Active Learning for a Classroom Observer who Can't Time Travel

Andres Felipe Zambrano
Graduate School of Education
University of Pennsylvania
Philadelphia, Pennsylvania, USA
azamb13@upenn.edu

Ryan S. Baker
Graduate School of Education
University of Pennsylvania
Philadelphia, Pennsylvania, USA
ryanshaunbaker@gmail.com

Andrew S. Lan
Manning College of Information & Computer Sciences
University of Massachusetts Amherst
Amherst, Massachusetts, USA
andrewlan@cs.umass.edu

***Abstract*— Classroom observation has been used to obtain training labels for affect detection, but is expensive for large representative samples. Active Learning (AL) methods have been proposed to address this challenge by identifying the specific samples that should be labeled to improve detector performance, based on a metric of informativeness. While previous work has investigated the potential benefits of AL methods in affect detection, they have considered scenarios that may not completely reflect reality, where an observer can code any student and time window within the entire data set. Unfortunately, actual use of such a method can only take place in the current time window -- classroom observers cannot time travel. This paper explores the potential benefit of AL methods in a scenario that more closely mimics the human coder's observation process in a real classroom -- where the coder can only observe behavior occurring at the current moment. Our experimental results show that AL methods slightly improve the performance indicators of binary detectors for concentration, confusion, and frustration compared to control sampling methods. However, there is no benefit for boredom detection. These findings have implications for the use of active learning-based data collection protocols for developing affect detectors.**

Index Terms*— active learning, affect detection, educational data mining, learning analytics

I. INTRODUCTION

The use of physical and physiological sensors [1]-[5] and interaction log data from student-computer interactions [6]-[9] has been proven effective in detecting affective states in the classroom. However, the transferability of affect detection models across different student populations is not always successful [10], [11]. Therefore, researchers need to collect large and representative samples to develop trustworthy detection models, thereby considerably increasing the implementation costs [12].

While approaches such as sensor-free detection are highly automated and reduce implementation costs, collecting labeled data still requires trained human observers [9]-[11] or self-reports [6]-[8], both of which have limitations. Frequent self-reports may disrupt students' engagement in their learning experience. For instance, the act of self-labeling an emotion has been shown to modify the physiological response of the person [13]. Moreover, young students may lack the ability to define and report on their affective states [14]. On the other hand, human observations, while able to mitigate these issues, also have a significant drawback due to the process required to code students' affective states in real classrooms [15], including training, scheduling, and deployment of observers, which creates an expensive bottleneck. Thus, while there is a wealth of data on student activity, collecting labeled data on affect presents a significant challenge. Therefore, exploring alternative approaches to develop machine learning models that can perform better with less annotated data is crucial.

To address the issue of data availability in affect detection, researchers have explored the use of semi-supervised learning for facial expression recognition, using datasets with mostly unlabeled samples [16]. This approach compares detectors based on their performance using exclusively labeled data versus incorporating labeled and unlabeled data during training. Results show that semi-supervised learning outperforms supervised methods that do not use unlabeled data. However, this approach has a limitation when the unlabeled data cannot be appropriately leveraged due to an insufficient number of labeled samples. Therefore, there is still a need to determine the most informative subset of samples to be labeled before applying these methods.

Active Learning (AL) [17], a subfield of Machine Learning (ML), has emerged as a potential solution to the challenge of developing accurate models with fewer labeled data samples. By enabling the ML model to choose which data to train on, AL methods select a smaller subset of carefully chosen data samples that can potentially lead to better detector performance at lower cost. This approach allows for a more targeted labeling effort, focused on the data samples that can be most helpful for the model, instead of coding all the students employing a predetermined order [15], which does not necessarily provide the most informative samples for the model. Ideally, this approach is conducted in real-time, when the samples are being collected, to make better observation decisions.

Previous studies have shown that using AL methods to sample data can improve the performance of affect detectors [11], [18]. However, these studies are based on an unrealistic assumption, that observations can be chosen across all available data, which consist of multiple classroom sessions, occurring across multiple days, in several schools. In reality, a coder cannot rely on future information or visit other classrooms or schools to decide which student to observe next. The coder must choose between the students in their current classroom, observing their affective state at the current moment. Thus, while these studies provide a promising starting point for exploring the use of AL for affect detection, their assumptions do not reflect the core challenges and limitations to what data the observer can collect at any given time. For these reasons, [18] acknowledged that the benefits of AL methods in real-world scenarios might be lower than the potential impact suggested by experimental results. Our goal is therefore to analyze the effectiveness of AL

methods in a more realistic scenario that takes into account the temporal nature of the sampling process.

In this paper, we investigate the potential of AL methods to improve affect detection by using a sequence of samples that mimics the observation process of a human observer in a real classroom. We analyze our results both in terms of future prediction within the current school, and for prediction of new students in unseen schools, examining the findings' degree of robustness across different schools. Going beyond previous studies that exclusively used Logistic Regression for detection with AL methods [11], [18], we also explore the use of Support Vector Machine (SVM), Random Forest (RF), and Multi-Layer Perceptron (MLP) classifiers. Our experimental results demonstrate that AL methods slightly improve the performance of binary detectors for three affective states in this more realistic scenario compared to control sampling methods. However, for boredom, AL methods do not appear to provide benefit to more straightforward sampling approaches.

II. Related Work

In this section, we will briefly introduce active learning methods and their previous applications in the context of affect detection. We will also review previous research on the generalizability of detectors across populations.

A. Active Learning in Student Affect Detection

Collecting labeled data for affect detection is generally expensive or disruptive to students . One solution is to attempt to collect the labels that are most informative for the model rather than trying to label all available/possible data. Active Learning (AL) provides a set of methods that determine which samples are likely to be most informative, based on metrics such as observation uncertainty, expected error reduction, expected variance reduction, and model change [19].

In the context of student affect detection, active learning has been used to study whether it is possible to improve detectors' performance without increasing resource use [11], [18]. According to Yang et al. [18], the most promising AL method is Linear Minimum Mean Squared Error (LMMSE), which identifies the data point least similar to the previously sampled points to achieve maximum reduction of the MSE. Another method is Uncertainty Sampling (UncS) [11], [18], which uses a detector trained by the previously sampled data to identify the following observation that has the highest level of uncertainty for the detector. Both LMMSE and UncS have shown a higher improvement in classifier performance than other more sophisticated AL methods for real-world benchmark datasets [19]-[21]. In this work, we employ both LMMSE and UncS.

A major challenge for AL methods is the cold start problem; when the number of previous observations is low (i.e. at the beginning of the AL process), the algorithm is less able to determine which observations would be informative. This issue may lead to the model sampling non-informative data, negatively impacting subsequent sampling decisions. One possible solution to cold start is to only start the AL process once sufficient initial observations have been collected. This, however, depends on determining the initial batch size required by AL methods to define informativeness accurately. Yang et al. [18] found that an initial batch of 20 observations is sufficient for AL methods to outperform random sampling. In this work, we use the same initial batch size of 20 observations.

An alternative approach to address the cold start issue is to incorporate data collected from other related models to provide additional observations at the beginning of the AL process. This approach, known as warm start [22], was investigated by [11] for engaged concentration detection and showed mixed results. Engaged concentration refers to the affective state associated with flow [23]. Karumbaiah and her colleagues developed models of affective states for new schools, building off of models previously developed for other schools. They found that when developing detectors for suburban schools, UncS and random sampling showed better results when warm start was incorporated, regardless of whether the warm start was conducted using data from a suburban school or urban school. However, the performance difference (AUC ROC) between LMMSE detectors with and without warm start was less than 0.02 points. This could have occurred because the data observed in previous schools was not informative enough to be used as the initial batch for warm starting detectors in a new target school. To address this limitation, we propose exploring the use of AL methods to sample a better initial batch for warm starting.

B. Role of Population in Affect Detection

Previous research has investigated the generalizability of affect detectors across schools in different regions [10], [11]. Ocumpaugh et al. [10] demonstrated that detectors are not necessarily generalizable among urban, suburban, and rural populations through three-fold population-level cross-validation. The largest decrease in performance occurred when detectors trained using data from urban and suburban schools were tested in rural schools. Karumbaiah et al. [11], focusing on detectors of engaged concentration within the same dataset, showed that combining samples from suburban and urban schools and ignoring data from rural schools can improve detector performance when tested in suburban schools compared to using data from suburban populations exclusively. However, using a warm start with data acquired from suburban schools reduced the performance of detectors when tested in urban schools compared to using data from urban students exclusively.

The issue of generalizability in affect detection has also been investigated by considering factors such as ethnicity, race, gender, and culture. National and cultural differences have been associated with differences in the emergence, expression, and patterns of affect [24]-[26]. However, as far as we know, there is still a need to investigate the potential impact of national and cultural differences on the performance of affect detectors. Additionally, the facial expressions correlated with learning have been shown to vary between men and women [27]. However, Bosch et al. [28] demonstrated that there is not a considerable decrease in affect detector performance when detectors are applied to learners of different race, ethnicity, or gender than they were trained on. By contrast, Whitehill et al. [29] found that detectors of engaged concentration decreased in performance when transferred to students of a different racial

group, although performance still remained substantially higher than chance. In summary, these results suggest that while generalization of affect detectors appears to be impacted by differences in urbanicity, it is less clearly affected by differences in other demographic variables, within-country.

III. METHODS

This section describes the dataset used in this work, the methodology employed for sequencing the data to emulate a realistic scenario, the sampling processes evaluated, and the experiments conducted to address our research questions.

A. Data

The dataset used in this study was obtained from past affect research involving the ASSISTments platform [30]. This platform facilitates the assignment of content by teachers, provides automated feedback and support for student responses, and provides extensive reports on student performance for teachers. This dataset consists of 3111 affective state observations from 471 students collected in mathematics classrooms from six urban, suburban, and rural middle schools in the United States. The observations were collected using the BROMP protocol [15] for collecting observations of student affect and engagement in classrooms. Multiple research groups have used this dataset to explore affect detection [10], [11], [18], [31]. The dataset is publicly available at http://tiny.cc/affectdata. The details of the number of students, samples, and urbanicity of each school are shown in Table I. The original dataset contains observations from 10 schools, but for this study, only urban and suburban schools were included. Rural schools were excluded due to their poor generalizability when used to train models for schools from a different population (see literature review above).

TABLE I. STUDENTS AND SAMPLES (OBSERVATIONS) BY SCHOOL

School Id	Urbanicity	# Students	# Samples
A	Urban	216	734
B	Suburban	27	68
C	Suburban	88	1599
D	Suburban	37	97

Table II displays the distribution of affective states for each school. The majority of the samples, 82%, represent engaged concentration. Some schools have limited samples involving confusion and frustration. We excluded affective state/school combinations where the target affective state occurred less than five times. For example, School B has only one sample of confusion and two of frustration. Thus, we excluded School B from our tests for detecting confusion and frustration. Additionally, we excluded observations labeled with more than one affective state (7 observations from school A and 3 observations from school C).

TABLE II. AFFECTIVE STATE DISTRIBUTION

School Id	# Eng. Conc	# Confusion	# Boredom	# Frustration
A	420 (57.2%)	57 (7.8%)	201 (27.4%)	68 (9.3%)
B	54 (79.4%)	1 (1.5%)	11 (16.2%)	2 (2.9%)
C	1536 (96.1%)	6 (0.4%)	47 (2.9%)	13 (0.8%)
D	71 (73.2%)	10 (10.3%)	11 (11.3%)	5 (5.2%)

B. Data Sequencing

As previously mentioned, the affective state of each sample was initially coded using BROMP. BROMP codes were collected by trained and certified human observers [15], who underwent a process of pre-field training, field training, and inter-rater reliability testing, and had to achieve inter-rater reliability over 0.6 with a certified trainer in order to themselves receive certification. Observations were made following a specific order of students selected based on their location inside their classroom. Each student was observed for up to 20 seconds to determine the (first) affective state that he or she was experiencing; the coder continued to the next student as soon as the affective state was clearly observed. Therefore, although there is no available information on more than one student at the same instant, the sample of two consecutive students is only separated by approximately 20 seconds. We assume that samples of 2 to 5 consecutive students are comparable because their affective state is relatively likely to remain stable within this time window (Botelho et al. [32], studying affect duration within the same learning system). The data was sequenced to reflect the real-time observation process, following the original order of observations made by the coders for each school. This sequencing approach emulates the temporal nature of a real observation scenario and the potential choices a coder could face when deciding which student to observe next rather than assuming that the observer can select a student who is tens or hundreds of positions ahead of the current observation.

C. Data Sampling for Active Learning

We employed UncS and LMMSE to conduct the AL processes for detecting each affective state. Each AL process starts by defining the sequence of schools (not individual samples within a school) to be considered. For example, if a detector for school B uses data from schools A, C, and D, one valid sequence for the AL process could be to observe school A first, followed by Schools C and D. Depending on the first school of the sequence, the initial batch for selecting the samples of the subsequent school will vary. To eliminate any bias due to the order of schools, all possible permutations of the sequence of schools were considered and averaged. Once the sequence of schools of the current epoch is established, the AL process begins with an initial batch of 20 samples from the first school. This batch size was selected based on results obtained by [18]. For School B, due to the limited number of observations, we used an initial batch of 10 samples.

For each iteration of the AL process, the algorithm selects one observation from the next five possible samples based on the criteria specific to each AL methodology. The sample window then shifts to consider the next five samples in the next iteration. For example, if the current pool is students 1, 2, 3, 4, and 5, and student 4 is selected, the next pool will be students 5, 6, 7, 8,

and 9. This process continues until the model has acquired 50 additional observations for the school (or until there is no more data). We select this number based on the results of tests using 20 to 50 additional observations. When reducing the number of additional observations, the performance of the models decreased. After obtaining the 50 observations of the current school, the next school in the sequence is considered. The data sampled in the previous school is employed as a warm start for sampling observations of the new school. This entire process is repeated for each school in the sequence and all the possible permutations.

D. Experiments

The AL processes described above were employed to separately detect concentration, confusion, boredom, and frustration. Before conducting the analysis, we established a baseline for the performance indicators of the detectors. The baseline was determined through a four-fold school-level cross-validation that tested the performance of the binary classification model for each affective state in each school.

In this work, we use the Area Under the Receiver Operating Characteristic Curve (AUC ROC; AUC for short) as the performance indicator. AUC is an appropriate choice, given the high imbalance between the proportions of each affective state [33]. To establish a reasonably-effective baseline for the performance indicators, we identified which affective states could be detected in each school using all the available data from the three remaining schools. Affective state/school combinations where no machine learning (ML) method achieved an AUC higher than 0.55 were excluded because the detector's performance, regardless of the sampling method used, would be close to chance.

Two control conditions were established. The first control condition involves a random selection from the same interval we use for the AL methods considering the subsequent five available samples. The second control condition consists of observing all the available samples in the order defined by the observation sequence of each school. It is important to note that this second control condition does not involve using all data, as that could be expected to achieve the maximum performance for the data set (and does not help us analyze the benefits of active sampling for the same sample size). This condition uses the first 50 available samples after the initial batch following the real observation sequence of each school.

We conducted our analysis by comparing our four experimental and control conditions to select the students to observe (UncS, LMMSE, Random selection, and taking the first 50 observations). For each sampling process, all the possible permutations of the order of schools were considered. We utilized Logistic Regression (LR), Random Forests (RF), Multi-Layer Perceptron (MLP), and Support Vector Machine (SVM) to train binary classifiers for each affective state.

We also used the AL methods to sample observations from the same school where the detector is implemented, to understand the degree to which generalizability across schools impacts the results. Studying this also helps us see the degree to which negative results may be due to cold start, where the methods do not have enough data to determine which sample can be more informative. Warm start methods have been proposed to mitigate this problem [11]. Therefore, we explored employing the four sampling methods (two experimental, two control) as an alternative to provide a better warm start for a new sampling process in a target school to improve its detectors by using data from its students. In practice, of course, it would make sense to conduct a warm start for a new category of schools rather than attempting to collect data for each new school, unless a different data collection method such as self-report or teacher observations was used.

To conduct this experiment, we sampled 100 additional observations from the target school. The remaining data, which was not used for sampling the 100 additional observations, was set aside as the test set. Since only school C had detectors that performed better than chance and had sufficient data for both the sampling and test sets, we restricted our testing to this school.

IV. RESULTS

This section presents the results of the experiments that compared the two AL methods and the two control conditions for sampling. We begin by establishing the baseline performance of the detectors and identifying the ML techniques with the highest AUC scores for each affective state in each school. Then, we compare the AUC scores of the four sampling conditions without including data from the target school in the training process. Finally, we also compare the four sampling processes when data from the target school is considered.

A. Baseline

Before employing AL, we start our study by identifying which affective states can be successfully detected in each of the four schools using models trained on the other schools. Table III shows the baseline results (obtained, as discussed above, using four-fold school-level cross-validation). The ML technique with the highest performance is shown in parentheses. In all cases, either SVM or LR classifiers obtained the highest AUC. In general, the highest performance was seen for suburban school C, where AUC over 0.6 was obtained for all affective states. In suburban school D, three of four affective states obtained AUC over 0.6 and the fourth still remained over chance (0.55). For suburban school B, better than chance performance (AUC > 0.5) was seen for engaged concentration, boredom, and frustration. Finally, model performance was close to chance for all affective states for the urban school A.

Based on these results which suggest an upper bound on performance, our analyses on the four sampling alternatives below do not test on school A. Also based on these results, within school B we only test the concentration and boredom detectors. The affective state detector/school cases which are tested in the following experiments are shown in bold in Table III.

TABLE III. SCHOOL-LEVEL CROSS-VALIDATION AUC SCORES USING ALL AVAILABLE DATA. DETECTOR/SCHOOL COMBINATIONS THAT ARE TESTED IN THE FOLLOWING EXPERIMENTS ARE SHOWN IN BOLD

Test School	Eng. Conc	Confusion	Boredom	Frustration
A	0.54 (LR)	0.41 (LR)	0.53 (SVM)	0.52 (SVM)
B	**0.63 (LR)**	0.55 (LR)	**0.65 (LR)**	0.73 (SVM)
C	**0.64 (LR)**	**0.63 (LR)**	**0.63 (LR)**	**0.65 (LR)**
D	**0.70 (LR)**	**0.64 (LR)**	**0.72 (SVM)**	**0.56 (LR)**

All other schools' data led to successful model performance in each target school in at least some cases. In specific, though other schools' models performed poorly for school A, we include the data from school A for training the detectors for the remaining schools (B, C, and D), since omitting school A's data led to poorer performance (Mean change in AUC>0.05).

B. Comparing Sampling Methods in an Unobserved School

We test the four sampling methods on the selected schools for detecting each affective state. Table IV shows the mean and standard deviation of the AUC scores at the end of the sampling process, calculated among all the possible permutations for the order of the schools to observe. We consider 6 permutations, which is the number of possible alternatives to organize the three training schools, for each test school. Only the ML technique with the highest AUC score is reported for each detector in Table IV.

Table IV shows the results of the experiments, indicating that LMMSE was the best performing sampling method for detecting engaged concentration, confusion, and frustration. In the case of engaged concentration, LMMSE outperformed the control conditions for all schools, achieving an AUC score 0.015 higher. The largest differences were seen in the confusion detector for school C and the frustration detector for school D, where LMMSE achieved an AUC score 0.03 higher than the control conditions. However, it is important to note that the sample size in both cases was small, with only 6 and 5 data points in the test set for confusion and frustration, respectively.

No other cases had differences larger than 0.03, and the direction of differences was not consistent. For instance, the control conditions outperformed the active learning methods for boredom detection. Furthermore, when comparing the active learning results in Table IV to the baseline shown in Table III, it can be observed that concentration and frustration detectors achieve similar, and in some cases higher, performance using a reduced number of samples. This similarity in performance suggests that there is not a substantial difference between concentration and frustration detectors trained using all available samples and those trained in this experiment using a reduced number of observations. On the other hand, there is a clear reduction in the performance of the boredom detectors for all schools and the confusion detector for school D when using a reduced number of samples.

TABLE IV. MEAN (SD) OF AUC FOR ALL SAMPLING METHODS

Affective State	School	LMMSE	UncS	All	Random
Eng. Conc.	B (LR)	0.680 (0.008)	0.671 (0.015)	0.661 (<0.001)	0.661 (0.001)
	C (LR)	0.630 (0.001)	0.558 (0.008)	0.614 (<0.001)	0.612 (0.001)
	D (LR)	0.724 (0.016)	0.717 (0.008)	0.705 (<0.001)	0.708 (0.006)
Conf.	C (LR)	0.651 (0.001)	0.622 (0.003)	0.620 (<0.001)	0.548 (0.055)
	D (LR)	0.592 (0.005)	0.587 (0.001)	0.586 (<0.001)	0.586 (0.002)
Bore.	B (LR)	0.610 (0.003)	0.614 (0.022)	0.617 (<0.001)	0.614 (0.005)
	C (SVM)	0.540 (0.001)	0.560 (0.002)	0.558 (<0.001)	0.556 (0.003)
	D (SVM)	0.584 (0.013)	0.645 (0.026)	0.687 (0.001)	0.668 (0.050)
Frust.	C (LR)	0.634 (0.003)	0.634 (0.004)	0.626 (<0.001)	0.633 (0.006)
	D (LR)	0.625 (0.032)	0.579 (0.009)	0.589 (<0.001)	0.590 (0.003)

The best-performing ML techniques remain LR and SVM, while RF and MLP did not achieve the highest AUC in any cases. In contrast with the results shown in Table III, SVM became the ML technique with the highest performance for boredom detection in school C when using a reduced number of samples.

Table IV shows the final performance of each model; however, what occurs in cases when less data is available? We show how model performance shifts as additional data becomes available in Fig 1. For comparing the four sampling methods, we only consider the ML technique with the highest performance for each school/affective state test set combination, as indicated in Table IV. This figure shows as the data available increases, all of the sampling methods provide a similar improvement in detector performance. Table IV shows that LMMSE performs slightly better than the other sampling methods for concentration, confusion, and frustration detectors. There are no additional differences between sampling methods over time, although UncS sampling generally performs slightly worse than the control conditions. However, the difference with the other sampling methods is no more than ten samples.

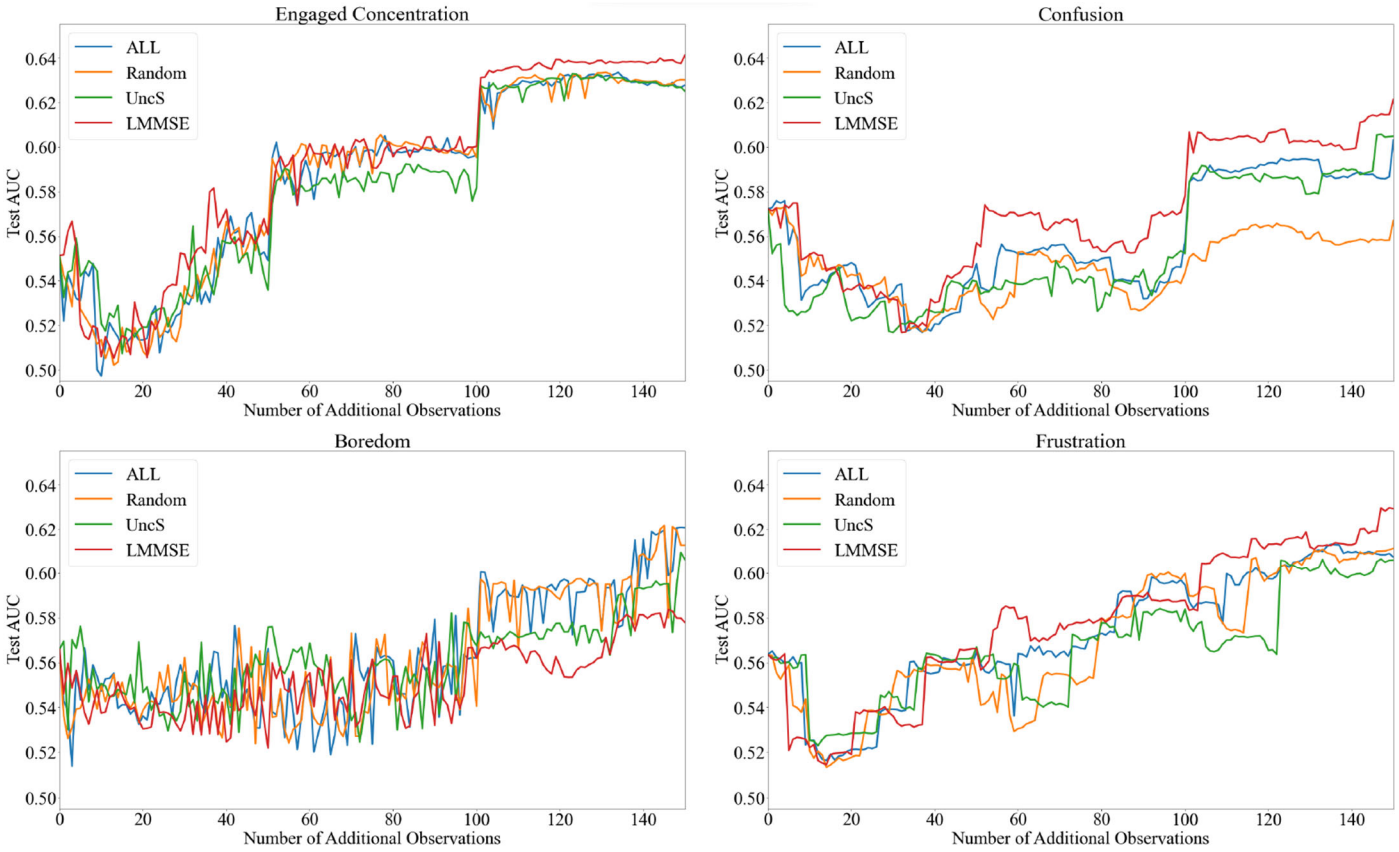

Fig. 1. Comparing AUC scores of the four sampling methods for detecting engaged concentration, confusion, boredom and frustration.

C. *Warm Start: Including Data of the Target School*

The results of the experiment in section IV-B suggest that AL, specifically LMMSE, promotes a slight improvement in the detectors' performance in entirely new schools. Other work has suggested that AL may be promising for warm start applications, where limited data from the target school is used to fine-tune a model initially trained on other schools [11], [18]. To test this, we also incorporate 100 additional observations sampled from the sequence of the target school after finishing the sampling processes analyzed in the previous experiment. This allows us to investigate whether the AL methods can lead to a better warm start and improve the performance of the detectors. For this experiment, we selected school C as the target school as it is the only one with sufficient data. These 100 additional observations are obtained from the original sequence of samples that considers the temporal nature of the data. Therefore, all the data considered for sampling and training the detectors temporally precedes the samples used for testing.

Fig 2 shows the AUC scores of the four sampling conditions. The data from the target school are incorporated after observation 150. Using AL methods for sampling again does not appear to yield a substantial improvement in the detection performance. Similarly to the previous results, LMMSE showed a slightly better performance for the concentration detector when considering data from the target school. The largest difference is still observed in the confusion detection, although, as mentioned before, the sample size for this affective state in school C is small (6 data points). On the other hand, LMMSE performed worse than the other sampling methods for boredom detection. No substantial differences were observed for frustration detection.

Additionally, with the exception of boredom, there is no substantial improvement in the AUC scores after observation 150. This result suggests that the observations from the other schools can be enough to reach the peak of the detectors' performance in the target school. Adding observations from students of the target school only improved the detector's performance for boredom.

V. Discussion and Conclusion

This paper investigates the use of AL methods for sampling data for affect detectors in a realistic setting. While Yang et al. [18] found that AL improved affect detection, they acknowledged that in a realistic scenario that considers the temporal nature of observations, the impact of these methods might be reduced. Our results show that AL methods still lead to a slight improvement in the performance of some detectors in a realistic scenario. However, there is not substantial improvement for any detector, and in some cases, such as boredom detection, observing all students is more appropriate. These mixed results suggest that further research with a larger number of samples and schools is needed to determine if AL methods could be beneficial for affect detectors, under realistic conditions, and whether the results depend on the

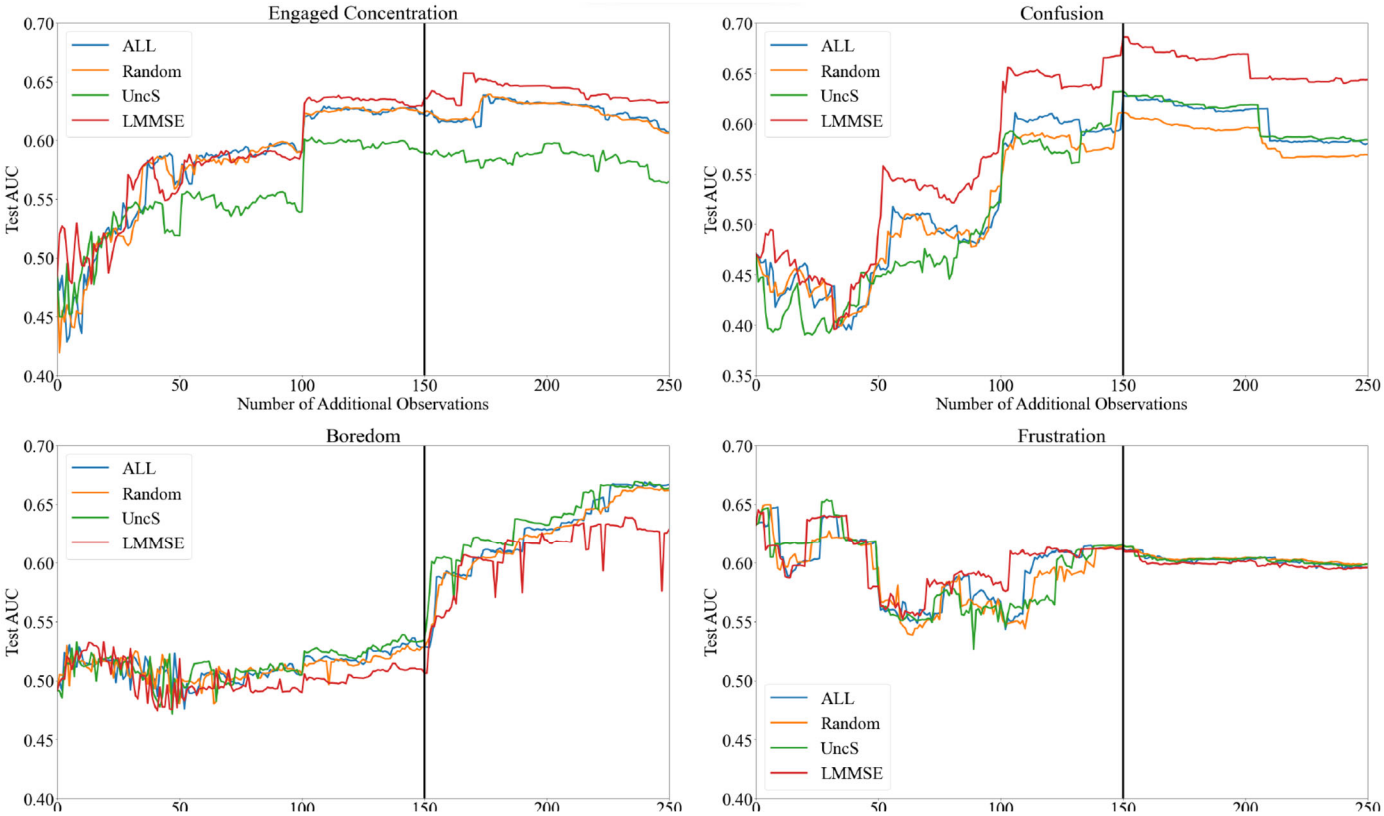

Fig. 2. AL process used as Warm Start for Affect Detection. Data of the target school is considered after observation 150.

schools being observed and test schools being demographically/regionally similar.

Going forward, studies should be conducted to further examine the realistic scenario proposed in this paper, including a larger number of urban and rural schools. Given the high imbalance between the number of samples of each affective state, studying affect in contexts with more frequent boredom, confusion, and frustration would allow a better comparison between different sampling methods for detecting affective states other than engaged concentration. Additionally, AL methods (mainly LMMSE) can be implemented in classrooms to collect data in real time and compare the performance of detectors trained using these data with detectors developed using data acquired with current existing protocols such as BROMP. To ensure a fair comparison, we suggest using all sampling methods in the same school, starting data collection with current protocols, and using this data to warm start the AL methods. In this way, all sampling methods used for comparison will have collected data for the same students, and the cold start issue of AL methods will be mitigated. In addition to this comparison, future work should also investigate the impact of window size. In this study, a window size of 5 was used due to the nature of the original dataset. Although this window size may be suitable for real-life implementation, it may not be optimal for every situation. Therefore, there is still a need to explore how to adjust this parameter.

Finally, further research should be conducted to determine which data is best to warm start AL methods in a new target school. As we observed, in some cases, adding data from the target school does not increase the performance achieved by detectors trained exclusively using data from previously observed schools. This result suggests that data from other schools can be informative for the target one. However, if these samples are not informative, using them as a warm start for AL would confuse the algorithm, leading to inappropriate data collection that would lead to poorer performance. Conducting larger-scale research of this nature will help identify when additional samples of the target school should be collected before employing AL methods.

ETHICAL IMPACT STATEMENT

In this study, we employed a publicly available dataset obtained with an observation protocol previously reviewed and approved by a research ethics committee, and that has been widely used in previous literature. One potential issue of employing this dataset is the generalizability of the results. We are not considering data from rural schools or schools in other regions or countries. In the future, it would be beneficial to conduct larger-scale research that demonstrates validity of the findings across wider populations.

ACKNOWLEDGMENT

This paper was written with the assistance of ChatGPT, which was used to improve the writing clarity and grammar of first

drafts written by humans. All outputs were reviewed and modified by two human authors prior to submission. Valdemar Švábenský conducted a software review to validate correctness and match to paper. This work was funded by the National Science Foundation through grant IIS-1917545. 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.

REFERENCES

[1] P. V. Rouast, M. T. P. Adam and R. Chiong, "Deep Learning for Human Affect Recognition: Insights and New Developments," in *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 524-543, 1 April-June 2021,

[2] N. Henderson, W. Min, J. Rowe and J. Lester, "Enhancing Multimodal Affect Recognition with Multi-Task Affective Dynamics Modeling," *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, Nara, Japan, 2021, pp. 1-8.

[3] J. DeFalco, J. Rowe, L. Paquette, V. Georgoulas-Sherry, K. Brawner, B. Mott, R. Baker, and J. Lester, "Detecting and addressing frustration in a serious game for military training," Int. J. Artif. Intell. Educ., vol. 28, no. 2, pp. 152–193, 2018.

[4] M. Song, Z. Yang, A. Baird, E. Parada-Cabaleiro, Z. Zhang, Z. Zhao, and B. Schuller, "Audiovisual analysis for recognising frustration during game-play: introducing the multimodal game frustration database," in Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction, 2019, pp. 517–523.

[5] N. Henderson, A. Emerson, J. Rowe and J. Lester, "Improving Sensor-Based Affect Detection with Multimodal Data Imputation," *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, Cambridge, UK, 2019, pp. 669-675

[6] S. Hutt, J. F. Grafsgaard, and S. K. D'Mello, 'Time to scale: Generalizable affect detection for tens of thousands of students across an entire school year', in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–14.

[7] J. L. Sabourin and J. C. Lester. Affect and engagement in game-based learning environments. IEEE Transactions on Affective Computing, 5(1):45-56, Jan. 2014.

[8] M. Wixon, I. Arroyo, K. Muldner, W. Burleson, D. Rai, and B. Woolf. The opportunities and limitations of scaling up sensor-free affect detection. In Proc. International Conference on Educational Data Mining, pages 145-152, July 2014.

[9] L. Paquette *et al.*, 'Sensor-Free or Sensor-Full: A Comparison of Data Modalities in Multi-Channel Affect Detection', *International Educational Data Mining Society*, 2016.

[10] J. Ocumpaugh, R. Baker, S. Gowda, N. Heffernan, and C. Heffernan, 'Population validity for Educational Data Mining models: A case study in affect detection', *British Journal of Educational Technology*, vol. 45, no. 3, pp. 487–501, 2014.

[11] S. Karumbaiah, A. Lan, S. Nagpal, R. S. Baker, A. Botelho, and N. Heffernan, 'Using Past Data to Warm Start Active Machine Learning: Does Context Matter?', in *LAK21: 11th International Learning Analytics and Knowledge Conference*, 2021, pp. 151–160.

[12] F. Hollands and I. Bakir. Efficiency of automated detectors of learner engagement and affect compared with traditional observation methods. Technical report, New York, NY: Center for Benefit-Cost Studies of Education, Teachers College, Columbia University, Aug. 2015.

[13] K. S. Kassam and W. B. Mendes, 'The effects of measuring emotion: Physiological reactions to emotional situations depend on whether someone is asking', *PloS one*, vol. 8, no. 6, pp. 649–659, 2013.

[14] S. C. Widen and J. A. Russell, 'Children acquire emotion categories gradually', *Cognitive development*, vol. 23, no. 2, pp. 29–312, 2008.

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

[16] S. Roy and A. Etemad, "Analysis of Semi-Supervised Methods for Facial Expression Recognition," *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, Nara, Japan, 2022, pp. 1-8.

[17] Settles., B. (2012). Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning,6(1):1–114.

[18] T.-Y. Yang, R. S. Baker, C. Studer, N. Heffernan, and A. S. Lan, 'Active learning for student affect detection', in Proceedings of the 12th International Conference on Educational Data Mining, EDM 2019, Montréal, Canada, July 2-5, 2019. International Educational Data Mining Society (IEDMS) 2019, 2019, pp. 208–217.

[19] Y. Yang and M. Loog, 'A benchmark and comparison of active learning for logistic regression', *Pattern Recognition*, vol. 83, pp. 401–415, 2018.

[20] A. S. Lan, M. Chiang, and C. Studer. An estimation and analysis framework for the Rasch model. In Proc. International Conference on Machine Learning, pp 2889-2897, July 2018.

[21] A. S. Lan, M. Chiang, and C. Studer. Linearized binary regression. In Proc. Conference on Information Sciences and Systems, pages 1-6, Mar. 2018.

[22] K. Konyushkova, R. Sznitman, and P. Fua, 'Learning active learning from data', *Advances in neural information processing systems*, vol. 30, 2017.

[23] M. Csikszentmihalyi, 'Flow and the psychology of discovery and invention', *HarperPerennial, New York*, vol. 39, pp. 1–16, 1997.

[24] S. Karumbaiah, R. B. Baker, J. Ocumpaugh, and A. Andres, 'A re-analysis and synthesis of data on affect dynamics in learning', *IEEE Transactions on Affective Computing*, 2021.

[25] Kitayama, S., Markus, H. R., & Kurokawa, M. (2000). Culture, emotion, and well-being: Good feelings in Japan and the United States. Cognition & Emotion, 14(1), 93-124.

[26] Tsai, J., Levenson, R. (1997). Cultural influences on emotional responding: Chinese Am. & European Am. dating couples during interpersonal conflict. J. Cross-Cultural Psych., 28(5), 600-25

[27] A. K. Vail, J. F. Grafsgaard, K. E. Boyer, E. N. Wiebe, and J. C. Lester, 'Gender differences in facial expressions of affect during learning', in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, 2016, pp. 65–73.

[28] N. Bosch, S. K. D'mello, J. Ocumpaugh, R. S. Baker, and V. Shute, 'Using video to automatically detect learner affect in computer-enabled classrooms', *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 6, no. 2, pp. 1–26, 2016.

[29] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, 'The faces of engagement: Automatic recognition of student engagement from facial expressions', *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 86–98, 2014.

[30] N. T. Heffernan and C. L. Heffernan, "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, vol. 24, no. 4, pp. 470–497, 2014.

[31] A. F. Botelho, R. S. Baker, and N. T. Heffernan, "Improving sensor-free affect detection using deep learning"', in International conference on artificial intelligence in education, 2017, pp. 40–51.

[32] A. F. Botelho, R. S. Baker, J. Ocumpaugh, and N. T. Heffernan, 'Studying Affect Dynamics and Chronometry Using Sensor-Free Detectors', *International Educational Data Mining Society*, 2018

[33] L. A. Jeni, J. F. Cohn and F. De La Torre, "Facing Imbalanced Data--Recommendations for the Use of Performance Metrics," *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, Geneva, Switzerland, 2013, pp. 245-25