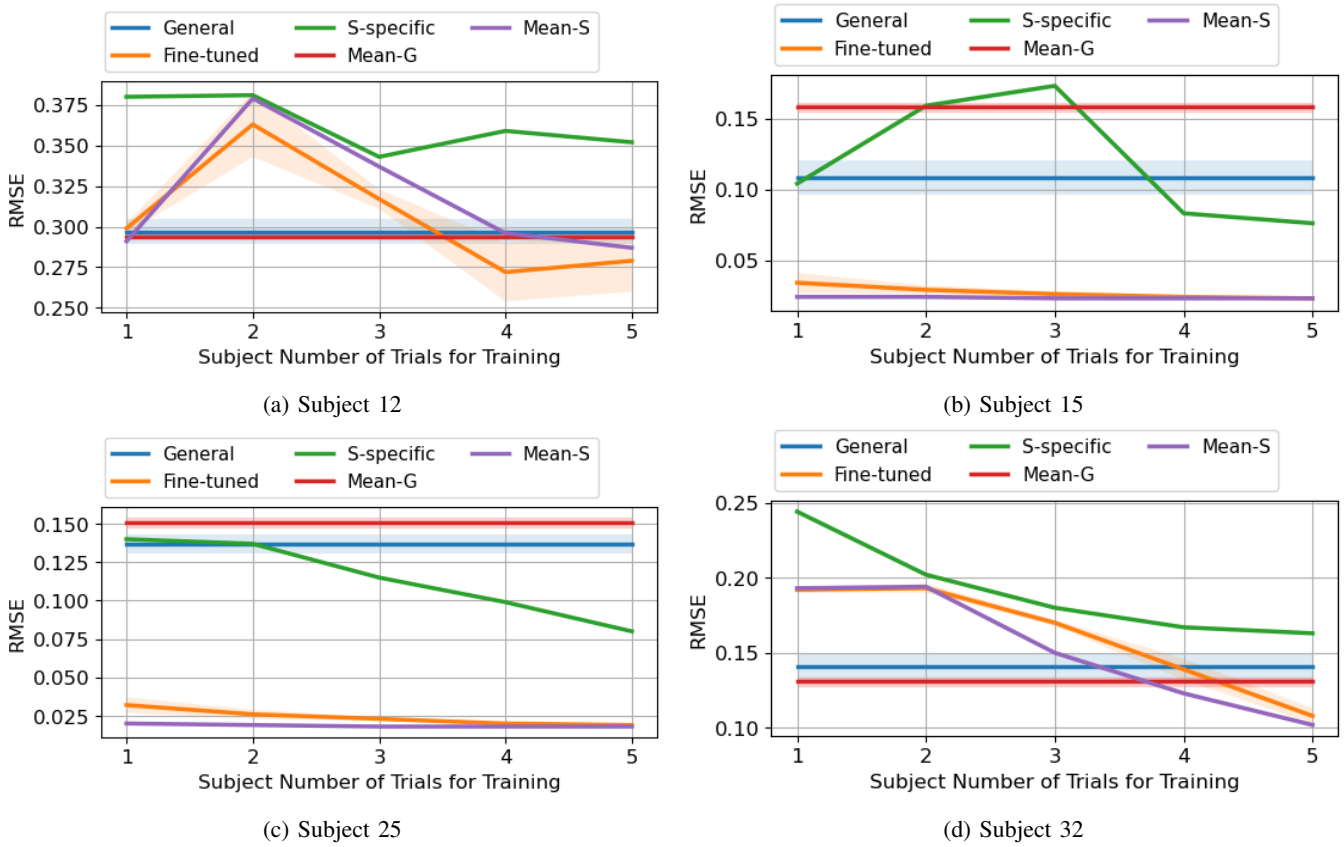
RSME is seen for MAE, where the *Mean-S* is the best model, *S-specific* model is the second-best, and the *Fine-tuned* is competitive to both.

In summary, we found that when subjects have low emotional variability, models that only depend on the current subject data (*S-Specific* and *Mean-S*) yield better performance than others. However, when the emotional variability is high, models that leverage a general dataset (*General*, *Fine-tuned*, and *Mean-G*) perform better. It is worth noticing that in all cases the *Fine-tuned* model is competitive in performance against the winning models, making it the best option when the emotional variability of the current subject is not known.

A. Data efficiency

We explore how the models perform when the training data from the testing subject is scarce. This can show which method is more data-efficient in terms of how much training data is needed from the testing subject to reach good performance.

Fig. 3 summarizes the result for the 4 testing subjects for the Comfort emotion. The testing data of each subject does not change, as this is still the last 3 trials of each subject. We vary the number of trials available from the subject for training (fine-tuning), going from only 1 trial to 5 trials. Across the subjects, we can see that using more training data from the subject improves performance for *Fine-tuned*, *S-specific*, and *Mean-S* models since these models can leverage that data. *General* and *Mean-G* models stay constant across the number of trials because they do not use the training data from the testing subject. The results for subjects 12 and 32 are shown in Fig. 3a and Fig. 3d respectively. We can see that, while *General* and *S-specific* models are competitive when 3 or fewer training trials are available from the subject, eventually, the *Fine-tuned* and *S-mean* models surpass them when more training data from the subject is available. This shows that a model that leverages data from the subject of interest performs better when more data is available. Results for subjects 15 and 25 are shown in Fig. 3b and Fig. 3c respectively. We can see that the *General* and *Mean-G* models are no longer competitive since these 2 subjects have low variability for their targets. The *Mean-S* model performs the best across the number of training trials available, however, the *Fine-tuned* model matches the *Mean-S* performance when more training trials are available.



(a) Subject 12

(b) Subject 15

(c) Subject 25

(d) Subject 32

Fig. 3: Data efficiency experiment results for the Comfort emotion where the number of trials for fine-tuning varies. (a) Subject 12, (b) Subject 15, (c) Subject 25, and (d) Subject 32

Based on the results described above, we can conclude that if the subject has high variability in their target scores, a model only using the general set data will perform better when training trials from the current subject are scarce. However, with more training trials available from the subject, models that leverage that data can perform better. If the variability in the subject target scores is low, then models only using the general set data perform poorly, as seen in subjects 25 and 32. In these cases, it is better to leverage any amount of training data available from the subject. Finally, we see across all subjects that the *Fine-tuned* model eventually surpasses all other models (or it becomes competitive) when more data is available from the subject.

V. Discussion and Conclusion

Effective human-robot collaboration hinges on a robot's ability to detect, learn from, and adapt to human psycho-physiological states. In this study, we collected data from a human-robot interaction to train a neural network for predicting human emotions. We found that fine-tuning a model pre-trained on general data yields the best results or becomes competitive compared to other baselines. Simple models, such as using the mean emotional response, are effective when emotional variability is low. Additionally, with more data available from the subject of interest, models that leverage this data perform better than fixed pre-trained general models. A key limitation of this study is the use of a small neural network, which may restrict performance. However, a larger network risks overfitting due to limited data. Future work could explore deeper and more complex networks (e.g. transformers) by collecting more data. Additionally, collecting more data can increase the number of testing subjects, potentially giving us more insights into the generalization of the proposed models. Another future direction is only fine-tuning portions of the network, rather than all of its components [25], [26]. This can reduce overfitting on the subject of interest. Finally, future work could explore using self-supervised learning techniques during pre-training instead of supervised learning, allowing the model to learn from unlabeled collected data, potentially creating features that generalize better to testing subjects during fine-tuning [27].

Acknowledgments. This work was partly supported by NSF awards #2326491 and #2125362. The views and conclusions contained herein are those of the authors and should not be interpreted as representing any sponsor's official policies or endorsements.

REFERENCES

[1] S. H. Fairclough, "Fundamentals of physiological computing," *Interacting with Computers*, vol. 21, pp. 133–145, Jan. 2009.

[2] NSF, "Information and Intelligent Systems (IIS): Core Programs." https://www.nsf.gov/pubs/2018/nsf18570/nsf18570.htm, 2019.

[3] D. Kulic and E. A. Croft, "Anxiety detection during human-robot interaction," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 616–621, IEEE, 2005.

[4] D. Kulić and E. Croft, "Physiological and subjective responses to articulated robot motion," *Robotica*, vol. 25, pp. 13–27, Jan. 2007.

[5] L. Tiberio, A. Cesta, and M. Belardinelli, "Psychophysiological Methods to Evaluate User's Response in Human Robot Interaction: A Review and Feasibility Study," *Robotics*, vol. 2, no. 2, pp. 92–121, 2013.

[6] M. Nokata, K. Ikuta, and H. Ishii, "Safety-optimizing method of human-care robot design and control," in *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No.02CH37292)*, vol. 2, pp. 1991–1996, IEEE, 2002.

[7] D. Kulic and E. A. Croft, "Real-time safety for human - robot interaction," in *ICAR '05. Proceedings., 12th International Conference on Advanced Robotics, 2005.*, vol. 2005, pp. 719–724, IEEE, 2005.

[8] D. Kulic and E. Croft, "Estimating Robot Induced Affective State using Hidden Markov Models," in *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 257–262, IEEE, Sept. 2006.

[9] D. Kulic and E. A. Croft, "Affective State Estimation for Human–Robot Interaction," *IEEE Transactions on Robotics*, vol. 23, pp. 991–1000, Oct. 2007.

[10] D. Novak, B. Beyeler, X. Omlin, and R. Riener, "Workload estimation in physical human-robot interaction using physiological measurements," *Interacting with Computers*, vol. 27, no. 6, pp. 616–629, 2015.

[11] I. Iturrate, R. Chavarriaga, L. Montesano, J. Minguez, and J. D. R. Millán, "Teaching brain-machine interfaces as an alternative paradigm to neuroprosthetics control," *Scientific Reports*, vol. 5, pp. 1–11, 2015.

[12] S. K. Ehrlich and G. Cheng, "A Feasibility Study for Validating Robot Actions Using EEG-Based Error-Related Potentials," *International Journal of Social Robotics*, vol. 11, pp. 271–283, Apr. 2019.

[13] A. F. Salazar-Gomez, J. Delpreto, S. Gil, F. H. Guenther, and D. Rus, "Correcting robot mistakes in real time using EEG signals," in *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 6570–6577, IEEE, 2017.

[14] F. Dehais, E. A. Sisbot, R. Alami, and M. Causse, "Physiological and subjective evaluation of a human-robot object hand-over task," *Applied Ergonomics*, vol. 42, no. 6, pp. 785–791, 2011.

[15] C. Savur, J. Heard, and F. Sahin, "Human comfortability index estimation in industrial human-robot collaboration task," *arXiv preprint arXiv:2308.14644*, 2023.

[16] C. Savur, J. Heard, and F. Sahin, "Human Comfortability Index Estimation in Industrial Human-Robot Collaboration," *IEEE Transactions on Human Machine Systems*, pp. 1–12, 2022.

[17] S. Kumar, C. Savur, and F. Sahin, "Dynamic Awareness of an Industrial Robotic Arm Using Time-of-Flight Laser-Ranging Sensors," *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2850–2857, 2018.

[18] S. Kumar, *Dynamic Speed and Separation Monitoring with On-Robot Ranging Sensor Arrays for Human and Industrial Robot Collaboration*. PhD thesis, Rochester Institute of Technology, 2020.

[19] E. Mejía-Mejía, J. M. May, R. Torres, and P. A. Kyriacou, "Pulse rate variability in cardiovascular health: A review on its applications and relationship with heart rate variability," *Physiological Measurement*, vol. 41, p. 07TR01, Aug. 2020.

[20] M. Kassner, W. Patera, and A. Bulling, "Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, (New York, NY, USA), pp. 1151–1160, ACM, Sept. 2014.

[21] C. Savur, S. Kumar, S. Arora, T. Hazbar, and F. Sahin, "HRC-SoS: Human robot collaboration experimentation platform as system of systems," in *2019 14th Annual Conference System of Systems Engineering, SoSE 2019*, pp. 206–211, IEEE, 2019.

[22] SCCN, "Lab Stream Layer (LSL)," 2018.

[23] C. Savur, S. Kumar, and F. Sahin, "A framework for monitoring human physiological response during human robot collaborative task," *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, vol. 2019-Octob, pp. 385–390, 2019.

[24] M. Crawshaw, "Multi-task learning with deep neural networks: A survey," *arXiv preprint arXiv:2009.09796*, 2020.

[25] T. L. Hayes, K. Kafle, R. Shrestha, M. Acharya, and C. Kanan, "Remind your neural network to prevent catastrophic forgetting," in *European Conference on Computer Vision*, pp. 466–483, Springer, 2020.

[26] M. Y. Harun, J. Gallardo, T. L. Hayes, R. Kemker, and C. Kanan, "Siesta: Efficient online continual learning with sleep," *Transactions on Machine Learning Research (TMLR)*, 2023.

[27] J. Gallardo, T. L. Hayes, and C. Kanan, "Self-supervised training enhances online continual learning," in *BMVC*, 2021.