From Reaction to Anticipation: Predicting Future Affect

Andres Felipe Zambrano[1], Ryan S. Baker[1], Sami Baral[2], Neil T. Heffernan[2], Andrew Lan[3]
[1]University of Pennsylvania
[2]Worcester Polytechnic Institute
[3]University of Massachusetts Amherst
azamb13@upenn.edu, {ryanshaunbaker, baralsami, neiltheffernaniii}@gmail.com,
andrewlan@cs.umass.edu

ABSTRACT
The educational data mining community has extensively investigated affect detection in learning platforms, finding associations between affective states and a wide range of learning outcomes. Based on these insights, several studies have used affect detectors to create interventions tailored to respond to when students are bored, confused, or frustrated. However, these detector-based interventions have depended on detecting affect when it occurs and therefore inherently respond to affective states after they have begun. This might not always be soon enough to avoid a negative experience for the student. In this paper, we aim to predict students' affective states in advance. Within our approach, we attempt to determine the maximum prediction window where detector performance remains sufficiently high, documenting the decay in performance when this prediction horizon is increased. Our results indicate that it is possible to predict confusion, frustration, and boredom in advance with performance over chance for prediction horizons of 120, 40, and 50 seconds, respectively. These findings open the door to designing more timely interventions.

Keywords
Affect detection, Affect forecasting, Affective computing, Educational data mining, Digital Learning.

1. INTRODUCTION
Since its beginnings, the educational data mining community has been studying the affective states that emerge during learning experiences mediated by digital platforms, attempting to identify, measure, analyze, and appropriately respond to them [16]. Previous research has demonstrated that affective states within digital learning platforms, such as intelligent tutoring systems and educational video games, are correlated with a variety of other important constructs, including self-efficacy [20], analytical reasoning [12], learning outcomes [4, 19, 24, 29], and college enrollment [32]. Historically, engaged concentration (flow) correlates positively with learning outcomes, while confusion and frustration have shown mixed and complex associations with these outcomes. In contrast, boredom has consistently shown a negative impact on learning (see review in [16]).

These insights have led to the use of affect detection in the design of customized interventions aimed at enhancing student engagement and minimizing the experience of affective states like confusion, frustration, and boredom [9, 10, 17, 23]. The core premise of all these research works is that recognizing and addressing students' affective states enhances their interaction with learning environments, making these interactions more engaging and effective [16]. However, existing research primarily targets the identification of affective states at the point when students are already experiencing it, potentially too late to prevent negative impacts.

The challenge of timely intervention is further complicated by the current reliance on methods to establish the ground truth for training these models which do not identify the exact moment when these affective states start (e.g. [2, 15]). Therefore, current detectors are dependent on training labels that might only identify an affective state after it has been occurring for some time, causing even more delayed detection and intervention. Additionally, both detection and intervention do not occur instantaneously, delaying the entire process even more.

Knowing these limitations of current affect detectors, we propose to reframe this task as a prediction of future affect. Early prediction of frustration or boredom enhances the probability of sustaining or quickly restoring a positive affective state rather than trying to reverse negative affect once it has already emerged. Based on this motivation, in this paper, we use machine learning techniques to predict affective states in advance within the ASSISTments learning platform. Specifically, we explore various time horizons for forecasting engaged concentration, confusion, frustration, and boredom, determining the maximum prediction window where performance remains sufficiently high. Additionally, we compare the prediction windows with the half-life of these affective states observed in previous research to assess the feasibility of interventions that can make proactive interventions rather than react.

2. RELATED WORK

2.1 Affect Detection and Learning Outcomes
There has been considerable work to use machine learning models to detect affective states [1, 5, 15, 31] and explore how these affective states are associated with different learning and educational outcomes [18, 24, 32]. These findings, as in research conducted with other methods (see review in [16]), have found relationships between affect and learning, with many replicating across learning environments. For example, Pardos et al. [24] used affect detectors within the ASSISTments platform to investigate the association between detected affect and state test scores. Their results indicated that both engaged concentration and frustration were positively correlated with learning outcomes, whereas confusion and boredom had a negative association with this outcome. Using data from the same platform, San Pedro et al. [32] used affect detectors to predict future college enrollment. They found that engaged concentration

positively predicts college attendance, while those students for whom the detectors identified higher levels of boredom or confusion were also less likely to enroll in college. Within another platform, Reasoning Mind, Kostyuk and colleagues [18] found that detected boredom and confusion were negatively associated with learning, whereas engaged concentration was positively associated with learning. However, these results are complicated somewhat by results such as [19], which found that the duration of confusion or frustration (detected in Cognitive Tutor Algebra) also matters for learning outcomes.

2.2 Detector-Based Interventions

Given that it is possible to detect affect, and affect is associated with learning, there has been considerable research focusing on using detectors to drive interventions that influence student affect. For instance, Padron-Rivera et al. [23] implemented a system to offer hints upon detecting students' confusion or frustration, arguing that such interventions could facilitate a return to engaged concentration while preventing boredom (in line with the affective state dynamics model in [11]). Additionally, they integrated congratulatory messages after correct answers to sustain student engagement. However, their work was not able to impact student affect.

D'Mello and Graesser [10] conducted a study comparing an affect-sensitive version of an intelligent tutoring system with a non-affective version. The affect-sensitive version was designed to recognize and respond to students' affective states, specifically boredom, confusion, and frustration, through pre-programmed emotional responses. Their system was successful at improving learning outcomes, but more for students with lower domain knowledge. In a third study along these lines, DeFalco et al. [9] detected student frustration, giving learners three different types of motivational messages designed around control-value theory, social identity, and self-efficacy. They found that giving frustrated students motivational messages focused on self-efficacy led to higher learning outcomes compared to a control group who did not receive messages. However, the intervention's success did not replicate in a subsequent learning environment designed to be less frustrating, suggesting that the impact of these messages may vary depending on the affective context of the learning environment.

Overall, automated interventions based on affect detection have shown promise but have not fully demonstrated that potential. One possible reason is that these systems may be intervening too late. If a student is already experiencing a negative emotion, it could be less likely for them to return to a more positive affect. Once a student is already frustrated or bored, that negative affect may be difficult to disrupt. By contrast, if an intervention is applied in an earlier stage when students are at risk of becoming frustrated or bored but have not yet done so (or who are just beginning to experience shifts), the chances of maintaining or returning quickly to positive affect are higher. Consequently, the primary objective of this research is to determine how much sooner affective states can be accurately predicted, so that interventions can be proactive about negative emotion rather than reactive to it.

2.3 Advanced Forecasting of Affect

To the best of our awareness, there has not yet been research on the advanced forecasting of affect in education, but efforts along these lines have been conducted in other domains. For example, both neural networks and random forest have been successfully used to forecast a speaker's future affect a few seconds later, from their current and past speech and image data [21, 33]. In addition, researchers have successfully predicted future stress levels from current and recent multimodal data [34, 35]. Though these efforts have involved very different data than digital learning platforms, they increase confidence that this challenge is feasible.

3. METHODS

3.1 Dataset

ASSISTments is a learning platform designed to enable teachers to assign content, offer automated feedback and support for student responses, and generate comprehensive reports on student performance. For this study, 9 middle-school mathematics teachers who frequently used ASSISTments [14] were recruited between 2021 and 2023, to assign problems using the platform and collect affective data from their students. A total of 312 middle-school students from the 9 teachers participated in the study, where they solved mathematics problems using the ASSISTments platform and reported their affective states. To support the replicability of our results and further experiments, the full data set and code used in this research can be found at https://osf.io/spg6v/.

To facilitate the collection of affective data, a self-reporting infrastructure was integrated into the ASSISTments platform, as shown in Figure 1. This infrastructure was designed to be minimally disruptive, ensuring that the primary focus of students remained on their mathematics learning. Upon the completion of a problem within their assigned mathematics assignment, students were prompted to report their affective state. The self-reports were designed based on past self-report approaches for affect [27] and iteratively designed with members of the target population. For this study, we focused on engaged concentration, confusion, boredom, and frustration, which are the most commonly studied affective states in online learning environments (see reviews in [2, 16]). The order of each affective state was randomized in the survey each time it was presented to the students. Students were randomly asked to report their affective states, either once or twice, for each assignment they completed. This decision was governed by a probabilistic algorithm, where there was a 10% chance of picking two problems from an assignment for affective reporting and a 90% chance of picking only one problem. The students were instructed to report their affective states immediately after completing the selected problem(s), thereby ensuring the timeliness and relevance of the affective data. We limited how often students were asked about their affective states to avoid fatigue, minimize potential disruptions to their learning experience, and avoid inadvertently inducing disinterest and boredom.

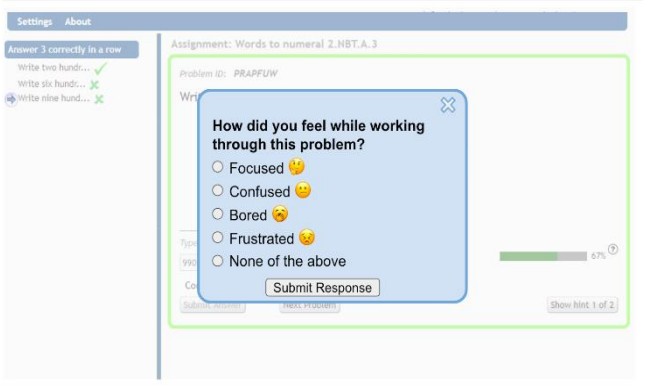

Figure 1. Self-Report survey within an assignment in the ASSISTments platform, asking students to report their affective state when working on an assignment.

This dataset collected using the ASSISTments infrastructure includes information about students' correct and incorrect answers, hint requests, and self-reported affective states. On average, each student solved 1.81 assignments (SD = 1.21), submitted 14.2 responses (SD = 10.5), requested hints 2.84 times (SD = 2.2), and self-reported their affective state 1.75 times (SD = 1.25).

The distribution of these reported affective states is detailed in Table 1. The most common affective state reported was *None* (29.0%), followed by engaged concentration (27.8%). Frustration was the least reported state, noted by only 8.3% of the reports. These proportions are similar to those observed in previous studies when students categorized their own affective states. For instance, Baker et al. [3] reported that students most frequently identified as feeling Neutral (29%) and Engaged Concentrated (20%) when they assessed their own affective states every 20 seconds while watching recordings of themselves using AutoTutor (a computerized tutor that mimics human tutors and converses with students in natural language, as described by [13]). In that study, students categorized Frustration, Confusion, and Boredom at rates of 11%, 18%, and 16%, respectively, which aligns to the distribution observed in our study.

Table 1. Distributions of affective states.

Affective State	# Samples	%
Eng. Concentration	151	27.8
Confusion	90	16.5
Frustration	46	8.3
Boredom	99	18.2
None	158	29.0
Total	544	100

Past studies where affect was identified by trained experts had a very different distribution of affective states than the student self-reports we observed. Generally, Engaged Concentration is the predominant affective state noted by trained human observers, exceeding 60% in multiple studies across various learning platforms, including ASSISTments [1, 3, 22, 24]. This difference in proportions between researcher-categorized affect and student self-reports has been found even within the exact same sample of students. For instance, [38] reported that substantially fewer positive emotions (concentration, focus, delight, and happiness) were observed by trained experts compared to self-reports collected simultaneously. Though experts disagree with self-report, it is unclear which method is more accurate, given the limitations of each approach (systematic error and bias for observers; demand, self-presentation, and lack of meta-awareness for self-report; see discussions in [26, 27]).

3.2 Prediction in Advance

The dataset was segmented into 5-second clips. This approach was adopted instead of the more commonly used 20-second segmentation to increase the granularity of the analysis. For each clip, we crafted 58 features, capturing diverse aspects of student interaction with the educational software. These aspects include the correctness of responses, the frequency of answers, hints requested, the time elapsed since the last action, and others. These features were tailored to reflect both general and skill-specific student interactions. Although the core clip is defined as 5 seconds, when predicting the label (typically from before the clip, since we are predicting in advance), the set of features also included data aggregated over the prior 20 seconds, 1 minute, and 3 minutes, and during the prior 3, 5, and 8 actions within the current study session. These features are inspired by similar features employed in previous studies that trained affect detectors using ASSISTments data [24, 36]. When aggregating data on the last 3, 5, or 8 actions, we do not consider actions of previous learning sessions that ended more than one hour earlier, because these older actions are fairly unlikely to influence the student's current affective state. Additionally, if the student has not performed any actions recently, it becomes impossible to calculate certain features, or they default to values of 0. Therefore, we filtered out periods of inactivity exceeding one hour.

In a model capable of predicting an affective state N minute in advance, the features must correspond to actions that occurred more than N minutes prior to the affective self-report of the students. For this reason, all affective states reported during the first N minutes after the first action of a studying session cannot be used for training and testing the models. This constraint, while necessary, does distort the sample somewhat, for students who solve a few or only one question in each studying session because those affective states will not be matched with a set of previous actions before the corresponding prediction horizon, reducing the number of samples available for developing the machine learning models.

Figure 2 presents the number of self-reports for each affective state with enough data (at least one action before the corresponding time horizon of N seconds within the same study session) for developing prediction models for each horizon ranging from 0 to 5 minutes in 10-second increments. As this time horizon increases, the number of samples with sufficient data to train and test a predictive model decreases. Based on the data available for each time horizon, we only considered horizons between 0 and 3 minutes in advance. This selection aimed to keep an adequate number of samples for training and testing the detectors while also establishing a sufficiently wide range of prediction horizons to investigate whether there is a decline in prediction performance as this horizon widens.

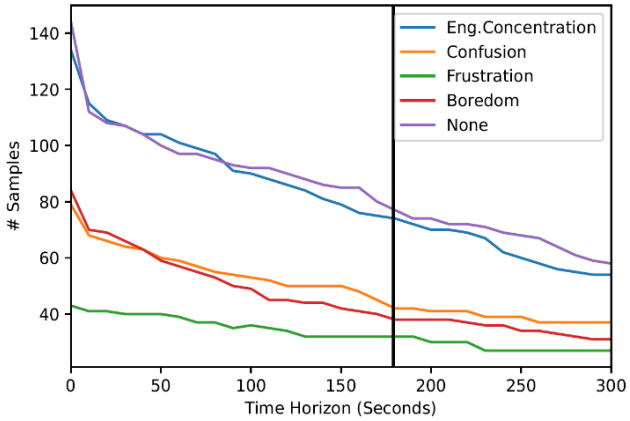

Figure 2. Number of samples for each affective state and each time horizon.

In Table 2, we present the distribution of samples that have sufficient prior data for a 3-minute temporal horizon to be possible. From the initial pool of 544 samples, only 265 (48.7%) met the criteria for the selected horizon. The distribution of affect in this reduced data set closely resembles the original dataset. Although more data is available for prediction horizons shorter than 3 minutes, we used this selected set of 265 samples for training and

validating all models across all the temporal horizons considered in this study (from 0 to 180 seconds in 10-second increments). This methodological choice allows us to attribute any observed changes in model performance specifically to the extension of the prediction horizon rather than to the influence of sample size variation or differences in affect at different times within the session.

Considering the dataset's relatively small size, we employ Logistic Regression (LR) and decision tree-based methods like Random Forests (RF) and Extreme Gradient Boosting (XGB) with the default parameter settings [8, 25]. For each affective state, separate binary classifiers were trained, as in [1, 24]. To validate these models, we applied a 4-fold stratified student-level cross-validation, selecting 4 folds to keep test sets large despite the small sample size. To enhance the robustness of our estimation, we repeated this process with 10 unique random seeds, thereby generating 40 unique combinations of training and testing sets. We assessed model performance using the Area under the Receiver Operating Characteristic Curve (AUC ROC), providing a comprehensive evaluation across various thresholds. This is particularly useful for evaluating the model's applicability in diverse interventions with varying cost-benefit trade-offs. The mean and standard deviation of the AUC were calculated across these 40 validation sets. Additionally, we examined the confusion matrix to identify and understand each detector's misclassification patterns. We evaluated the mean decrease impurity (MDI; [7]) feature importance of each model to understand how the most important features varied across different prediction horizons. This feature importance metric was chosen due to its straightforward computation, which relies on the division of decision trees, while reducing the risk of hiding important features that do not have a uniform association (either positive or negative) with the outcome and that are dependent on interactions with other features [7].

Table 2. Affect distributions for samples with enough data for training prediction models with a time horizon of at least 3 minutes.

Affective State	# Samples	%
Eng. Concentration	74	27.9
Confusion	42	15.8
Frustration	32	12.1
Boredom	39	14.7
None	78	29.4
Total	265	100

4. RESULTS

4.1 Affect Detection

Table 3 presents the baseline performance of affect detectors, operating with a zero-second time horizon (i.e. the actual clip time), using Random Forest (RF), XGBoost (XGB), and Logistic Regression (LR) models. The results indicate that frustration is the best predicted affective state using the decision tree-based models (RF and XGB), with the XGB model achieving the highest performance (AUC=0.727). Confusion and boredom are also effectively detected by the ML models, showing an AUC of 0.671 for confusion and 0.638 for boredom using RF and XGB, respectively. In contrast, the models could not accurately detect engaged concentration in this data set, showing an AUC of 0.503 (chance level).

Table 4 presents the confusion matrix for the detectors. Columns represent the real affective states reported by the students, while rows indicate the number of instances where the corresponding detector identifies each affective state. We set the classification threshold at 0.3, as the detectors' outputs tended to be below the conventional 0.5 threshold. Table 4 also includes, in parentheses, the percentage of each actual affective state that was identified by the detectors as the ground truth affect in that column (whether correctly or incorrectly). For example, the confusion matrix reveals that the Engaged Concentration detector incorrectly identifies 49.5% of actual Boredom instances and 45.9% of None instances as Concentration, both higher percentages than its correct identification of actual Engaged Concentration instances (36.7%). This suggests that the Engaged Concentration detector's performance is compromised by misidentifying Boredom and None instances. One possible interpretation is that the None category may include some cases which would have been categorized as Engaged Concentration by experts, and that students do not fully understand the distinction between these affective states (and therefore are actually incorrectly categorizing their own affect). Similarly, the Boredom detector also misclassifies 10.5% of None instances and 8.2% of Engaged Concentration instances as Boredom.

Table 3. Detection of current affective state. 4-fold student-level cross-validation AUC of affect detectors employing different ML techniques. Best performing models for each affective state are shown in bold. Standard deviation of performance metrics across the 4 folds are shown in parenthesis.

Affective State	RF	XGB	LR
Eng. Con	**0.503 (0.070)**	0.483 (0.068)	0.439 (0.072)
Conf	**0.671 (0.065)**	0.668 (0.079)	0.597 (0.074)
Fru	0.688 (0.095)	**0.727 (0.087)**	0.556 (0.115)
Bor	0.627 (0.085)	**0.638 (0.082)**	0.570 (0.075)

Table 4. Confusion Matrix of Detections. Columns correspond to self-reported affect and rows correspond to detector outputs. In parentheses, the percentage of each actual affective state that was identified by the detectors as the ground truth affect in that column (whether correctly or incorrectly).

Detector	Eng. Con	Conf	Fru	Bor	None
Eng. Con	27.2 (36.7)	11.1 (26.4)	10.2 (31.9)	19.3 (49.5)	35.8 (45.9)
Conf	7.2 (9.7)	10.1 (24.4)	8.4 (26.3)	2.5 (6.4)	6.1 (7.1)
Fru	3.1 (4.2)	4.2 (10.0)	6.6 (20.6)	0.4 (1.0)	3.4 (4.4)
Bor	6.1 (8.2)	2.5 (6.0)	1.7 (5.3)	8.4 (21.5)	8.2 (10.5)

The confusion and frustration detectors successfully identified Engaged Concentration, Boredom, and None as neither frustration nor confusion. None of these three affective states is misidentified by the confusion or frustration detectors at a rate higher than 10%. However, the confusion detector identifies 24.4% of confusion and 26.3% of frustration instances as confusion. Similarly, but at a lesser level, the frustration detector identifies 20.6% of the frustration instances and 10% of confusion instances as frustration. These results suggest that some students might be having trouble distinguishing between these two affective states or that both share

similar patterns that make them be identified together. Indeed, some recent work has argued that these two affective states should be lumped together during detection [28].

4.2 Affect Prediction in Advance

Figures 3 through 5 present the performance of predictive models for each affective state with different time horizons ranging between 0 to 3 minutes with 10-second increments. In each figure, a range representing +1 and -1 standard deviations, as well as the chance level performance, are included. For this analysis, we exclude Engaged Concentration because, as shown in Table 3, the detection and prediction models for Engaged Concentration do not surpass chance-level performance.

The confusion detectors trained using RF perform above 0.6 AUC for temporal horizons within the 1-minute range, as shown in Figure 3. For all horizons less than 2 minutes, performance was consistently more than one standard deviation above chance. In the case of frustration detectors (trained using XGB), performance remained at least one standard deviation above chance for prediction horizons up to 40 seconds in advance, as shown in Figure 4. In all these cases, the models achieved an AUC of over 0.6. The predictive models for Boredom (trained using XGB; see Figure 5) show comparable outcomes. In this case, for prediction horizons up to 50 seconds, all models had an AUC above 0.6, with performance at least one standard deviation better than chance.

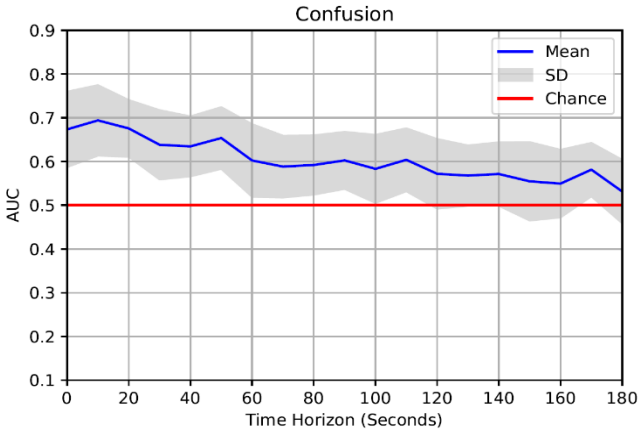

Figure 3. Confusion prediction with different time horizons using RF.

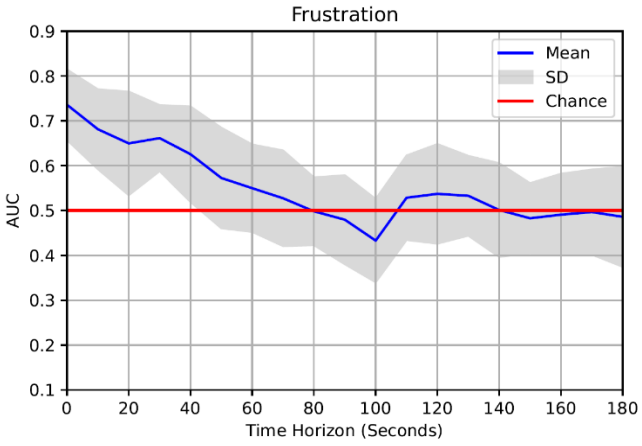

Figure 4. Frustration prediction with different time horizons using XGB.

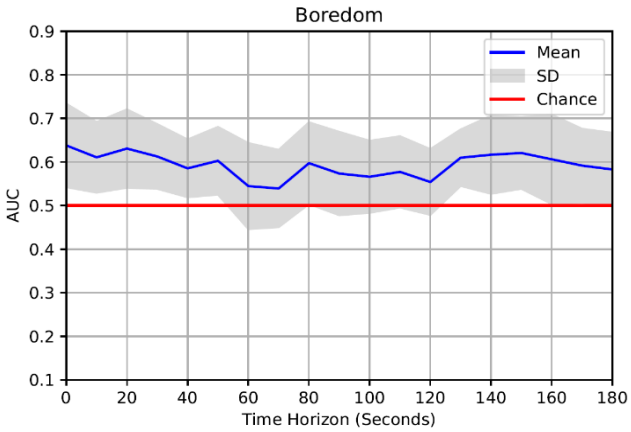

Figure 5. Boredom prediction with different time horizons using XGB.

4.3 Most Important Features

We evaluated the most important features for each affect prediction considering the horizons of 0 (current affect), 60, and 120 seconds. As shown in Table 5, the time spent in the assignment appeared as an important feature across all prediction horizons for confusion. Errors in the current problem type (i.e. multiple-choice question) and time since the last error were important for the predictions of 0 and 60 seconds but not for 120 seconds. Time since the last correct answer and time spent in the last attempt appeared as important features for 60 and 120 seconds but not for detecting the current affect. In the case of frustration detection, the number of errors in the last 8 attempts appeared as an important feature across all time horizons. However, in contrast with the results observed for the confusion prediction, the detector of current frustration does not share any other feature with the 60-second and 120-second prediction models. Finally, the boredom detection and prediction models shared 3 of the top 5 most important features (time since requesting the last hint, number of errors in multiple-choice questions, and attempts during the last minute).

5. DISCUSSION AND CONCLUSIONS

According to our results, confusion, frustration, and boredom can be predicted in advance with performance over chance for time horizons of 120, 40, and 50 seconds, respectively. For each affective state, and particularly for confusion and boredom, the most important features for both prediction (60 and 120 second horizons) and detection (0 second horizon) models were similar. This result suggests that both models might be capturing a similar signal, just varying the timing of the prediction, indicating that the labels are likely autocorrelated. This finding reinforces the argument that these affective states can be anticipated. However, one limitation to our interpretation is that we are uncertain when each affective state instance actually began, a difficult thing to be certain of with any ground truth method. Thus, it is important to be careful in the interpretation of these results. Nevertheless, it is unlikely that all of our predictive success is due to capturing earlier onset of the later affective state. Botelho et al. [6] found that for the ASSISTments platform, confusion can persist for 40 seconds to 1 minute, frustration for 2 minutes, and boredom for even 5 minutes. Comparing this with our prediction windows, we see that our 120-second prediction window for confusion exceeds its half-life. This suggests that we are likely to be predicting at least some confusion before it actually occurs. In contrast, our prediction windows for frustration and boredom are shorter than their potential durations. For this

Table 5. Top 5 most important features for the top-performing prediction model for each affective state for a prediction horizon of 0 (detection), 60, and 120 seconds. Common features across different time horizons are shown in bold.

Detector	0 seconds	60 seconds	120 seconds
Conf	**Time Spent in the Assignment**	**Time Spent in the Assignment**	**Time Spent in the Assignment**
	Errors in Current Problem Type	**Errors in Current Problem Type**	Time Spent in the Last 8 Attempts
	Time since Last Error	**Time since Last Error**	Time Spent in the Last 3 Attempts
	Time in Questions of the Same Skill	**Time since Last Correct Answer**	**Time since Last Correct Answer**
	Errors in Questions of the Same Skill	**Time Spent in the Last Attempt**	**Time Spent in the Last Attempt**
Frust	**Errors in Last 8 Attempts**	**Errors in Last 8 Attempts**	**Errors in Last 8 Attempts**
	Errors in Questions of the Same Skill	**Hints in the Assignment**	**Hints in the Assignment**
	Time since Last Hint	**Hints in Last 8 Attempts**	**Hints in Last 8 Attempts**
	Hints Requested for Skill	**Attempts in Current Problem Type**	**Attempts in Current Problem Type**
	Time in Questions of the Same Skill	Attempts in Check-All Questions	Errors in Last 20 seconds
Bored	**Time since Last Hint**	**Time since Last Hint**	**Time since Last Hint**
	Errors in Multiple Choice Questions	**Errors in Multiple Choice Questions**	**Errors in Multiple Choice Questions**
	Attempts in Last Minute	**Attempts in Last Minute**	**Attempts in Last Minute**
	Attempts in Multiple Choice Questions	**Errors in Last 8 Attempts**	**Errors in Last 8 Attempts**
	Errors in Current Problem Type	Time since Last Error	Attempts in Same Problem Type

reason, students might be already experiencing those affective states, in some cases, when the prediction models determine that students would report them in the next minute. Although this alternate interpretation of the model functionality does not correspond to the original goal of intervening before students feel frustration or boredom, even in this case this approach remains useful because it allows earlier detection than what previous detectors can do.

The potential autocorrelation of these signals suggests that future work could compare the predictive improvements of ML-based detectors against models that solely use autocorrelation based on the labels' time series. If autocorrelation models alone (using the collected labels exclusively) provide substantial predictive accuracy, it could indicate that labels of previous instances should be included in the feature set of ML-based models. In this case, we excluded the labels from the feature set because we want to be able to use detectors without continuously asking students about their affective states. However, the predicted labels in previous instances could also be considered when making later predictions, as is seen in some neural network topologies, such as the Long-Short Term Memory networks that have shown promising performance for affect detection [5].

One limitation of our study that prevented us from exploring neural networks was the limited sample size available for training the machine learning models. The reduced number of self-report requests for each student, to reduce issues from repeating the same question excessively, resulted in a smaller sample size than what previous detector research based on self-reports has typically gathered [37, 38]. Using a larger dataset collected among a larger number of learners or increasing the granularity of the data acquisition would likely lead to better model performance. This ground truth with higher granularity or larger number of samples would enable a more precise tuning of the hyperparameters in the ML models and use other machine learning techniques like artificial neural networks [5], potentially enhancing the performance of the models. Additionally, future work can also explore data augmentation techniques as an alternative to obtaining more training samples.

One surprise in our findings is the relatively low frequency of engaged concentration and the unusually poor performance of engaged concentration detection compared to previous engaged concentration detection in ASSISTments (and other systems as well) [3, 24, 31]. One possible interpretation is that offering *None* as a response option in students' self-reports might be skewing the performance of our models, particularly engaged concentration. Engaged concentration is commonly the dominant affective state in various learning environments [3, 22, 24, 38], as the only affective state with a non-negative valence studied in most environments [2, 30]. Typically, engaged concentration is low activation [3], so it is possible that some students are not even realizing that they are engaged. Alternatively, it may be that past studies based on expert judgment confused engaged concentration with the neutral affective state or an absence of affect. Earlier research also indicates that positive emotions, especially engaged concentration, are reported less frequently when individuals assess their own affective state [3, 38]. The reduced effectiveness of engaged concentration detection

in this study, along with its tendency to identify instances of boredom and *None* more than actual engaged concentration, suggests that self-reporting might be less reliable for identifying engaged concentration compared to expert labeling methods.

Despite these limitations, models that can predict (or early detect) confusion, frustration, and boredom can be useful, as they can lead to intervention before a student's negative affect (particularly boredom) lead to problematic behaviors such as gaming the system, which are detrimental to their learning outcomes [3]. This is particularly important in systems with high latency as such delays further postpone interventions, amplifying the risk of negative outcomes stemming from late responses. For instance, early detection of frustration or boredom enables the learning platform to suggest breaks or deliver motivational messages to the students. This early detection can also facilitate other interventions, such as switching to a different learning activity in the next problem or even increasing the difficulty of the subsequent activities if the system detects that boredom will appear soon, but there is a low risk of future confusion or frustration. This approach would help prevent students from becoming disinterested in their studies, a typical outcome of boredom, or unresolved confusion or frustration, which lead to poorer learning. As such, detecting affect early or in advance may help us to develop learning systems that better support learners' motivation and learning.

6. ACKNOWLEDGMENTS

We thank the National Science Foundation, award IIS-1917545, for their support. We also thank Xiner Liu for conducting a software review to validate correctness and match to paper. Andres Felipe Zambrano thanks the Ministerio de Ciencia, Tecnología e Innovación and the Fulbright-Colombia commission for supporting his doctoral studies through the Fulbright-MinCiencias 2022 scholarship.

7. REFERENCES

[1] Baker, R.S., Gowda, S.M., Wixon, M., Kalka, J., Wagner, A.Z., Salvi, A., Aleven, V., Kusbit, G.W., Ocumpaugh, J. and Rossi, L. 2012. Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra. *International Educational Data Mining Society*. (2012).

[2] Baker, R.S., Ocumpaugh, J.L. and Andres, J. 2020. BROMP quantitative field observations: A review. *Learning Science: Theory, Research, and Practice. New York, NY: McGraw-Hill*. (2020), 127–156.

[3] Baker, R.S.J. d., D'Mello, S.K., Rodrigo, Ma.M.T. and Graesser, A.C. 2010. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*. 68, 4 (Apr. 2010), 223–241.

[4] Bosch, N. and D'Mello, S. 2017. The Affective Experience of Novice Computer Programmers. *International Journal of Artificial Intelligence in Education*. 27, 1 (Mar. 2017), 181–206.

[5] Botelho, A.F., Baker, R.S. and Heffernan, N.T. 2017. Improving Sensor-Free Affect Detection Using Deep Learning. *Artificial Intelligence in Education: 18th International Conference* (2017), 40–51.

[6] Botelho, A.F., Baker, R.S., Ocumpaugh, J. and Heffernan, N.T. 2018. Studying Affect Dynamics and Chronometry Using Sensor-Free Detectors. *International Educational Data Mining Society*. (2018).

[7] Breiman, L. 2017. *Classification and regression trees*. Routledge.

[8] Chen, T. and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (2016), 785–794.

[9] DeFalco, J.A., Rowe, J.P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B.W., Baker, R.S. and Lester, J.C. 2018. Detecting and Addressing Frustration in a Serious Game for Military Training. *International Journal of Artificial Intelligence in Education*. 28, 2 (Jun. 2018), 152–193.

[10] D'mello, S. and Graesser, A. 2013. AutoTutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Trans. Interact. Intell. Syst.* 2, 4 (Jan. 2013).

[11] D'Mello, S. and Graesser, A. 2012. Dynamics of affective states during complex learning. *Learning and Instruction*. 22, 2 (2012), 145–157.

[12] D'Mello, S., Person, N. and Lehman, B. 2009. Antecedent-consequent relationships and cyclical patterns between affective states and problem solving outcomes. *Artificial Intelligence in Education* (2009), 57–64.

[13] Graesser, A.C., Person, N., Harter, D., Group, T.R., and others 2000. Teaching tactics in AutoTutor. *Modelling human teaching tactics and strategies. International Journal of Artificial Intelligence in Education*. 11, (2000), 1020–1029.

[14] Heffernan, N.T. and Heffernan, C.L. 2014. The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching. *International Journal of Artificial Intelligence in Education*. 24, 4 (Dec. 2014), 470–497.

[15] Hutt, S., Grafsgaard, J.F. and D'Mello, S.K. 2019. Time to Scale: Generalizable Affect Detection for Tens of Thousands of Students across An Entire School Year. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2019), 1–14.

[16] Karumbaiah, S., Baker, R., Tao, Y. and Liu, Z. 2022. How does Students' Affect in Virtual Learning Relate to Their Outcomes? A Systematic Review Challenging the Positive-Negative Dichotomy. *LAK22: 12th International Learning Analytics and Knowledge Conference* (New York, NY, USA, 2022), 24–33.

[17] Karumbaiah, S., Lizarralde, R., Allessio, D., Woolf, B., Arroyo, I. and Wixon, N. 2017. Addressing Student Behavior and Affect with Empathy and Growth Mindset. *International Conference on Educational Data Mining* (2017), 96–103.

[18] Kostyuk, V., Almeda, Ma.V. and Baker, R.S. 2018. Correlating affect and behavior in reasoning mind with state test achievement. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (New York, NY, USA, 2018), 26–30.

[19] Liu, Z., Pataranutaporn, V., Ocumpaugh, J. and Baker, R. 2013. Sequences of frustration and confusion, and learning. *International Conference on Educational Data Mining* (2013), 114–120.

[20] McQuiggan, S.W. and Lester, J.C. 2009. Modelling affect expression and recognition in an interactive learning environment. *International Journal of Learning Technology*. 4, 3–4 (2009), 216–233.

[21] Noroozi, F., Akrami, N. and Anbarjafari, G. 2017. Speech-based emotion recognition and next reaction prediction. *2017 25th Signal Processing and Communications Applications Conference (SIU)* (2017), 1–4.

[22] Ocumpaugh, J., Andres, J.M., Baker, R., DeFalco, J., Paquette, L., Rowe, J., Mott, B., Lester, J., Georgoulas, V., Brawner, K. and Sottilare, R. 2017. Affect Dynamics in Military Trainees Using vMedic: From Engaged Concentration to Boredom to Confusion. *Artificial Intelligence in Education: 18th International Conference* (2017), 238–249.

[23] Padron-Rivera, G., Joaquin-Salas, C., Patoni-Nieves, J.-L. and Bravo-Perez, J.-C. 2018. Patterns in Poor Learning Engagement in Students While They Are Solving Mathematics Exercises in an Affective Tutoring System Related to Frustration. *Pattern Recognition: 10th Mexican Conference* (2018), 169–177.

[24] Pardos, Z.A., Baker, R.S., San Pedro, M.O., Gowda, S.M. and Gowda, S.M. 2014. Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes. *Journal of Learning Analytics*. 1, 1 (2014), 107–128.

[25] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., and others 2011. Scikit-learn: Machine learning in Python. *Journal of machine Learning research*. 12, (2011), 2825–2830.

[26] Porayska-Pomsta, K., Mavrikis, M., D'Mello, S., Conati, C. and Baker, R.Sj. 2013. Knowledge elicitation methods for affect modelling in education. *International Journal of Artificial Intelligence in Education*. 22, 3 (2013), 107–140.

[27] Rebolledo-Mendez, G., Huerta-Pacheco, N.S., Baker, R.S. and du Boulay, B. 2022. Meta-Affective Behaviour within an Intelligent Tutoring System for Mathematics. *International Journal of Artificial Intelligence in Education*. 32, 1 (Mar. 2022), 174–195.

[28] Richey, J.E., Zhang, J., Das, R., Andres-Bray, J.M., Scruggs, R., Mogessie, M., Baker, R.S. and McLaren, B.M. 2021. Gaming and Confrustion Explain Learning Advantages for a Math Digital Learning Game. In *International conference on artificial intelligence in education* (2021), 342–355.

[29] Rodrigo, M.M.T., Baker, R.Sj. and Nabos, J.Q. 2010. The relationships between sequences of affective states and learner achievement. *Proceedings of the 18th international conference on computers in education* (2010), 56–60.

[30] Russell, J.A. 2003. Core affect and the psychological construction of emotion. *Psychological Review*. 110, 1 (2003), 145–172.

[31] Sabourin, J., Mott, B. and Lester, J.C. 2011. Modeling learner affect with theoretically grounded dynamic Bayesian networks. *Affective Computing and Intelligent Interaction: 4th International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part I 4* (2011), 286–295.

[32] San Pedro, M.O., Baker, R., Bowers, A. and Heffernan, N. 2013. Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. *International Conference on Educational Data Mining* (2013).

[33] Shahriar, S. and Kim, Y. 2019. Audio-Visual Emotion Forecasting: Characterizing and Predicting Future Emotion Using Deep Learning. *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)* (2019), 1–7.

[34] Taylor, S., Jaques, N., Nosakhare, E., Sano, A. and Picard, R. 2020. Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health. *IEEE Transactions on Affective Computing*. 11, 2 (2020), 200–213.

[35] Umematsu, T., Sano, A., Taylor, S. and Picard, R.W. 2019. Improving Students' Daily Life Stress Forecasting using LSTM Neural Networks. *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)* (2019), 1–4.

[36] Wang, Y., Heffernan, N.T. and Heffernan, C. 2015. Towards better affect detectors: effect of missing skills, class features and common wrong answers. *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (New York, NY, USA, 2015), 31–35.

[37] Wixon, M., Arroyo, I., Muldner, K., Burleson, W., Rai, D. and Woolf, B. 2014. The opportunities and limitations of scaling up sensor-free affect detection. *International Conference on Educational Data Mining* (2014), 145–152.

[38] Zambrano, A.F., Nasiar, N., Ocumpaugh, J., Goslen, A., Zhang, J., Rowe, J., Esiason, J., Vandenberg, J., and Hutt, S. 2024. Says Who? How different ground truth measures of emotion impact student affective modeling. *Proceedings of the 17th International Conference on Educational Data Mining* (2024).

8. APPENDIX I: LIST OF FEATURES

The complete list of features employed in this study is:

- Number of skills practiced by the student.
- Time taken in the last response.
- Average time taken in the last 3 responses.
- Average time taken in the last 5 responses.
- Average time taken in the last 8 responses.
- Time taken in the current assignment.
- Days since the student started the current assignment.
- Total attempts.
- Percentage of wrong answers in the last 3 responses.
- Percentage of wrong answers in the last 5 responses.
- Percentage of wrong answers in the last 8 responses.
- Hints requested in the current assignment.
- Hints requested in the last 3 responses.
- Hints requested in the last 5 responses.
- Hints requested in the last 8 responses.
- Total errors.
- Total hints requested.
- Attempts in the last 20 seconds.
- Attempts in the last minute.
- Attempts in the last 3 minutes.
- Errors in the last 20 seconds.
- Errors in the last minute.
- Errors in the last 3 minutes.
- Hints requested in the last 20 seconds.

- Hints requested in the last minute.
- Hints requested in the last 3 minutes.
- Time since the last attempt.
- Time since the last error.
- Time since the last correct answer.
- Time since the last hint requested.
- Is the current problem a Match problem?
- Is the current problem a numeric value problem?
- Is the current problem a multiple-choice problem?
- Is the current problem an algebraic expression problem?
- Is the current problem a check all that apply problem?
- Attempts in the current problem type.
- Errors in the current problem type.
- Hints requested in the current problem type.
- Attempts in the match problems.
- Attempts in numeric value entry problems.
- Attempts in multiple-choice problems.
- Attempts in algebraic expression problems.
- Attempts in check-all that apply problems.
- Total errors in match problems.
- Total errors in numeric value entry problems.
- Total errors in multiple-choice problems.
- Total errors in algebraic expression problems.
- Total errors in check-all that apply problems.
- Total time solving problems of the current skill.
- Is this problem the first time the student has practiced this skill?
- First time practicing a skill in the last 20 seconds.
- First time practicing a skill in the last minute.
- First time practicing a skill in the last 3 minutes.
- Error in the first time the student practiced the skill.
- Total responses of the student in problems of the current skill.
- Total errors of the student practicing the current skill.
- Total hints requested by the student practicing the current skill.
- Action outside school hours (before 8 in the morning or after 5 in the afternoon).