Semi-supervised Model Personalization for Improved Detection of Learner's Emotional Engagement

Nese Alyuz[1], Eda Okur[1], Ece Oktay[1], Utku Genc[1], Sinem Aslan[1], Sinem Emine Mete[1], Bert Arnrich[2], Asli Arslan Esme[1]

[1]Intel Corporation, Turkey
{nese.alyuz.civitci, eda.okur, ece.oktay, utku.genc, sinem.aslan, sinem.mete, asli.arslan.esme}@intel.com

[2]Bogazici University, Turkey
bert.arnrich@boun.edu.tr

ABSTRACT

Affective states play a crucial role in learning. Existing Intelligent Tutoring Systems (ITSs) fail to track affective states of learners accurately. Without an accurate detection of such states, ITSs are limited in providing truly personalized learning experience. In our longitudinal research, we have been working towards developing an empathic autonomous 'tutor' closely monitoring students in real-time using multiple sources of data to understand their affective states corresponding to emotional engagement. We focus on detecting learning related states (i.e., 'Satisfied', 'Bored', and 'Confused'). We have collected 210 hours of data through authentic classroom pilots of 17 sessions. We collected information from two modalities: (1) appearance, which is collected from the camera, and (2) context-performance, that is derived from the content platform. The learning content of the content platform consists of two section types: (1) instructional where students watch instructional videos and (2) assessment where students solve exercise questions. Since there are individual differences in expressing affective states, the detection of emotional engagement needs to be customized for each individual. In this paper, we propose a hierarchical semi-supervised model adaptation method to achieve highly accurate emotional engagement detectors. In the initial calibration phase, a personalized context-performance classifier is obtained. In the online usage phase, the appearance classifier is automatically personalized using the labels generated by the context-performance model. The experimental results show that personalization enables performance improvement of our generic emotional engagement detectors. The proposed semi-supervised hierarchical personalization method result in 89.23% and 75.20% F1 measures for the instructional and assessment sections, respectively.

CCS Concepts

• **Human-centered computing→Empirical studies in HCI**
• **Applied computing→Learning management systems**

Keywords

Emotional engagement detection, adaptive learning, personalization, affective computing, intelligent tutoring systems.

1. INTRODUCTION

Educational systems should provide personalized learning ("accommodate-for-each") rather than a "one-size-fits-all" approach [1]. Intelligent Tutoring Systems (ITSs) are becoming popular due to their promising capabilities for personalizing learner experience [2]**.** ITSs allow (1) monitoring students' learning process by tracking their interactions with the content platform, (2) creating a learning profile for each student, and (3) providing real-time feedback for any learning difficulties [3], [4]**.** However, existing ITSs fail to track affective states of learners very accurately. Without an accurate detection on such states, ITSs are limited in providing truly personalized learning experience.

In the related literature, there are a limited number of studies focusing on detecting affective states of learners [5] to realize Affective Tutoring Systems [6]. For example, in [7], the binary classification problem of whether a student was *interested* or not while interacting with learning activities was investigated. In [8], [9], and [10], automatic recognition of *frustration* was investigated. In [11], students' postures are used to track *boredom* and *flow*. In [12], affective states of *boredom*, *confusion*, *frustration*, *delight*, and *engagement* were researched. In [13], affective state detection was investigated for detecting *confidence*, *frustration*, *excitement*, and *interest*. The majority of these studies focus on generic affective state models. However, individuals differ in their emotional experience depending on their personal characteristics (e.g. gender, age) and personality traits [14], and their reactions are moderated by their personality: Individuals scoring high in neuroticism, are more likely to experience negative emotions and to view the world negatively [15]. On the other hand, individuals scoring high in extroversion experience more positive emotions [16]. As individual differences play an important role in affect recognition, AI models should embrace more personalized approaches. In order to target individual differences, person-dependent modeling is usually applied with varying portions of individual data [17]: (1) An available generic model is refined by adding individual training data, or (2) only individual data is used to create the model. The first method allows to capture general tendencies with the generic model and to fine-tune individual effects with the personal data. In the second method, a truly personal model can be achieved; but it has the drawback of requiring large amount of labeled training data for each new person.

In our longitudinal research, we have been working towards developing an empathic autonomous 'tutor' closely monitoring students in real-time using multiple sources of data to understand their emotional engagement. We are currently using two different modalities to infer a learner's emotional engagement: (1) Appearance modality, where a camera is utilized to capture visual information; and (2) context-performance modality, where context

ICMI'16, November 12–16, 2016, Tokyo, Japan
© 2016 ACM. 978-1-4503-4556-9/16/11...$15.00
http://dx.doi.org/10.1145/2993148.2993166

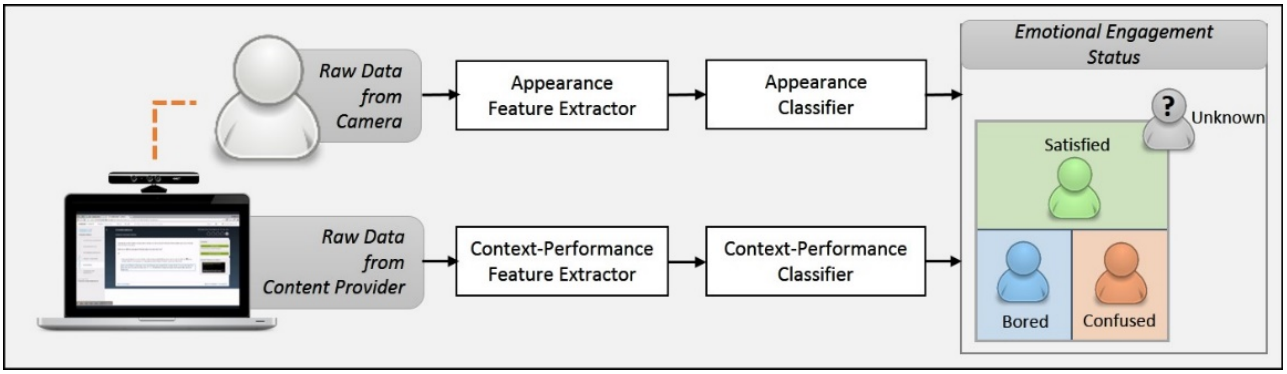

Raw Data from Camera
Appearance Feature Extractor
Appearance Classifier
Emotional Engagement Status
Unknown
Satisfied
Raw Data from Content Provider
Context-Performance Feature Extractor
Context-Performance Classifier
Bored
Confused

Figure 1. Overall scheme of the generic emotional engagement detector.

and performance information is gathered from a content platform. The general scheme of the system is given in Figure 1: For each modality, we implemented a separate feature extractor, and then trained a separate generic supervised classifier, and the overall emotional engagement is defined using the two classifier outputs. In our research, motivated by the circumplex model [18], we focus on the affective states of 'Satisfied', 'Bored', and 'Confused' (Figure 1, right).

In this paper, we propose a semi-supervised hierarchical method to overcome the drawback of providing large amounts of training data for truly personalized models that recognize learners' emotional engagement. We observed that the context-performance classifier achieves high accuracies even with a limited amount of person-specific training set, whereas the appearance classifier needs a larger set of labeled data to achieve a high accuracy. We utilize this effect in our proposed method: We start with a calibration phase, which includes personalization of the context-performance classifier. The second level of the hierarchical method includes automatic personalization of the appearance model, where the personalized contextual classifier is utilized as the label predictor. We compare the performance of our semi-supervised hierarchical approach with generic models and fully personalized models.

This paper is organized as follows: In Section 2, the proposed method is explained in detail, whereas Section 3 summarizes the experimental results obtained. In Section 4, conclusions and future directions are outlined.

2. METHODOLOGY OVERVIEW

Our aim is to develop a multi-modal system that can detect a learner's emotional engagement. For improving the performance of our emotional engagement detector, we obtain personalized empirically shown (see Section 3.2 and 3.4 for results), the improvement achieved by personalization is evident for both models expressing each individual's characteristics: As of the modalities considered: For context-performance modality, the classifier achieves high accuracies with a limited amount of person-specific training data; whereas for the appearance modality, classifier more personal data is needed to achieve similar results. Motivated by these findings, in this paper, we propose a personalization approach that adapts the emotional engagement models in a hierarchical manner: First, the context-performance classifier is personalized in a calibration phase. The labeled personal data classifier is then used in the online-usage phase to automatically provide labels (as a one-sided co-training is used to improve the context-performance classifier through model personalization. The personalized context-performance approach [19]) for the personalization of the appearance classifier. After the calibration phase, no more labels are required. The different phases (calibration and online-usage), the corresponding training sets, and the output models are illustrated in Figure 2: At the beginning of the calibration phase, we have a generic context-performance classifier trained in an offline manner using the initial training set. This set is collected from different students. In the calibration phase, the context-performance classifier is personalized using labeled subject-specific data. In the online usage phase, subject-specific data is collected, and the personalized context-performance classifier acts as the automatic label generator. The automatically labeled subject-specific data is then used to personalize the appearance classifier. In the following subsections, further details about each modality, classifier, and personalization strategy is given.

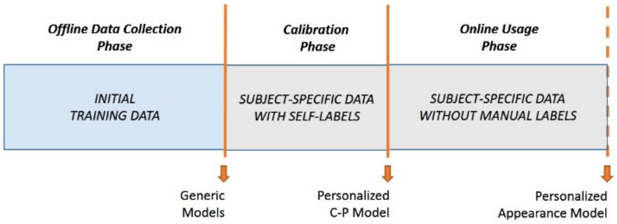

Offline Data Collection Phase
Calibration Phase
Online Usage Phase
INITIAL TRAINING DATA
SUBJECT-SPECIFIC DATA WITH SELF-LABELS
SUBJECT-SPECIFIC DATA WITHOUT MANUAL LABELS
Generic Models
Personalized C-P Model
Personalized Appearance Model

Figure 2. Timeline for the usage of different training sets and the corresponding output models.

2.1 Data Modalities & Feature Extraction

The learning content is provided by a content platform, in the form of two section types of *instructional* (where students watch instructional videos) and *assessment* (where students solve exercise questions). The computing device used for content retrieval is equipped with a 3D camera (i.e., Intel® RealSense™ F200 Camera). As visualized in Figure 1, we consider two sources of information as the two modalities (as in [20]): (1) Appearance modality which is acquired through the camera, and (2) context-performance modality which is gathered from the content platform and includes data related to the learning content or the profile of the learner. To extract features for each modality, the raw data are segmented into windows of 8-seconds length: In [21], various window sizes between 2-180 seconds were tested, and 8-seconds length was empirically found out to be suitable for the engagement detection problem. Therefore, in our further experiments we utilized windows of 8 seconds. Moreover, considering the continuous nature of the video data, we used a sliding window with an overlap of 4-secs.

2.1.1 Appearance Features

The videos of students were recorded with Intel® RealSense™ F200 Camera, and they include the RGB and depth streams of the student (including face and upper body). The frame-wise raw data are fed into the Intel® RealSense™ SDK [22] to extract face location and head position in the 3D space, 2D/3D positions of 78 facial landmarks, head pose, 22 facial expressions, and seven basic facial emotions. These are employed in the extraction of *segment-wise* features necessary for engagement detection. The extracted appearance features include various L-estimator statistical values (e.g. tri-mean of head velocity) and energy calculations (e.g. trend of pose energy), related to head position and pose, to facial expressions, and to seven basic emotions. The groupings of higher-level appearance features used in this paper are given in Table 1.

2.1.2 Context and Performance Features

Context features are extracted partly from the user profile and session information (i.e. gender, age, time of day), in addition to the data provided by the content platform (i.e. video duration, exercise/trial number, time within session). They are utilized for enabling context-awareness, and they are related to the educational content, to the environment, or to the student in general. They are present in both section types of *instructional* and *assessment*. The performance features are extracted from the user profile data containing user characteristics provided by the content platform. Note that the performance features are present only for the assessment sections; and they are related to grade, time spent, number of trials, or number of hints taken for a question. Since contextual and performance features are extracted using the same sources of information (i.e., the content platform), we employed data fusion at feature level and obtained a single context-performance feature set. The groupings of context-performance features are given in Table 1, together with feature counts and some exemplary features.

Table 1. Appearance (Appr) and context-performance (C-P) feature subgroups and corresponding feature counts.

Appr Feature Groups (Counts)	Examples
Tracking ratio (2)	Position and pose tracking
Head position / pose (128)	trend of pose energy, standard deviation of head position, etc.
Facial expressions (32)	Number of eye raisers per segment, mean of smile, etc.
Seven basic emotions (28)	Mean of anger intensity, number of joyful segments etc.

C-P Feature Groups (Counts)	Examples
Time related (6)	Time from beginning, video/attempt duration, etc.
Trial related (3)	Trial number, number of trials until success, etc.
Hint related (5)	Number of hints used on attempt, on question, etc.
Grade related (7)	Grade, correct attempt percentage, etc.
Other (3)	Gender, question number from beginning, etc.

2.2 Uni-modal Classification

As seen in Figure 1, the outputs of modality-specific feature extractors are fed into the corresponding uni-modal classifiers. As uni-modal classifier for both modalities, we utilized the Random Forest classification method [23]: In the Random Forest algorithm, a multiple number of decision trees are trained using random sets of training data and features. During testing, the test sample is predicted with all trained decision trees. Multiple decisions from individual trees are then analyzed with majority voting to generate a final prediction. Considering the advantages of Random Forests (e.g., no overtraining risk, no need for cross validation), we trained two separate forest with 100 trees each: (1) Appearance classifier, and (2) context-performance classifier.

2.3 Confidence Calibration

During the calibration phase, our generic context-performance classifier needs to separate samples with low confidence to decide when to request self-labels. Moreover, during the online usage phase, our personalized context-performance classifier needs to provide labels with high confidence since they are used as auto-labels for personalizing the appearance classifier. Therefore, having well-calibrated confidence values is crucial for our proposed personalization strategy. For each test sample, the Random Forest classifier outputs a final prediction, achieved by applying majority voting over 100 trees; and a probability score, calculated as the ratio of trees with the output of the majority vote. This probability score can be taken as a measure of confidence in the prediction, because samples with more trees agreeing on the same prediction are more likely to be classified correctly. However, these probabilities are usually not well-calibrated and can be statistically unreliable [24]. In the literature, calibration procedures for Random Forests, and in general for machine learning, have been examined in previous studies [24], [25], [26]. We employed one of the most popular calibration methods known as isotonic regression [26]: In this method, a general form of binning is suggested such that no specific number of bins or limits for the bin size is required. For further details of this algorithm, see [26].

2.4 Uni-modal Personalization

As shown in [27] and [17], person-specific models achieve significant improvement over person-independent models for the emotion detection problem. In this paper, it is empirically shown that this also holds for the detection of learner's emotional engagement states (see Section 3.4). In our research, generally we envision to achieve personalized models using active learning strategies: At randomized time points, the student will be asked to provide his/her current affective state through giving self-labels. At the end of each session, the provided labels and the corresponding features will be added to the training data and retraining will take place to yield a more person-specific model after each session. For model personalization, the labeled person-specific data can either be added to the training set of the generic model (i.e., 'Adapted'), or it can be utilized alone (i.e., 'Personal'). For the experiments in this paper, we experimented with both the 'Adapted' and the 'Personal' approach. We utilized *ground truth* labels to investigate the improvements we can achieve by personalization strategies. For these experiments, we employed all subject-specific training samples to see the upper limit for the performance of such a personalized model.

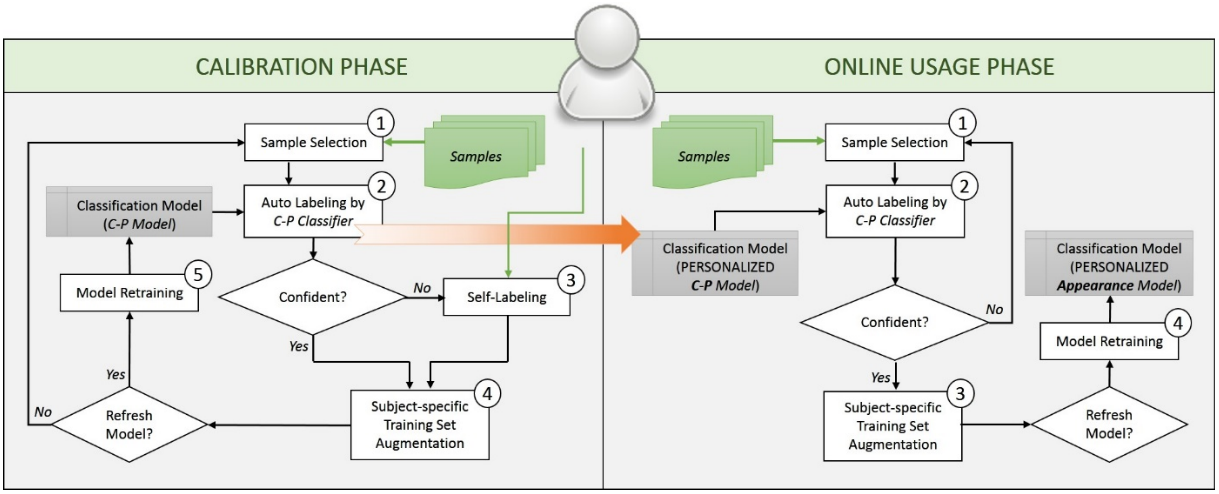
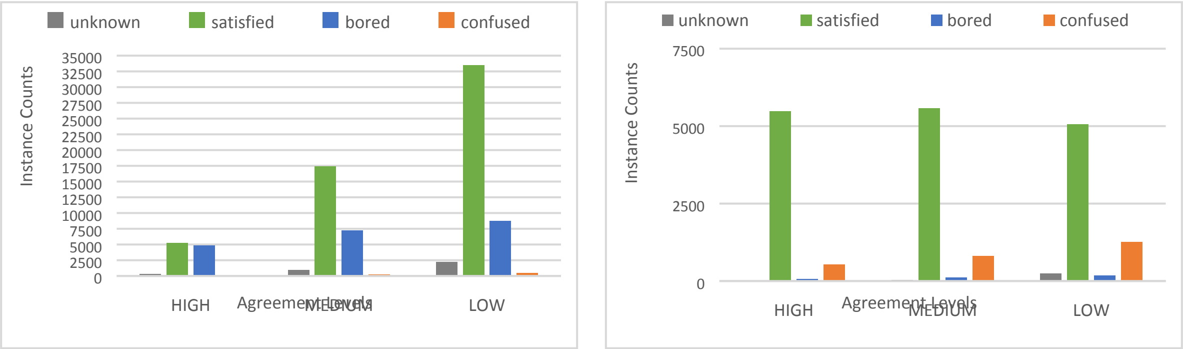


Figure 3. Calibration and online usage phases for the proposed hierarchical personalization approach.



(a) Instructional

(b) Assessment

Figure 4. Distribution of samples for different agreement levels of (1) High (5/5), (2) Medium (4/5), and (3) Low (3/5), for (a) Instructional, and for (b) Assessment sections.

2.5 Hierarchical Personalization

Modality-specific models can be personalized using active learning strategies. Although both modalities benefit from personalization (see Section 3.4), context-performance converges with very limited data, whereas the appearance classifier requires more labeled data to reach the accuracy levels of the context-performance. Motivated by the importance of the appearance modality, which will be available even when no specific content platform is utilized (e.g., the student is watching any video or reading any article on the web), we propose a hierarchical personalization approach. The overall scheme of the proposed approach is given in Figure 3: First, during a calibration phase, the labels are predicted in real time using the generic context-performance classifier. For random predictions with low confidence, labels are requested. The confidently predicted samples and the self-labeled instances are utilized to augment the training set with subject-specific features and the corresponding labels. From time to time (i.e., end of each session), the context-performance classifier is retrained to obtain improved models. In the online usage phase, the personalized context-performance classifier is utilized to generate labels necessary for appearance model personalization. Once again, the confidence of the context-performance classifier is assessed to choose confidently predicted samples. Employing both the self-labeled data of the calibration phase and the automatically labeled data during the online usage phase, the appearance models are retrained for personalization. With this strategy, the aim is to personalize the appearance model without intervening with the student (i.e., requesting self-labels).

Table 2. Engagement detection results (F1-measures) for *instructional* and *assessment* sections on Appearance and Context-Performance modalities, using: (1) the 'Generic', (2) 'Adapted', and (3) the 'Personal' models.

Section Type	Classes	AVERAGE TRAINING SIZE/MODEL			CONTEXT-PERFORMANCE CLASSIFIER (%)			APPEARANCE CLASSIFIER (%)		
		Generic	Adapted	Personal	Generic	Adapted	Personal	Generic	Adapted	Personal
INSTRUCTIONAL	Unknown	967	1018	51	9.62	72.97	85.38	10.73	24.85	33.04
	Satisfied	967	2272	1305	55.76	96.12	97.18	61.04	87.63	89.65
	Bored	967	1542	575	39.68	93.33	94.41	44.93	70.91	73.54
	OVERALL	2901	4832	1931	49.50	96.13	97.32	55.79	85.44	89.30
ASSESSMENT	Unknown	1886	2210	324	27.94	72.02	72.75	33.53	47.21	49.48
	Satisfied	1886	2883	997	76.32	94.04	94.39	60.58	83.43	83.79
	Confused	1886	2044	158	46.59	82.05	85.01	17.12	37.64	44.04
	OVERALL	5658	7137	1479	63.41	90.24	90.89	48.12	75.25	76.37

3. EXPERIMENTAL RESULTS

3.1 Data Collection and Labeling

We ran authentic classroom pilots with 9th grade students (age of 14-15) using an online math learning platform which provides *instructional* videos and *assessment* items. In total, data were collected from 20 students in 17 one-hour sessions to generate 210 hours of data. During each session, the videos of students was recorded with a 3D camera (i.e., Intel® RealSense™ Camera F200) and the context and performance logs were collected through the content platform. Since our aim was to investigate the personalization strategies and the amount of available personal data was important for the personalization experiments, we selected students who attended most of the sessions (i.e., twice a week), and carried out our experiments with the data from nine students.

For the supervised training phase of our models and the performance evaluation of our system, ground truth labels were necessary. Following the labeling methodology of [28], each recording was labeled by five different experts with a background in psychology or education, who defined segments based on observed state changes. Instances for feature extraction were defined as sliding windows of 8-seconds with overlaps of 4-seconds. As the Krippendorff's alpha computed among multiple labelers were low (0.4) [28], highlighting the subjective nature of the affective labeling task, final labels were assigned to each instance by applying majority voting together with validity filtering [20]: The ratio of majority votes was computed and the instances were grouped as of high (5/5), medium (4/5), or low (3/5) agreement. If there is no majority among labelers (i.e., the majority votes is below 3/5), the instance is labeled as instances of disagreement. In Figure 4, the data distributions of agreement levels are given. By examining these distributions, we decided to use high-to-medium agreement samples for the instructional. For the assessment, however, we chose high-low agreement samples due to limited sample counts, and included disagreement samples as instances of 'Unknown' class. Furthermore, we examined class-specific distributions for different section types, and we decided not to use 'Confused' class for the instructional, and 'Bored' class for the assessment sections due to too few samples.

3.2 Generic Classification Results

Table 2 provides a summary of the results produced by the generic model (see columns entitled with "Generic"): Discrete Random Forest classifiers are trained for each section type (instructional vs. assessment) and for each modality (appearance vs. context-performance). In order to reduce the effects of overfitting to each test subject, leave-one-subject-out cross-validation approach was applied: The training samples of all the other students were utilized to construct the training set of that test subject's classifiers. Moreover, for the generic models, we employed balanced sample sets for each class: From the training set, 10-fold random selection is applied to construct training sets and the averaged results are reported on test sets: As results, average F1 measures are reported to incorporate both the precision and recall metrics.

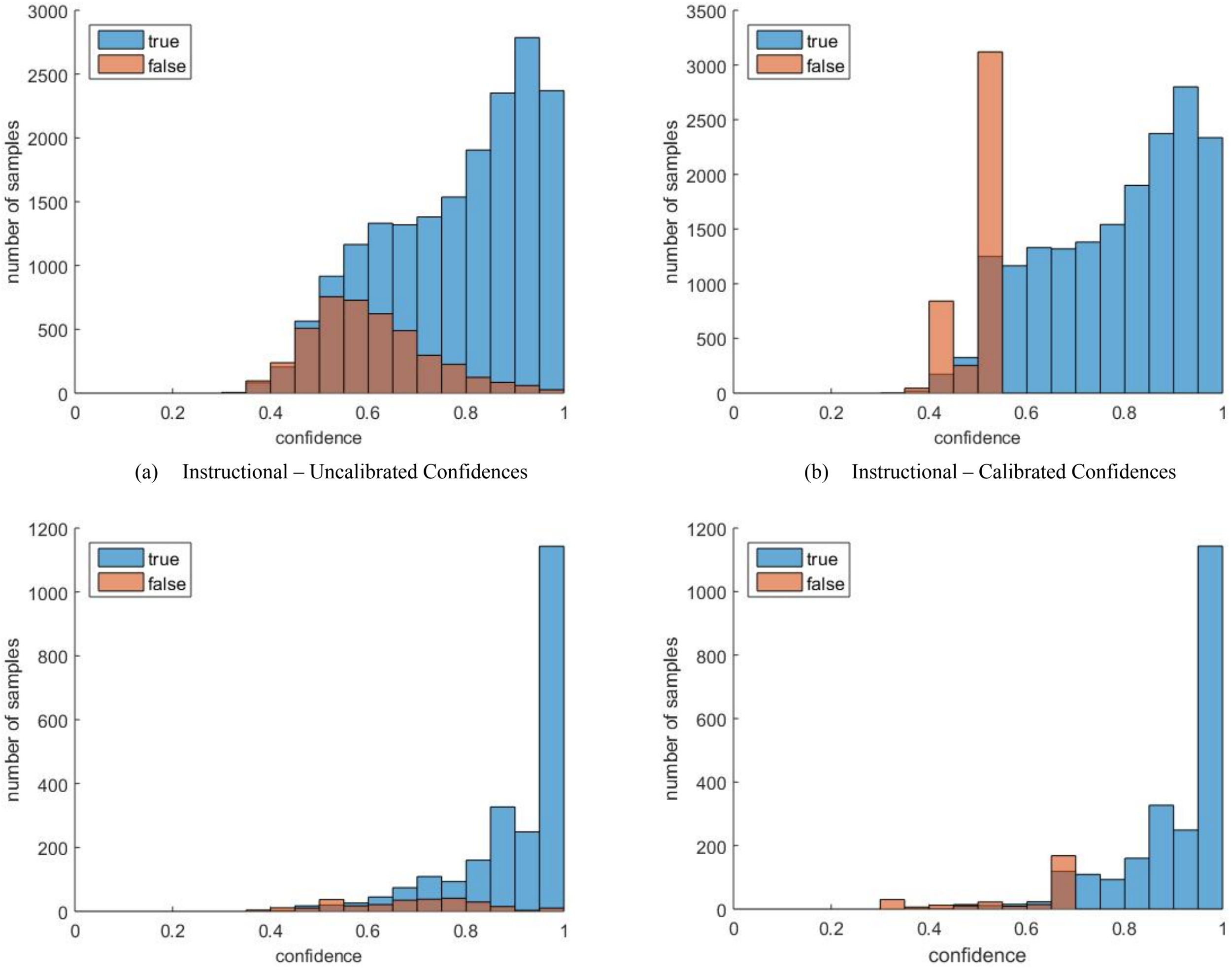
By examining the results, one can easily notice that context-performance classifier performance is low during instructional sessions (49.50%), since valuable performance-related features extractable for assessment sections are not present in addition to limited training sizes. For the assessment sections, improved results are reported for the context-performance modality (63.41%). When we compare the results for the two modalities, we see different trends: For the instructional sections, appearance classifier (55.79%) performs better than the context-performance classifier (49.50%). For the assessment, context-performance modality (63.41%) yields higher accuracies than the appearance (48.12%).

3.3 Confidence Calibration Results

In order to observe the effects of calibration on confidence scores, we plotted the histograms, i.e. the number of samples at each confidence value interval, both for the true and false predictions of our context-performance classifier. In Figure 4, confidence value distributions are presented for both instructional and assessment sections, where we can compare the results before (on the left) and after (on the right) the confidence calibration using isotonic regression method [25]: Without confidence calibration, it is visible that true and false predictions are not separable by employing thresholding over the confidence scores. However, after calibration, thresholding can be applied. In our dataset, we empirically estimated confidence thresholds on the training sets: Samples with confidence values above *0.55* and *0.70* can be assumed highly confident for instructional (Figure 4(b)) and assessment (Figure 4(d)) sections, respectively.

3.4 Uni-modal Personalization Results

In this paper, ground truth labels are employed when constructing the person-specific labeled sets for the personalization experiments.



(a) Instructional – Uncalibrated Confidences (b) Instructional – Calibrated Confidences

(c) Assessment – Uncalibrated Confidences (d) Assessment – Calibrated Confidences

Figure 4. Confidence distributions for the Context-Performance classifier with and without confidence calibration are given for different section types: (a), (b) for *instructional*; (c), (d) for *assessment*.

For model personalization, we experimented with two approaches: (1) ‘Adapted’, where the person-specific data is augmented to the training set of the generic model; and (2) ‘Personal’, where the person-specific data is used alone in the model training phase. These results mainly set the upper limits for the performance of our personalization approaches. In Table 2, the average number of training sizes (columns 3-5) and the averaged F1 measures for the context-performance (columns 6-8) and appearance (columns 9-11) modalities are given. As the overall results (last rows) show, both the adapted and the personal models achieve higher accuracies when compared to generic models, indicating that the information residing in both the context-performance and the appearance modalities are specific to each individual. Moreover, when personal results of different modalities are compared, it is seen that the improvement for the context-performance classifier (from 49.50% to 97.32% for instructional, and from 63.41% to 90.89% for assessment) is more evident and high accuracies can be achieved even with limited amount of personal data. For the appearance modality, with the same amount of personal data, lower accuracies can be achieved (89.30% for instructional, and 76.37% for assessment).

3.5 Hierarchical Personalization Results

Motivated by the highly performing personalized context-performance classifier, we proposed a model personalization approach for the appearance classifier which does not require self-labels: As explained in Section 2.5, the proposed hierarchical personalization approach utilizes the personalized context-performance classifier to automatically label data for the personalization of the appearance classifier. In the hierarchical approach, the highly confident samples labeled by the ‘Personal’ context-performance classifier are included in person-specific training sets. We have also experimented with a hybrid approach, where the unconfident classified samples are fed into the training phase using ground truth labels. This corresponds to requesting self-labels only for the samples that are unconfidently labeled by the personalized context-performance classifier. The results for the appearance modality, where three different personalization approaches are tested are given in Table 3: (1) ‘Hierarchical’, corresponding to the proposed hierarchical personalization scheme; (2) ‘Hybrid’, which incorporates ground-truth labels to the hierarchical approach for unconfident samples; and (3) ‘Full’, where ground-truth labels for all training samples are utilized to construct personal training sets and gives the upper limit for the performance. The ‘Hierarchical’, ‘Hybrid’, and ‘Full’ results are given both for the ‘Adapted’ and the ‘Personal’ approach, where the person-specific samples are either added to the generic sets or utilized alone for model training. As the results indicate, similar results are achieved with either the ‘Adapted’ or the ‘Personal’ approaches. When ‘Personal’ results are investigated to compare the proposed ‘Hierarchical’ method with the ‘Full’ models (setting the upper limits), it is visible that similar results are obtained: For the instructional sections, ‘Hierarchical’ models are obtained using 0.55 confidence threshold (i.e., 99% of samples are confidently labeled), and they achieve 89.23%. For the assessment sections, ‘Hierarchical’ models achieve 75.20% F1 measure using a threshold of 0.70 (i.e., 80% of samples labeled as confident). These results are very similar to the results achieved by the ‘Full’ models, where the ground truth labels are utilized: For instructional and assessment, 89.30% and 76.37% F1 measures are obtained, respectively. When the ‘Hybrid’ results for the ‘Personal’ approach are investigated, we can say that similar results are obtained by the ‘Hierarchical’ approach.

4. CONCLUSIONS AND FUTURE WORK

In this work, our aim is to infer the emotional engagement of a learner using two modalities, namely appearance and context-performance. In order to target individual differences, we propose a semi-supervised hierarchical method, which overcomes the drawback of providing large amounts of training data necessary for a truly personalized model: Since the context-performance classifier achieves high accuracies even with a limited amount of training data, in an initial calibration phase, we train a personalized context-performance classifier. Utilizing the personalized context-performance classifier as the label predictor during the online usage phase, we personalize the appearance model with automatically labeled person-specific data. In our experiments, we investigated two personalization strategies using the ground truth labels: (1) ‘Adapted’, where the initial training set is augmented with the person-specific data; and (2) ‘Personal’, where only the personal data is utilized in model retraining. Moreover, for a better understanding of how the engagement detector performs, we treated different section types of instructional and assessment separately. For the evaluation of the proposed hierarchical personalization approach, we compared results with those of the generic and the personalized models (‘Adapted’ and ‘Personal’).

For our experiments, we collected data from 9th grade students during a semester while they were using an online math learning platform. The collected data were labeled by experts and their multiple decisions were processed by majority voting and validity filtering to provide the ground truth labels for the affective states of ‘Satisfied’, ‘Bored’, and ‘Confused’.

In a leave-one-subject-out cross validation, the modality-specific generic models indicated that the appearance modality is more informative for the instructional sections (55.79% vs. 49.50% F1 measure), whereas for the assessment sections the context-performance modality becomes more representative (63.41% vs. 48.12% F1 measure). As the full personalization experiments showed (i.e., ‘Adapted’ and ‘Personal’), information included in both of the modalities are person-specific, thus model personalization helps to achieve high performance for emotional engagement detection. Even with the limited amount of data utilized in the ‘Personal’ strategy, context-performance classifier achieve F1 measures of 97.32% during instructional sessions and 90.89% during assessment sessions. For the appearance classifier, 89.30% and 76.37% F1 measures are achieved for instructional and assessment sections, respectively. These results for the fully personalized models indicate that the context-performance models can be personalized to act as the label predictor for the appearance modality, which needs a larger amount of labeled data to achieve high performance levels as the context-performance modality. We evaluated the proposed hierarchical personalization approach (‘Hierarchical’) and compared the results with the fully personalized models, which set the upper limits with the available data. We also experimented with the ‘Hybrid’ approach, using ground truth labels when the label predictor is not confident, and compared it with the results of the ‘Hierarchical’ approach. As the results indicated, with the proposed hierarchical personalization approach, the performance levels of the fully personalized models were achieved: 89.23% and 75.20% F1 measures are obtained for the instructional and the assessment sections, respectively. The results are similar for the ‘Hybrid’, indicating that using only the automatic labels generated by the context-performance model is sufficient.

For the final system, we envision to use self-labels to obtain labeled person-specific sets. We are currently designing a new data collection pilot, where self-labels are requested from the students at randomized time points. Therefore, we will be evaluating the performance of the hierarchical personalization approach using self-labels as proposed. Moreover, we are planning on including bio-sensors as an additional modality and understand whether any bio-sensor data can be beneficial for automatic label generation. The new pilot will enable us to conduct future experiments on an extended set of students. Moreover, on the extended database, we will be further investigating supervised and semi-supervised approaches for model personalization.

Table 3. Engagement detection results (F1-measures) for *instructional* and *assessment* sections on Appearance (Appr.) using different strategies: (1) ‘Generic’, (2) ‘Adapted-Hierarchical’, (3) ‘Adapted-Hybrid’, (4) ‘Adapted-Full’, (5) ‘Personal-Hierarchical’, (6) ‘Personal-Hybrid’, and (7) ‘Personal-Full’.

Section Type	Classes	GENERIC	ADAPTED			PERSONAL		
			Hierarchical	Hybrid	Full	Hierarchical	Hybrid	Full
INSTRUCTIONAL	Unknown	10.73	25.31	25.21	24.85	29.44	30.55	33.04
	Satisfied	61.04	87.78	87.77	87.63	89.34	89.23	89.65
	Bored	44.93	70.83	70.77	70.91	72.92	72.74	73.54
	OVERALL	55.79	85.48	85.56	85.44	89.23	89.06	89.30
ASSESSMENT	Unknown	33.53	43.52	47.85	47.21	42.25	49.07	49.48
	Satisfied	60.58	82.65	83.15	83.43	83.17	83.84	83.79
	Confused	17.12	33.92	38.16	37.64	38.88	44.86	44.04
	OVERALL	48.12	73.92	75.22	75.25	75.20	76.40	76.37

5. REFERENCES

[1] S. Aslan and C. M. Reigeluth, " A trip to the past and future of educational computing: Understanding its evolution," *Contemporary Educational Technology,* vol. 2, no. 1, pp. 1-17, 2011.

[2] M. Martinez, "Designing learning objects to personalize learning," *The Instructional Use of Learning Objects,* pp. 151-171, 2002.

[3] G. Paviotti, P. G. Rossi and D. Zarka, Intelligent tutoring systems: an overview, Pensa Multimedia, 2012.

[4] F. F. Burton, Foundations of Intelligent Tutoring Systems, 2013.

[5] B. Woolf, W. Burleson, I. Arroyo, T. Dragon, D. Cooper and R. Picard, "Affect-aware tutors: recognising and responding to student affect," *Int. Journal of Learning Technology,* vol. 4, no. 3, pp. 129-164, 2009.

[6] M. B. Ammar, M. Neji, A. M. Alimi and G. Gouardères, "The affective tutoring system," *Expert Systems and Applications,* vol. 37, no. 4, pp. 3013-3023, 2010.

[7] A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," in *Int. Conf. on Multimedia*, 2005.

[8] A. Kapoor, W. Burleson and R. W. Picard, "Automatic prediction of frustration," *Int. Journal of Human-Computer Studies,* vol. 65, no. 8, pp. 724-736, 2007.

[9] M. E. Hoque, D. J. McDuff and R. W. Picard, "Exploring temporal patterns in classifying frustrated and delighted smiles," *Transactions on Affective Computing,* vol. 65, no. 8, pp. 323-334, 2012.

[10] J. F. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe and J. C. Lester, "Automatically recognizing facial indicators of frustration: a learning-centric analysis," in *Affective Computing and Intelligent Interaction*, 2013.

[11] S. D'Mello, P. Chipman and A. Graesser, "Posture as a predictor of learner's affective engagement," in *Annual Cognitive Science Society*, 2007.

[12] N. Bosch, S. D'Mello, R. Baker, J. Ocumpaugh, V. Shute, M. Ventura and W. Zhao, "Automatic detection of learning-centered afective states in the wild," in *Int. Conf. on Intelligent User Interfaces*, 2015.

[13] I. Arroyo, D. G. Cooper, W. Burleson, B. P. Woolf, K. Muldner and R. Christopherson, "Emotion sensors go to school," *Artificial Intelligence in Education (AIED),* vol. 200, pp. 17-24, 2009.

[14] C. Kappeler-Setz, "Multimodal emotion and stress recognition (dissertation)," ETH, 2012.

[15] D. Watson and L. A. Clark, "Negative affectivity: the disposition to experience aversive emotional states," *Psychological Bulletin,* vol. 96, no. 3, pp. 465-490, 1984.

[16] R. E. Lucas and F. Fujita, "Factors Influencing the Relation Between Extraversion and Pleasant Affect," *Journal of Personality and Social Psychology,* vol. 79, no. 6, pp. 1039-1056, 2000.

[17] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic and K. Scherer, "The first facial expression recognition and analysis challenge," in *Int. Conf. on Automatic Face and Gesture Recognition and Workshops*, 2011.

[18] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology,* vol. 39, no. 6, p. 1161, 1980.

[19] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Annual Conf. on Computational Learning Theory*, 1998.

[20] N. Alyuz, E. Okur, E. Oktay, U. Genc, S. Aslan, S. E. Mete, D. Stanhill, B. Arnrich and A. A. Esme, "Towards an emotional engagement model: Can affective states of a learner be automatically detected in a 1:1 learning scenario?," in *ACM Conf. on User Modeling, Adaptation and Personalization (UMAP) - Workshops*, 2016.

[21] S. Aslan, Z. Cataltepe, I. Diner, O. Dundar, A. A. Esme, R. Ferens, G. Kamhi, E. Oktay, C. Soysal and M. Yener, "Learner Engagement Measurement and Classification in 1: 1 Learning," in *Int. Conf. on Machine Learning and Applications*, 2014.

[22] Intel Corporation, "Intel RealSense SDK: Design Guidelines," 2014: https://software.intel.com/sites/default/files/managed/27/50/Intel%20RealSense%20SDK%20Design%20Guidelines%20F200%20v2.pdf.

[23] C. Chen, A. Liaw and L. Breiman, Using random forest to learn imbalanced data, University of California, Berkeley, 2004.

[24] H. Boström, "Calibrating random forests," in *Int. Conf. on Machine Learning and Applications*, 2008.

[25] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in Large Margin Classifiers,* vol. 10, no. 3, pp. 61-74, 1999.

[26] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Int. Conf. on Machine Learning*, 2005.

[27] J. Chen, X. Kiu, P. Tu and A. Aragones, "Learning person-specific models for facial expression and action unit recognition," *Pattern Recognition Letters,* vol. 34, 2013.

[28] S. Aslan, S. E. Mete, E. Okur, E. Oktay, N. Alyuz, U. Genc, D. Stanhill and A. A. Esme, "Human Expert Labeling Process (HELP): Towards a reliable higher-order user state labeling by human experts," in *Int. Conf. on Intelligent Tutoring Systems (ITS) - Workshops*, 2016.