Personalized Productive Engagement Recognition in Robot-Mediated Collaborative Learning

Vetha Vikashini C.R.*
SMART Lab, Department of Computer Science
New York University Abu Dhabi
Abu Dhabi, UAE
vethssvikas1@gmail.com

Hanan Salam*
SMART Lab, Department of Computer Science
New York University Abu Dhabi
Abu Dhabi, UAE
hanan.salam@nyu.edu

Jauwairia Nasir
CHILI Lab, Ecole Polytechnique Fédérale de Lausanne
Lausanne, CH
jauwairia.nasir@epfl.ch

Barbara Bruno
CHILI Lab, Ecole Polytechnique Fédérale de Lausanne
Lausanne, CH
barbara.bruno@epfl.ch

Oya Celiktutan
Department of Engineering
King's College London
London, UK
oya.celiktutan@kcl.ac.uk

ABSTRACT

In this paper, we propose and compare personalized models for Productive Engagement (PE) recognition. PE is defined as the level of engagement that maximizes learning. Previously, in the context of robot-mediated collaborative learning, a framework of productive engagement was developed by utilizing multimodal data of 32 dyads and learning profiles, namely, Expressive Explorers (EE), Calm Tinkerers (CT), and Silent Wanderers (SW) were identified which categorize learners according to their learning gain. Within the same framework, a PE score was constructed in a non-supervised manner for real-time evaluation. Here, we use these profiles and the PE score within an AutoML deep learning framework to personalize PE models. We investigate two approaches for this purpose: (1) Single-task Deep Neural Architecture Search (ST-NAS), and (2) Multitask NAS (MT-NAS). In the former approach, personalized models for each learner profile are learned from multimodal features and compared to non-personalized models. In the MT-NAS approach, we investigate whether jointly classifying the learners' profiles with the engagement score through multi-task learning would serve as an implicit personalization of PE. Moreover, we compare the predictive power of two types of features: incremental and non-incremental features. Non-incremental features correspond to features computed from the participant's behaviours in fixed time windows. Incremental features are computed by accounting to the behaviour from the beginning of the learning activity till the time window where productive engagement is observed. Our experimental results show that (1) personalized models improve the recognition performance with respect to non-personalized models when training models for the gainer vs. non-gainer groups, (2) multitask NAS (implicit personalization) also outperforms non-personalized models, (3) the speech modality has high contribution towards prediction, and (4) non-incremental features outperform the incremental ones overall.

CCS CONCEPTS

• **Human-centered computing** → **User models**; **Interaction techniques**; • **Applied computing** → *Interactive learning environments*; • **Computing methodologies** → **Artificial intelligence**; **Supervised learning by regression**; **Multi-task learning**; **Classification and regression trees**.

KEYWORDS

Engagement Prediction; Personalized Affective Computing; Embodied Interaction; Human-robot/Agent Interaction; Social Signals; Personalization; Social Robotics in Education

ACM Reference Format:
Vetha Vikashini C.R., Hanan Salam, Jauwairia Nasir, Barbara Bruno, and Oya Celiktutan. 2022. Personalized Productive Engagement Recognition in Robot-Mediated Collaborative Learning. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3536221.3556569

1 INTRODUCTION

With the increase in technological advancements, commercial availability of several robots, and the readiness to integrate technology in various fields including education, we can indeed witness a rise of social robotics in learning settings [15]. In what way can these robots be beneficial for advancing the pedagogical goal is still an open question.

One line of research that stands out in educational Human Robot Interaction (HRI) and Intelligent Tutoring Systems (ITS) is using robots to personalize learning strategies to individual needs in order to cater for the learning goal as not everyone learns in the same way [6, 18, 31]. Further, another line of research focuses more

*Both authors contributed equally to this paper.

ICMI '22, November 7–11, 2022, Bengaluru, India

ACM ISBN 978-1-4503-9390-4/22/11…$15.00
https://doi.org/10.1145/3536221.3556569

on equipping robots with the ability to infer whether the learners are engaged in the learning activity at hand. We believe this goes hand in hand with personalization as *the better the robot is aware about the students individual characteristics, the better it can detect the engagement state of the learners, which itself can be manifested in several ways*. Similarly, *the better this engagement state is inferred, the better personalized learning interactions the robot can offer*. Existing approaches in engagement recognition employ non-verbal behaviours as predictors of engagement [38]. The most recent approaches also make use of deep learning architectures to train engagement models which has proven to outperform traditional machine learning models [3, 36]. The current approaches consider one-fits-all paradigms which consist of training static models without taking into consideration specific characteristics of the individuals present in the dataset. Such models are simple to train, however, their accuracy in predicting the users engagement state is compromised. Moreover, previous research in behavioural and social sciences have revealed differences among different individuals in conveying their engagement state. Zacherman and Foubert [44] and Conner [5] highlight the need for customized models, and the potential of such models in offering more accurate decisions.

The concept of Productive Engagement (PE) was introduced by Nasir et al. [23] with the aim to conceptualize engagement in educational settings. It is defined as *the level of engagement that maximizes learning*, or in other words, *engagement that is conducive to learning* where the authors present it as a "hidden hypothesis that links multimodal behaviors of the users to learning and performance". In contrast to existing work in engagement conceptualization, Nasir et al. [23] argue that over-engagement, similar to under-engagement, can lead to decreased learning outcomes. The proposed concept was validated in the context of a robot-mediated human-human collaborative learning activity implemented within the *JUSThink* platform by Nasir et al. [28], where the employed approach surfaces multiple ways (precisely sets of multimodal behaviors given by log, audio, and video data) in which different individuals learn [25]. Through an in-depth quantitative and qualitative analysis, three learner profiles are identified, namely, *Expressive Explorers*, *Calm Tinkerers*, and *Silent Wanderers* where the first two profiles correspond to those who have high learning gain, while the last one corresponds to those who exhibit a low learning gain [27]. The idea of conceptualizing engagement with a focus on the learning aspect in educational settings and the aforementioned outcomes indeed show potential in advancing the literature in educational HRI and multimodal learning analytics. Furthermore, we believe that personalization would further enhance learner models given that personalization techniques catering for the unique requirements of learners have been found to be beneficial as mentioned previously.

Until recently, the Human-Machine Interaction (HMI) community has focused on one-fits-all approaches. Few approaches have attempted to train personalized models for affective and personality computing tasks such as mood [42], engagement [35], and emotion recognition [40], or personality traits prediction [39]. Among the employed machine learning methods for learning personalized models for HMI tasks, multitask learning was used to learn individual-specific models for mood and stress prediction. Through weight sharing, multitask learning has the ability to learn individual user models while leveraging data across the population. In a recent work, Salam et al. [39] proposed personalized models of big five personality traits via Efficient Neural Architecture Search [14] and it was shown to outperform state-of-the-art approaches in personality computing.

In this paper, we then seek to combine the positive aspects of *personalization* and adaptation with a deeper understanding of what *productively engaged* learners look like. We extend the approach proposed by Salam et al. [39] to the task of productive engagement prediction. Hence, we propose to learn personalized deep architectures for different learner profiles identified by Nasir et al. [27] using single-task and multi-task Efficient Neural Architecture Search (ENAS). Via multitask ENAS, we investigate whether simultaneously predicting productive engagement scores and classifying learners profiles can efficiently offer an implicit personalization of productive engagement models.

Previous research by Oertel et al. [29] has shown that engagement is a dynamic process that varies in time via a dynamic evolution mechanism between (two or more) interaction parties. Existing approaches in engagement recognition have focused on using user behavioural features extracted in fixed time windows or time-frames and using these features to recognize the engagement state in the time window in question. However, the user's prior behaviours can be indicative of the their current state of mind [17, 32]. The dataset used by Nasir et al. [26] includes a set of incremental features computed from the beginning of the learning activity till the time window where productive engagement is observed rather than focusing on the behaviours within the time window in question. In this work, we conduct a comparative study between incremental and non-incremental features. Such analysis allows to investigate the effect of past behaviours on the predictive power of engagement recognition models.

In summary, the main contributions of this work are:

(1) We build personalized models for automatic prediction of productive engagement in robot-mediated collaborative gamified learning. To this end, we use a Neural Architecture Search framework for designing and training separate models per learner profile. We evaluate our approach with a rich open source data set of multimodal features developed by Nasir et al. [26] including log data, gaze, affect, and speech, both singly as well as their early fusion.
(2) We investigate multitask learning for implicitly accounting for different learner profiles in the model.
(3) We compare incremental and non-incremental features serving to investigate whether accounting for multimodal behaviour from the beginning of the learning activity until the time window where productive engagement is observed has an effect on the current learner's productive engagement state.

The rest of the paper is organized as follows: Section 2 reviews related work in learner's mental states analysis in education settings, including engagement recognition, and personalized models in Human-Machine Interaction. Section 3 introduces the productive engagement dataset used in our framework. Section 4 presents the proposed approach. Section 5 reports the performance evaluation of the proposed approach. Finally, section 6 concludes the paper.

2 RELATED WORK

In this section, we review relevant work in learner's mental states analysis in education settings, including engagement recognition, and personalized models in Human-Machine Interaction (HMI).

2.1 Learner's Mental States Analysis in Education Settings

Personalizing an intelligent system's actions and decisions to individual differences in education settings is compulsory for achieving a better learning performance and a higher level of learner's satisfaction. The first step to build personalization mechanisms is the understanding of human observable behaviours. To this effect, there have been some works by Alyuz et al. [2], Gupta et al. [12], Kamath et al. [16], Mustafa et al. [22], Pham and Wang [30], which focused on automatically detecting learner's mental states associated with learning such as satisfaction, confusion, engagement or boredom.

Different studies have approached to the problem within the context of in-class teaching. For instance, Alyuz et al. [2] collected data from 20 students (14-15 years) who partook a math course over the course of several months. Each student worked independently in the class using a laptop, and was recorded using a 3D camera. The recordings were annotated by experts in educational psychology with respect to the affective states of excited, calm, bored, confused and unknown. From the recordings, they extracted two types of features, namely, appearance features and contextual features. While appearance features were composed of face location, head pose, facial gestures and seven basic facial emotions (e.g., happiness, sadness, etc.), contextual features were extracted from (i) user profiles including age, gender; (ii) session information including video duration, time within a session; and (iii) performance features including number of trials until success, number of used hints, grade.

A line of work by Gupta et al. [12], Kamath et al. [16], Mustafa et al. [22], Pham and Wang [30] has focused on predicting learner's mental states during MOOCs. Dhall et al. [8] introduced a subchallenge for predicting learner's engagement level, from disengaged to highly engaged, in large-scale real-life video recordings in the 2018 Emotion Recognition in the Wild (EmotiW) challenge. Among these methods, Pham and Wang [30] modelled human behaviours from physiological signals, which is out of scope of this work. Kamath et al. [16] recorded 23 students using a web camera mounted on a computer screen while viewing a video lecture for a duration of 10 minutes. Annotations regarding student's engagement states, namely, not engaged, nominally engaged, and very engaged, were collected from external observers recruited via a crowdsourcing service. For recognising engagement states, first Histogram of Gradients (HoGs) were extracted from the face region, and then fed into the instance-weighted Support Vector Machines together with Multiple Kernel Learning framework where the importance of a particular sample in the training data was obtained from the crowdsourced annotations. In their following work [12], the previously collected dataset was enriched by incorporating video recordings from up to 119 students and additional annotations such as boredom, confusion, and frustration, again collected via crowdsourcing. The enlarged dataset, called DAiSEE dataset, was used to train/fine-tune widely used CNN models for image classification (e.g., InceptionNet V3) and video classification (e.g., C3D, LRCN - Long-term Recurrent Convolutional Networks). Mustafa et al. [22] took an approach similar to what was followed by Gupta et al. [12] for data collection and annotation, where a total of 75 participants were recorded via a video-conferencing setup, and a team of 5 annotators provided ratings with respect to four engagement levels ranging from completely disengaged to highly engaged. They proposed to use a Deep Multi-Instance Learning (DMIL) framework where the task of engagement prediction was formalised as a regression problem from weakly labeled data. The DMIL was trained using facial features that were Local Binary Patterns extracted from three orthogonal planes, and outperformed classical regression methods such Support Vector Regression for engagement intensity estimation.

In HRI, some studies also formalized the task of engagement prediction as a regression problem. For instance, Del Duchetto et al. [7] proposed a novel regression model (utilizing CNN and LSTM networks) to compute a single scalar engagement from video streams, obtained from the point of view of an interacting robot. Similarly in the work of Rossi et al. [34], different feature selection and regression models were compared to predict a user's engagement state along three dimensions: affective, cognitive and behavioural. The study found that characterising each dimension separately in terms of features and regression leads to better results compared to a model directly combining the three dimensions.

One line of research that stands out in educational Human Robot Interaction (HRI) and Intelligent Tutoring Systems (ITS) is using robots to personalize learning strategies to individual needs in order to cater for the learning goal as not everyone learns in the same way [6]. For example, Leyzberg et al. [18] employed a setting where participants try to solve grid-based puzzles with a personalized or a non-personalized tutor. When comparing the time it took for the students to solve the puzzles, the authors observed improvement in the post-test of those students who dealt with a personalized learning interaction. Then, Ramachandran et al. [31] showed that when a robot tutor provides breaks to students according to their performance gain or performance drop rather than at fixed times, it has a positive effect on the learning gains of the students.

2.2 Personalized Models in Human-Machine Interaction

Until recently, the Human-Machine Interaction (HMI) community has focused on one-fits-all approaches. Few approaches have attempted to train personalized models for affective and personality computing tasks such as pain, mood, and emotion recognition, or personality traits prediction. Personalized models can be trained to take into account characteristic and behavioural differences among (1) individual users, or a (2) group of users. Different characteristics include the users age, gender, culture, or even personality. Behavioural differences entail clustering behavioral patterns based on correlations of the users behavioural cues and the target task.

Personalizing models for individual users involves training models on each user's data. Such personalization allows to tailor the models towards the specific individual assuming that different individuals behave differently and have unique ways in conveying their state. Multitask learning was used in the literature for this

purpose due to its ability to learn individual user models while leveraging data across the population through weight sharing. Example approaches that used individual-level multitask learning for personalization include the work of Jaques et al. [13], Taylor et al. [42] who used multitask learning to train individual models for mood, stress, and health prediction. Another work that employed multitask learning with Gaussian process regression models to personalize self-reported pain prediction is by Liu et al. [20].

Within the area of facial expression recognition, personalization is performed to take into account facial appearance differences among individuals. While facial expressions are considered universal, the differences in facial shapes and textures affect the accuracy of expression recognition models. Consequently, training individual-level adaptive models allows a better modeling of facial expressions variability. An example approach for individual-level personalized expression recognition include supervised domain adaption with mixture of experts [10]. Another approach by Shahabinejad et al. [40] proposed a CNN architecture to learn and propagate individual deep facial features followed by a spatial attention map, which is then provided as an input to another CNN. For the task of personality computing, Shao et al. [41] proposed to learn individual-specific graph representations for personality traits recognition in a human-human interaction scenario. Individual-specific CNN architecture is learned from the conversational partner's (speaker) non-verbal cues to predict the target individual's facial reactions (listener). The learned individual-specific CNN's parameters and layer weights are then used to create a person-specific graph representation which is then provided as an input to a residual gated graph convolution neural network for personality prediction.

In the context of educative settings, few approaches for personalization were proposed. The study by Alyuz et al. [2] demonstrated that learning student-specific personalized models for learners' confusion and satisfaction (engagement emotional states) classification outperformed one-fits-all generic. Two different personalization approaches were compared in this work: the adapted approach augments the one-fits-all model's training data with student-specific data; the personal approach uses person-specific data to train the model. The personal approach trained with the contextual features outperformed the other models overall. In another work, Alyuz et al. [1] proposed a semi-supervised model for personalizing engagement emotional states (satisfied, bored, confused). Students were instructed to provide their engagement emotional states at randomised times during the course of the learning task.

Personalizing models with respect to differences among a group of users entails dividing users into different profiles characterizing each group, and then training models that make use of the users' profile information. Such information can be used at the data-level or the model-level. At the data-level, personalized models can be trained by creating profile-specific datasets which are used to train the models. At the model-level, the users profiles can be used within the models learning process. Profile-wise personalization was explored in the literature in few HMI contexts for a number of tasks. For instance, in a multi-party HRI context, Salam et al. [37] used users personality scores as features to predict individual and group engagement with a robot. In an HRI autism therapy framework, child-specific deep learning models of valence, arousal, and engagement were trained [35]. Culture and gender profiling information were used within specific layers of the deep architecture to nest the children based on these profiles. This was followed by individual network layers for each child. In another work by Rudovic et al. [36], a deep learning architecture, called CultureNet, was introduced, which used culture data to tailor the model towards each culture and child.

2.2.1 Neural Architecture Search. Neural architecture search (NAS) is a technique of automated machine learning that aims to automate the design of artificial neural networks (ANN) architectures, which were shown to be on par or outperform manually-designed architectures [9, 33]. NAS mainly works by evaluating a large number of architectures across a search space (defining the ANN type) using a search strategy (approach used for the search space exploration) and selecting the optimal architecture for a certain task via a performance estimation strategy. Among the existing search algorithms we can find NASNet [11], Progressive NAS (PNAS) [19], and Efficient NAS (ENAS) [14]. NAS has been applied in the literature for training personalized models in various domains of application. These include personalized human pose estimation [43], efficient object recognition [4], and heart rate estimation from faces [21], among others. In a recent work, Salam et al. [39] proposed to learn gender-wise and age-wise personalized models of big five personality traits. ENAS was employed to automatically learn deep learning architectures for different user profiles from multimodal behavioural features. This work extends the approach of Salam et al. [39] to the task of productive engagement prediction based on different learners' profiles. Additionally, multitask ENAS is also proposed for implicitly personalizing productive engagement models. Implicit personalization is investigated by jointly classifying learners' profiles and predicting productive engagement scores.

3 PE-HRI-TEMPORAL DATASET

We use the open source PE-HRI-temporal dataset developed by Nasir et al. [26] generated from a study done with the JUSThink platform [28] where 68 children (in teams of two i.e., 34 teams) aged 9 to 12 years interact with a collaborative learning platform consisting of two screens and a QTrobot acting as a guide and a mediator. The learning aim of this activity was to impart the knowledge of the minimum spanning tree problem. This was presented in a scenario based map of Switzerland. The dataset comprises of 28 multi-modal behaviours which were extracted from log, video (affect and gaze), and audio (speech) data (as seen in Section 4.2) where the average duration of the interaction data per team is 20 minutes that is organized in windows of 10 seconds. Hence, this gives a total of 5048 windows. In addition to this, it also contains the performance metrics and the learning gains of the teams. We must note here that while the dataset provides 5048 data points with 34 teams; the learner profiles (discussed later in Section 4.1) are only available for 32 teams [27]; hence, the data used in this work consists of 4676 data points from the dataset.

Productive Engagement Scores – For the *Productive Engagement Score*, we make use of the metric proposed by Nasir et al. [24] to characterize productive engagement in real-time. Briefly, the

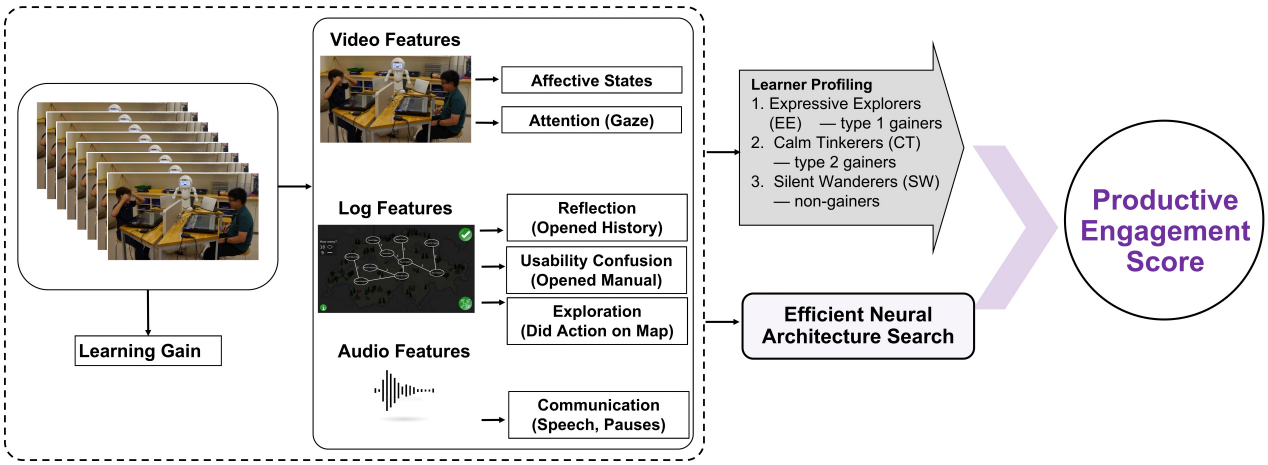

Figure 1: Overview of the proposed approach: Learners are clustered into three profiles based on their multimodal behavioural features [26] and learning gain: Expressive Explorers (Type 1 gainers), Calm Tinkerers (Type 2 gainers), and Silent Wanderers (non-gainers) [27]. Behavioural features and learners' profiles are then fed as inputs to the Efficient Neural Architecture Search (ENAS) framework to automatically search for the optimal architecture for each profile and for each modality to predict productive engagement.

score is generated using a linear combination of speech behaviors that have been found most discriminating between those who learn and those who do not, for example, the amount of interjections or overlap in speech of the two team members or the amount of long pauses that the teams made in their speech activity. For more details, see the work of Nasir et al. [24]. This score is what the regression models, in this paper, will try to predict.

4 PERSONALIZED PRODUCTIVE ENGAGEMENT PREDICTION

We propose to learn personalized models of productive engagement of teams of children (a team is composed of two members) involved in a robot-mediated learning task. Our approach can be decomposed into three steps: (1) Learners' profiling, (2) Features extraction, and (3) Efficient Neural Architecture Search (ENAS). ENAS is investigated for this task within two frameworks: (1) single-task profile-level personalization, and (2) multitask learning personalization. In this work, we rely on the learners' profiles and features extracted by Nasir et al. [27] and discussed below. The workflow of the proposed approach is depicted in Figure 1.

4.1 Learners' Profiling

The first step of the proposed approach is learners' profiling. Nasir et al. [27] applied a clustering approach to multimodal learners behaviours and their associated learning gain metrics followed by comparing the two sets of clusters obtained in terms of the teams they consist of; thus, allowing for the identification of three types of learner profiles: Expressive Explorers (EE), Calm Tinkerers (CT) and Silent Wanderers (SW). While EE and CT are gainers (high learning outcomes), SW falls in the non-gainer category (low learning outcomes). These profiles are used in this work to personalize the productive engagement models. The following is a brief description of the profiles. For a detailed description, the reader is referred to the work of Nasir et al. [27].

(1) **Expressive Explorers** belong to the category of high gainers. Teams belonging to this profile demonstrate effective communication, high reflection periods and a pronounced exploratory approach when involved in a learning task. They tend to exhibit high expressiveness of their emotional state (higher arousal and negative valence).
(2) **Calm Tinkerers** also belong to the category of high gainers. Teams belonging to this profile are similar to the EE profile in terms of communication, reflection, and exploration. However, they exhibit a relatively calm emotional state characterized by lower arousal and negative valence.
(3) **Silent Wanderers** belong to the non-gainers category. This profile is characterized by poorer communication, lower reflection periods, and high frustration states.

4.2 Features

In this work, a set of multimodal behaviour features are explored for training personalized productive engagement models. These features are provided with the publicly available dataset used in this work (please refer to Section 3). They include:

(1) **Log Features**: These features represent behaviours describing the team's interaction with the learning activity. These include features characterising reflection (e.g., the team members opened the history of their submissions, and reflected on them), usability confusion (e.g., the team members opened the instructions manual), and exploration (e.g., team's actions to solve the learning activity problem on the learning platform).

(2) **Affect Features**: These features represent the team members' displayed affective state along the valence and arousal dimensions.
(3) **Gaze Features**: In order to gauge the attention of the learners in the collaborative settings, gaze features represent the team members gaze behaviours (e.g. looking at the partner or the robot).
(4) **Speech Features**: These features represent the audio-related team members behaviours such as silence, pauses and overlaps.
(5) **Time Feature**: The time window of which the features are computed for a given team. This is measured with respect to the total duration of the task. The time information enables the investigation of temporal dynamics of engagement, and whether including the point in time (i.e., beginning, middle, or end of learning activity) of which the multimodal behaviours are computed contribute to the predictive power of the productive engagement models.

In the dataset, two types of features were computed:

(1) **Non-Incremental Features**: The average value of a feature in that particular time window.
(2) **Incremental Features**: The average value of a feature from the beginning of the learning activity until that particular time window.

Incremental features serve to investigate whether accounting for the team's behaviour from the beginning of the learning activity till the time window where productive engagement is observed have an effect on the current productive engagement state.

4.3 Personalized Neural Architecture Search Strategy

In order to personalize productive engagement models, we propose two frameworks: (1) Profile-level personalization, and (2) Multi-task learning personalization. These frameworks are described below:

Profile-level Personalization: Profile-level personalization entails explicitly using the learners' profiles to create different datasets for each profile, and using Efficient Neural Architecture Search (ENAS) to learn adaptive architectures per profile. The learner profiles are used to separate the learners into different groups. An adaptive neural architecture is then automatically designed and trained for each profile using the extracted features (log, affect, gaze, speech, time) as input to ENAS. Proposed by Jin et al. [14], ENAS employs a search space defined by network morphism operations such as new layers insertion, existing layers expansion, or skip connections addition. Bayesian optimization is used to guide an Efficient exploration of the search space. We use Auto-Keras, an open-source AutoML implementation of ENAS, to implement and train our models which was developed by Jin et al. [14].

Implementation Details – A default architecture composed of two dense layers with 32 units was used. An 85 – 15% split strategy was used to divide the training dataset into training and validation sets. Mean squared error is used as loss function, and each network is trained with ADAM optimiser. The number of epochs was set to 100. The number of trials was set to 10. A trial is a parameter of Neural Architecture Search. It corresponds to the maximum number of different models to try.

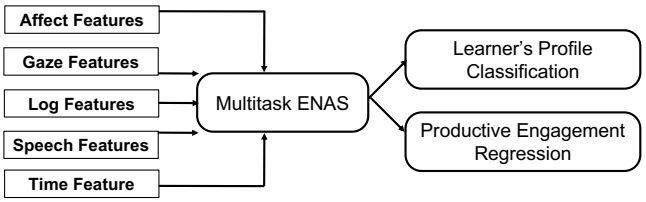

Figure 2: Multi-task learning personalization framework.

Multi-task Learning Personalization: Multi-task learning personalization entails using multi-task ENAS to learn an adaptive model for jointly predicting productive engagement and classifying learner profiles, thus implicitly using the learner profiles information to drive productive engagement model via weight sharing. Here, we used the AutoModel module of Auto-Keras. In the multi-task architecture, the output is composed of a regression head for predicting the PE score, and a classification head for classifying the three learner profiles. The best models were searched by employing network morphism operations such as inserting new layers, expanding existing layers, or adding skip connections. Figure 2 depicts the multi-task ENAS PE personalization framework.

Implementation Details – Similar to the single-task personalization framework, an 85 – 15% split strategy was used to divide the training dataset into training and validation sets. For classification, categorical cross entropy was used as the loss function and for the regression task, mean squared error was used as the loss function. The number of epochs was set to 100. The number of trials was set to 10.

5 EXPERIMENTS AND RESULTS

In this section, we describe the used evaluation metrics and present the evaluation results of the proposed approach, as compared to a non-personalised approach (i.e., baseline). Please note the test data for personalised and non-personalised models are the same.

5.1 Evaluation Metrics

The proposed productive engagement regression models are assessed using Pearson Correlation Coefficient (PCC) and Root Mean Square Error (RMSE). Let ES_{tr} and ES_{pr} be the true and predicted productive engagement scores, respectively, $RMSE(ES_{tr}, ES_{pr})$ is given by:

$$RMSE(ES_{tr}, ES_{pr}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(ES_{tr}(i) - ES_{pr}(i))^2} \quad (1)$$

where N is the number of samples in the test set.

In addition to PCC and RMSE, accuracy and Cross-Entropy loss (CE) are used to assess the classification task in the multi-task framework. Cross-Entropy loss (log loss) is a measure of the model's classification performance based on the output probability. CE increases with the divergence of the predicted probability from the true class label. The CE loss is computed by computing a separate

Table 1: Performance evaluation of personalized (P) and non-personalized (NP) models using incremental (INC) and non-incremental (NON INC) features. The results are reported in terms of RMSE and PCC as follows: RMSE (PCC*). (*) corresponds to statistically significant results (p-value ≤ 0.05). Gainers (G); Non-Gainers (NG). Bold: best performance among the features (vertically); Underline: best performance among the personalized and non-personalized models using the best performing features.

		INC		NON INC	
	Features	**P**	**NP**	**P**	**NP**
EE-SW (G vs. NG)	**All**	0.041 (0.264*)	0.05 (0.066*)	0.03 (0.408*)	0.041 (0.254*)
	Log	0.042 (0.336*)	0.042 (-0.151*)	0.03 (0.426*)	0.038 (0.028)
	Gaze	0.047 (0.099*)	0.049 (0.148*)	0.03 (0.402*)	0.04 (0.026)
	Affect	0.05 (0.283*)	0.045 (-0.033)	0.029 (0.410*)	0.041 (0.149*)
	Speech	**0.031 (0.363*)**	**0.038 (0.297*)**	**0.025 (0.500*)**	**0.037 (0.448*)**
CT-SW (G vs. NG)	**All**	0.079 (0.248*)	**0.139 (0.181*)**	0.038 (0.382*)	0.051 (0.185*)
	Log	0.035 (0.315*)	0.046 (0.012)	0.033 (0.397*)	0.043 (-0.008)
	Gaze	0.055 (0.239*)	0.055 (-0.074*)	0.031 (0.373*)	0.043 (-0.008)
	Affect	**0.06 (0.319*)**	0.041 (-0.064*)	0.033 (0.389*)	0.042 (0.194*)
	Speech	0.032 (0.303*)	0.038 (-0.171*)	**0.031 (0.494*)**	**0.037 (0.377*)**
EE-CT (G vs. G)	**All**	0.069 (0.063*)	**0.039 (0.284*)**	0.037 (0.162*)	0.03 (0.075*)
	Log	0.043 (0.055*)	0.04 (0.113*)	0.035 (0.101*)	0.036 (0.144*)
	Gaze	0.06 (-0.040*)	0.044 (0.039*)	0.034 (0.013)	0.035 (-0.086*)
	Affect	0.05 (0.027)	0.045 (-0.056*)	0.035 (0.067*)	0.034 (-0.043*)
	Speech	**0.035 (0.192*)**	0.031 (0.139*)	**0.032 (0.323*)**	**0.029 (0.413*)**
EE-CT-SW	**All**	0.062 (0.217*)	0.039 (0.194*)	0.035 (0.344*)	0.037 (0.217*)
	Log	0.041 (0.256*)	0.057 (-0.090*)	0.033 (0.355*)	0.052 (0.049*)
	Gaze	0.054 (0.115*)	0.059 (-0.004)	0.032 (0.319*)	0.04 (0.092*)
	Affect	0.044 (0.060*)	0.042 (0.079*)	0.032 (0.342*)	0.039 (0.054*)
	Speech	**0.033 (0.306*)**	**0.033 (0.352*)**	**0.029 (0.458*)**	**0.029 (0.400*)**

loss for each class label per observation and summing the result.

$$CE = -\sum_{c=1}^{M} y_{o,c} \log \left(p_{o,c}\right. \tag{2}$$

where M is the number of classes (three classes in our case, namely, CC, SW, and EE), log is the natural log, y is a binary indicator (0/1) of whether class label c is the correct classification label for observation o, and p is the predicted probability that observation o is of class c.

5.2 Performance Evaluation

We evaluate the proposed single-task and multi-task personalization frameworks using a leave-one-team-out strategy.

5.2.1 Profile-level Personalization. Table 1 presents the comparison of the proposed single-task personalized productive engagement approach to non-personalized (one-fits-all) approach for different feature modalities individually (Log, Gaze, Affect, Speech) as well as the early fusion of these modalities. For each of these experiments, we compare the performance using incremental and non-incremental features. The one-fits-all models are learned using the combined data of the different learners profiles as input to ENAS. The personalized models are learned using each learner profile data separately as input to ENAS. Different personalized and one-fits-all models are trained and compared by (1) taking all the three profiles into account, and (2) eliminating one of the profiles from the data. This resulted in four types of experiments:

(1) EE-SW: The Calm Tinkerers profile is excluded from the data used to learn the models. This constitutes models that consider learner profiles of high learning gain versus low learning gain.
(2) EE-CT: The Silent Wanderers profile is excluded from the data used to learn the models. This constitutes models that consider two types of learner profiles exhibiting high learning gain, with differences in the behavioural manifestation of such gain.
(3) CT-SW: The Expressive Explorers profile is excluded from the data used to learn the models. This constitutes models that consider learner profiles of high learning gain versus low learning gain.
(4) EE-CT-SW: All the learners profiles data are used to learn the models.

From the table, we can notice that when the models account for differences among gainers and non-gainers learning profiles, personalized models outperform non-personalized models using both the incremental and the non-incremental features, and for all the unimodal and multimodal features. With the EE-SW profiles, the models trained on the speech modality outperform the other

modalities and the fusion of all modalities (PCC = 0.363 [P, Speech] vs. 0.297 [NP, Speech] with the incremental speech features; PCC = 0.5 [P, Speech] vs. 0.448 [NP, Speech] with the non-incremental speech features). With the CT-SW profiles, the affect modality outperforms the rest using the incremental features (PCC = 0.319 [P, Affect] vs. 0.181 [NP, All]), while the speech modality outperforms the rest with the non-incremental features (PCC = 0.494 [P, Speech] vs. 0.377 [NP, Speech]).

When the models account for differences among the gainers profiles (EE-CT), the non-personalized models perform better than the personalized ones for both the incremental and the non-incremental features. Except for the non-personalized model trained on incremental features where the multimodal model performs the best in the two-by-two gainers profile, the speech features perform the best for all the other models (PCC = 0.192 [P, Speech] vs. PCC = 0.284 [NP, All] with the incremental speech features; PCC = 0.323 [P, Speech] vs. PCC = 0.413 [NP, Speech] with the non-incremental speech features).

When the models account for differences among the three learner profiles, personalized models outperform non-personalized models using the non-incremental features for all the unimodal and multimodal features. The speech modality performs the best (PCC = 0.458 [P, Speech] vs. 0.400 [NP, Speech]). In the case of incremental features, personalized models perform better except for the affect and speech modalities. The best performing modality is again the speech (PCC = 0.352 [NP, Speech] vs. 0.306 [P, Speech]).

Overall, we can observe that non-incremental features perform better than incremental ones. This could be due to the nature of the incremental features: as the time evolves, the average of each feature at time *t* becomes very similar to the average at *t-1*; hence, giving us almost identical data points associated with unique PE score values. This similarity between the data points can then lead to more errors in prediction. It is also observed that the speech modality outperforms the other modalities. This is not surprising as speech modality was also found to be most discriminant modality between gainers and non-gainers by Nasir et al. [27]. Additionally, the PE score is generated with speech based features (see [24] for details), albeit these speech features not included in our feature set here, that may explain the results we observe with speech modality. On taking a closer look into the results, we can conclude that models trained on data of gainers vs. non-gainers benefit from personalization while models trained on data of type 1 gainers vs. type 2 gainers show an opposite trend. In the case of models trained on data including all profiles, personalization provides a better outcome when compared to non-personalized analysis on average.

In addition to the overall model evaluation presented in Table 1, we report the results of each personalized learner profile-wise model on the corresponding profile, i.e., the results of training on the data corresponding to a single learner profile and predicting the engagement score of the teams belonging to the learner profile in question. These profile-wise results are reported in Table 2. The main aim is to further examine the performance of the personalized models on the different profiles. This also allows us to compare the performance of the incremental and non-incremental for predicting the productive engagement scores for each profile. In all the three profiles, the speech modality performs the best (EE: PCC = 0.309 [Non-inc, Speech] vs. 0.272 [Inc, Speech], CT: PCC = 0.341 [Non-inc, Speech] vs. 0.106 [Inc, Speech], SW: PCC = 0.304 [Inc, Speech] vs. 0.285 [Non-inc, Speech]).

Table 2: Performance evaluation of each personalized learner profile-wise model on the corresponding profile, i.e. the results of training on the data corresponding to a single learner profile and predicting the engagement score of the teams belonging to the learner profile in question. The results are reported in terms of RMSE and PCC as follows: RMSE (PCC*). (*) corresponds to statistically significant results (p-value ≤ 0.05).

Profile	Features	INC	NON INC
EE	**All**	0.042 (0.039)	0.032 (0.148)
	Log	0.047 (0.069*)	0.032 (0.126*)
	Gaze	0.054 (-0.145*)	0.033 (0.055*)
	Affect	0.044 (0.060*)	0.032 (0.080*)
	Speech	**0.034 (0.272*)**	**0.028 (0.309*)**
CT	**All**	0.104 (0.081*)	0.045 (0.174*)
	Log	0.037 (0.026)	0.038 (0.051*)
	Gaze	**0.068 (0.106*)**	0.036 (-0.085*)
	Affect	0.057 (-0.102*)	0.039 (0.030)
	Speech	0.036 (0.102*)	**0.037 (0.341*)**
SW	**All**	0.037 (0.236*)	0.026 (0.189*)
	Log	0.032 (0.263*)	0.024 (0.157*)
	Gaze	0.032 (0.172*)	0.022 (0.252*)
	Affect	0.063 (0.207*)	0.022 (0.214*)
	Speech	**0.025 (0.304*)**	**0.02 (0.285*)**

For the gainer groups, the non-incremental features outperforms incremental features whereas for the non-gainer group, incremental features show a better performance. These outcomes are coherent with the previously discussed results and thus can also be explained by the reasoning's given above.

5.2.2 Multi-task Learning Personalization. Via the multitask learning personalization framework, we aim to investigate whether the joint classification of learner profiles and the prediction of productive engagement scores can lead to an implicit personalization of productive engagement through weight sharing. Since we have four kinds of data instances (log, gaze, affect, and speech) each conveying a different form of information, we followed a multi-modal approach.

Within the multi-modal approach, we performed two main tests: 1) Taking the four features as four separate inputs and discarding the time feature in every input (TI-MT-NAS), and 2) Taking the four features as four separate inputs after removing time from each input and taking time as the fifth input (TD-MT-NAS). The main aim of these two experiments is to (i) compare and check whether the time modality is adding any bias to the predictions of the PE score, and (ii) check whether the prediction of user learner profiles labels influences the prediction of PE score.

Table 3 presents the results of the multi-task experiments. Since we have performed both classification and regression, we report two types of losses: RMSE loss for regression and Cross-Entropy

Table 3: Multi-task results. The results are reported in terms of RMSE and PCC for productive engagement prediction. Accuracy and classification loss (cross-entropy) are reported for the classification of learner profiles. (*) corresponds to statistically significant results (p-value ≤ 0.05). Time-Independent (TI), Time-Dependent (TD).

	TI-MT-NAS		TD-MT-NAS		EE-CT-SW (NP) All features	
	NON-INC	INC	NON-INC	INC	NON-INC	INC
MSE	**0.038**	0.104	**0.036**	0.099	0.037	0.039
PCC	**0.232***	0.035*	**0.266***	0.022	0.217*	0.194*
Class. Acc.	**0.53**	0.465	**0.512**	0.448	––	––
Class. Loss	**1.047**	2.093	**1.066**	2.451	––	––

loss for classification. From these results, we can observe that in the time-independent experiment, the non-incremental features are performing better than the incremental ones (PCC = 0.232 vs. 0.035). For the five modalities experiment, the non-incremental is performing better than incremental (PCC = 0.266 vs. 0.022). Concerning the classification accuracy of the learner profiles, we observe that for the TI-MT-NAS experiment, the non-incremental features perform better than the incremental ones (ACC = 0.53 vs. 0.465). The same pattern is noted for the TD-MT-NAS experiment (ACC = 0.512 vs. 0.448).

On comparing the two sets of experiments together, we can see that the five modalities experiment is performing better than the time-independent case in terms of PCC and performs less better in terms of classification accuracy. But there is no significant difference between the two. This indicates that time does not add any specific bias to the results. When we compare the MT results to the non-personalized three-by-three analysis for all features, we can see that MT performs better compared to the latter. Similar to the past two experiment sets, the trend that the non-incremental features are outperforming the incremental ones is noted here too.

6 CONCLUSION

In this paper, we propose and compare personalized models for Productive Engagement (PE) recognition in the context of robot-mediated learning scenario. Three learner profiles identified via a clustering technique from multimodal behaviours are used within an AutoML deep learning framework to personalize productive engagement models. We investigate two approaches for this purpose: (1) Single-task Deep Neural Architecture Search (NAS) (ST-NAS), and (2) Multitask NAS (MT-NAS).

In the former approach, personalized models for each learner profile are trained and compared to non-personalized models. Moreover, we analyze the performance by training two-by-two personalized models (EE-CT, EE-SW, and CT-SW) in order to underpin whether some profiles add noise to the training. We notice that personalization outperforms non-personalized models in the case of three-by-three models and for the two-by-two models in case of the gainer-non-gainer combination but shows the contrary for gainer-gainer groups. Our experimental results show that personalized models improve the recognition performance with respect to non-personalized models. The speech modality is the most informative feature in the prediction of the PE score.

In the MT-NAS approach, we investigate whether jointly classifying the learners' profiles with the engagement score through multi-task learning would serve as an implicit personalization of the productive engagement. These set of experiments prove that the time modality does not add any bias to the prediction of PE score. In both these approaches (ST-NAS and MT-NAS), it is noted that the non-incremental features performs much better than the incremental ones.

As a future work, it might be interesting to compare the performance with other personalisation strategies based on SVMs, such as Selective Transfer Machines [Chu et al., PAMI 2017].

ACKNOWLEDGMENTS

This research was supported by NYUAD internal grant and by the Center of AI & Robotics (CAIR) grant. The work of Jauwairia Nasir has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 765955. The work of Oya Celiktutan was supported by the LISI project, funded by the UKRI EPSRC (Grant Ref.: EP/V010875/1).

REFERENCES

[1] Nese Alyuz, Eda Okur, Ece Oktay, Utku Genc, Sinem Aslan, Sinem Emine Mete, Bert Arnrich, and Asli Arslan Esme. 2016. Semi-supervised model personalization for improved detection of learner's emotional engagement. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 100–107.

[2] Nese Alyuz, Eda Okur, Ece Oktay, Utku Genc, Sinem Aslan, Sinem Emine Mete, David Stanhill, Bert Arnrich, and Asli Arslan Esme. 2016. Towards an emotional engagement model: Can affective states of a learner be automatically detected in a 1: 1 learning scenario. In *Proceedings of the 6th Workshop on Personalization Approaches in Learning Environments (PALE 2016)*. 24th conference on User Modeling, Adaptation, and Personalization (UMAP 2016), CEUR workshop proceedings, this volume.

[3] Prakhar Bhardwaj, PK Gupta, Harsh Panwar, Mohammad Khubeb Siddiqui, Ruben Morales-Menendez, and Anubha Bhaik. 2021. Application of Deep Learning on Student Engagement in e-learning environments. *Computers & Electrical Engineering* 93 (2021), 107277.

[4] Hanlin Chen, Baochang Zhang, Xiawu Zheng, Jianzhuang Liu, Rongrong Ji, David Doermann, Guodong Guo, et al. 2021. Binarized neural architecture search for efficient object recognition. *International Journal of Computer Vision* 129, 2 (2021), 501–516.

[5] Jerusha O Conner. 2009. Student engagement in an independent research project: The influence of cohort culture. *Journal of Advanced Academics* 21, 1 (2009), 8–38.

[6] Diana Cordova and Mark Lepper. 1996. Intrinsic Motivation and the Process of Learning: Beneficial Effects of Contextualization, Personalization, and Choice. *Journal of Educational Psychology* 88 (12 1996), 715–730. https://doi.org/10.1037/0022-0663.88.4.715

[7] Francesco Del Duchetto, Paul Baxter, and Marc Hanheide. 2020. Are You Still With Me? Continuous Engagement Assessment From a Robot's Point of View. *Frontiers in Robotics and AI* 7 (2020), 116. https://doi.org/10.3389/frobt.2020.00116

[8] Abhinav Dhall, Amanjot Kaur, Roland Goecke, and Tom Gedeon. 2018. EmotiW 2018: Audio-Video, Student Engagement and Group-Level Affect Prediction. *CoRR* abs/1808.07773 (2018). http://arxiv.org/abs/1808.07773

[9] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. 2019. Neural architecture search: A survey. *The Journal of Machine Learning Research* 20, 1 (2019), 1997–2017.

[10] Michael Feffer, Rosalind W Picard, et al. 2018. A mixture of personalized experts for human affect estimation. In *International conference on machine learning and data mining in pattern recognition*. Springer, 316–330.

[11] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Tobias Springenberg, Manuel Blum, and Frank Hutter. 2019. Auto-sklearn: efficient and robust automated machine learning. In *Automated Machine Learning*. Springer, Cham, 113–134.

[12] Abhay Gupta, Richik Jaiswal, Sagar Adhikari, and Vineeth Balasubramanian. 2016. DAISEE: Dataset for Affective States in E-Learning Environments. *CoRR* abs/1609.01885 (2016). arXiv:1609.01885 http://arxiv.org/abs/1609.01885

[13] Natasha Jaques, Sara Taylor, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. 2016. Multi-task learning for predicting health, stress, and happiness. In *NIPS Workshop on Machine Learning for Healthcare*.

[14] Haifeng Jin, Qingquan Song, and Xia Hu. 2019. Auto-Keras: An Efficient Neural Architecture Search System. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 1946–1956.

[15] Wafa Johal. 2020. Research Trends in Social Robots for Learning. *Current Robotics Reports* 1 (2020), 1–9. https://doi.org/10.1007/s43154-020-00008-3

[16] A. Kamath, A. Biswas, and V. Balasubramanian. 2016. A crowdsourced approach to student engagement recognition in e-learning environments. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1–9. https://doi.org/10.1109/WACV.2016.7477618

[17] Manu Kapur. 2011. Temporality matters: Advancing a method for analyzing problem-solving processes in a computer-supported collaborative environment. *International Journal of Computer-Supported Collaborative Learning* 6, 1 (2011), 39–56.

[18] Dan Leyzberg, Samuel Spaulding, and Brian Scassellati. 2014. Personalizing robot tutors to individuals' learning differences. In *ACM/IEEE International Conference on Human-Robot Interaction*. 423–430. https://doi.org/10.1145/2559636.2559671

[19] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. 2018. Progressive neural architecture search. In *Proceedings of the European conference on computer vision (ECCV)*. 19–34.

[20] Dianbo Liu, Peng Fengjiao, Rosalind Picard, et al. 2017. DeepFaceLIFT: interpretable personalized models for automatic estimation of self-reported pain. In *IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing*. PMLR, 1–16.

[21] Hao Lu and Hu Han. 2021. NAS-HR: Neural architecture search for heart rate estimation from face videos. *Virtual Reality & Intelligent Hardware* 3, 1 (2021), 33–42.

[22] Aamir Mustafa, Amanjot Kaur, Love Mehta, and Abhinav Dhall. 2018. Prediction and Localization of Student Engagement in the Wild. *CoRR* abs/1804.00858 (2018). arXiv:1804.00858 http://arxiv.org/abs/1804.00858

[23] Jauwairia Nasir, Barbara Bruno, Mohamed Chetouani, and Pierre Dillenbourg. 2021. What if Social Robots Look for Productive Engagement? *International Journal of Social Robotics* (2021), 1–17.

[24] Jauwairia Nasir, Barbara Bruno, Mohamed Chetouani, and Pierre Dillenbourg. 2022. A Speech-based Productive Engagement Metric for Real-time Human-Robot Interaction in Collaborative Educational Contexts. *IEEE Transactions on Affective Computing* (2022). http://infoscience.epfl.ch/record/294035

[25] Jauwairia Nasir, Barbara Bruno, and Pierre Dillenbourg. 2020. Is There 'ONE Way' of Learning? A Data-Driven Approach. In *Companion Publication of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20 Companion)*. Association for Computing Machinery, New York, NY, USA, 388–391. https://doi.org/10.1145/3395035.3425200

[26] Jauwairia Nasir, Barbara Bruno, and Pierre Dillenbourg. 2021. *PE-HRI-temporal: A Multimodal Temporal Dataset in a robot mediated Collaborative Educational Setting*. https://doi.org/10.5281/zenodo.5576058

[27] Jauwairia Nasir, Aditi Kothiyal, Barbara Bruno, and Pierre Dillenbourg. 2022. Many are the ways to learn identifying multi-modal behavioral profiles of collaborative learning in constructivist activities. *International Journal of Computer-Supported Collaborative Learning* (2022), 1–39.

[28] Jauwairia Nasir, Utku Norman, Barbara Bruno, and Pierre Dillenbourg. 2020. When positive perception of the robot has no effect on learning. In *2020 29th IEEE international conference on robot and human interactive communication (RO-MAN)*. IEEE, 313–320.

[29] Catharine Oertel, Ginevra Castellano, Mohamed Chetouani, Jauwairia Nasir, Mohammad Obaid, Catherine Pelachaud, and Christopher Peters. 2020. Engagement in human-agent interaction: An overview. *Frontiers in Robotics and AI* 7 (2020), 92.

[30] Phuong Pham and Jingtao Wang. 2016. Adaptive Review for Mobile MOOC Learning via Implicit Physiological Signal Sensing. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (Tokyo, Japan) *(ICMI 2016)*. ACM, New York, NY, USA, 37–44. https://doi.org/10.1145/2993148.2993197

[31] A Ramachandran, C.-M. Huang, and B Scassellati. 2017. Give Me a Break!: Personalized Timing Strategies to Promote Learning in Robot-Child Tutoring. *12th Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI 2017* Part F1271 (2017), 146–155. https://doi.org/10.1145/a2909824.3020209

[32] Peter Reimann. 2009. Time is precious: Variable-and event-centred approaches to process analysis in CSCL research. *International Journal of Computer-Supported Collaborative Learning* 4, 3 (2009), 239–257.

[33] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. 2020. A comprehensive survey of neural architecture search: Challenges and solutions. *arXiv preprint arXiv:2006.02903* (2020).

[34] Alessandra Rossi, Mario Raiano, and Silvia Rossi. 2021. Affective, Cognitive and Behavioural Engagement Detection for Human-robot Interaction in a Bartending Scenario. In *2021 30th IEEE International Conference on Robot Human Interactive Communication (RO-MAN)*. 208–213. https://doi.org/10.1109/RO-MAN50785.2021.9515435

[35] Ognjen Rudovic, Jaeryoung Lee, Miles Dai, Björn Schuller, and Rosalind W Picard. 2018. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics* 3, 19 (2018).

[36] Ognjen Rudovic, Yuria Utsumi, Jaeryoung Lee, Javier Hernandez, Eduardo Castelló Ferrer, Björn Schuller, and Rosalind W Picard. 2018. Culturenet: A deep learning approach for engagement intensity estimation from face images of children with autism. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 339–346.

[37] Hanan Salam, Oya Celiktutan, Isabelle Hupont, Hatice Gunes, and Mohamed Chetouani. 2016. Fully automatic analysis of engagement and its relationship to personality in human-robot interactions. *IEEE Access* 5 (2016), 705–721.

[38] Hanan Salam and Mohamed Chetouani. 2015. Engagement detection based on mutli-party cues for human robot interaction. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 341–347.

[39] Hanan Salam, Viswonathan Manoranjan, Jian Jiang, and Oya Celiktutan. 2022. Learning Personalised Models for Automatic Self-Reported Personality Recognition. In *Understanding Social Behavior in Dyadic and Small Group Interactions*. PMLR, 53–73.

[40] Mostafa Shahabinejad, Yang Wang, Yuanhao Yu, Jin Tang, and Jiani Li. 2021. Toward Personalized Emotion Recognition: A Face Recognition Based Attention Method for Facial Emotion Recognition. In *Proceedings of IEEE International Conference on Face & Gesture*.

[41] Zilong Shao, Siyang Song, Shashank Jaiswal, Linlin Shen, Michel Valstar, and Hatice Gunes. 2021. Personality Recognition by Modelling Person-specific Cognitive Processes using Graph Representation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 357–366.

[42] Sara Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. 2017. Personalized multitask learning for predicting tomorrow's mood, stress, and health. *IEEE Transactions on Affective Computing* 11, 2 (2017), 200–213.

[43] Lumin Xu, Yingda Guan, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. 2021. ViPNAS: Efficient Video Pose Estimation via Neural Architecture Search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16072–16081.

[44] Avi Zacherman and John Foubert. 2014. The relationship between engagement in cocurricular activities and academic performance: Exploring gender differences. *Journal of Student Affairs Research and Practice* 51, 2 (2014), 157–169.