Using Past Data to Warm Start Active Machine Learning: Does Context Matter?

Shamya Karumbaiah
University of Pennsylvania, United States

Andrew Lan
University of Massachusetts Amherst, United States

Sachit Nagpal
New York University, United States

Ryan S. Baker
University of Pennsylvania, United States

Anthony Botelho
Worcester Polytechnic Institute, United States

Neil Heffernan
Worcester Polytechnic Institute, United States

ABSTRACT

Despite the abundance of data generated from students' activities in virtual learning environments, the use of supervised machine learning in learning analytics is limited by the availability of labeled data, which can be difficult to collect for complex educational constructs. In a previous study, a subfield of machine learning called Active Learning (AL) was explored to improve the data labeling efficiency. AL trains a model and uses it, in parallel, to choose the next data sample to get labeled from a human expert. Due to the complexity of educational constructs and data, AL has suffered from the cold-start problem where the model does not have access to sufficient data yet to choose the best next sample to learn from. In this paper, we explore the use of past data to warm start the AL training process. We also critically examine the implications of differing contexts (urbanicity) in which the past data was collected. To this end, we use authentic affect labels collected through human observations in middle school mathematics classrooms to simulate the development of AL-based detectors of engaged concentration. We experiment with two AL methods (uncertainty sampling, L-MMSE) and random sampling for data selection. Our results suggest that using past data to warm start AL training could be effective for some methods based on the target population's urbanicity. We provide recommendations on the data selection method and the quantity of past data to use when warm starting AL training in the urban and suburban schools.

CCS CONCEPTS

• **Active learning**; • **User characteristics**; • **Education**;

KEYWORDS

Affect detection, Model generalization, Warm start, Urbanicity

ACM Reference Format:
Shamya Karumbaiah, Andrew Lan, Sachit Nagpal, Ryan S. Baker, Anthony Botelho, and Neil Heffernan. 2021. Using Past Data to Warm Start Active Machine Learning: Does Context Matter?. In *LAK21: 11th International Learning Analytics and Knowledge Conference (LAK21), April 12–16, 2021, Irvine, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3448139.3448154

LAK21, April 12–16, 2021, Irvine, CA, USA

ACM ISBN 978-1-4503-8935-8/21/04…$15.00
https://doi.org/10.1145/3448139.3448154

1 INTRODUCTION

New forms of digital data captured in various learning settings have made it possible to build meaningful models to understand and optimize learning [18]. Interaction logs, and sensors, among others, make it easy to generate abundant data from students' learning activity [2]. Yet, in many cases, the modeling effort is limited by the availability of ground truth labels of complex educational constructs (which are used as target variables in supervised ML) related to student affect, cognition, behavior, and other sociocultural factors [18]. In some cases, it is possible to attain labels at little to no cost (e.g., college enrollment, in-system behavior, and test performance). But for other constructs, the success in using ML depends on human annotation that is time-consuming, expensive, and sometimes difficult depending on the pedagogical context and the complexity of the construct being modeled. For instance, if the data labels are collected in authentic settings like physical classrooms, then fieldwork opportunities are limited in time, resource-intensive, and involve several tedious tasks such as background verification of the observers, approvals from the school administration and institutional review boards, and obtaining written consent from students and parents. Even video replay coding still takes a substantial amount of time per label [25]. Thus, despite having an abundance of student activity data overall, the limitations in collecting human labeled data are pushing us to find new ways to develop better performing ML models with a smaller amount of annotated data [31].

A potential solution lies in a subfield of ML called Active Learning (AL) that tries to learn a good model from fewer data samples by letting the ML model choose the data it trains from – thus, focusing the labeling efforts on a smaller subset of carefully selected data samples [28]. This subfield is not to be confused with the instructional approach of active learning in education [3]. In this paper, AL refers to the ML-based data selection algorithms aimed at improving the data labeling efficiency (discussed further in Section 2.1). AL works by training a model and choosing data iteratively: in each iteration, it first uses the current model to choose which data point to use next (i.e., on which data point to query for a human-generated label), and then uses the label to update the model. In contrast to simple classification problems (e.g., classifying apples from oranges in an image), complex settings like education pose some challenges to the adoption of AL. First, the labels could be highly subjective (e.g., self-reports of student emotions). Second, the

data can be highly noisy (e.g., video data from a physical classroom). Third, the input feature set could be large (e.g., hundreds of features summarizing student activity in a virtual learning environment). Thus, the complexity involved in the ML tasks in application fields like education may require AL to seek significantly more samples to reach reasonable quality. As such, AL has found limited use in LA thus far, especially in the cold start situation, where the model doesn't have access to sufficient data yet [28].

In this paper, we investigate the use of past data on the same construct to overcome the cold-start problem when using AL. Using past data to warm start the AL training process, even for the same construct, may not be straightforward given the diversity in the student population [12]. Considerable research shows that demographic factors are often related to differences in educational outcomes [6]. Thus, we also examine whether the differing context of the past data used to warm start AL has an implication in building a model in the target population. This is important because LA models need to ensure population validity as they attempt to meet the needs of all students [22]. Given the feasibility challenges around collecting learning data in schools, the first data that is collected may in many cases come from convenience samples of middle-class students (see discussion in [14]) or another highly accessible student population where it is easy to collect labeled data. Since AL follows a greedy approach to optimize data collection, using data from a dominant student population could drive the model training process for a different population of learners to a suboptimal solution - a biased model. Hence, it is necessary to critically investigate the role of differing contexts of the past data if we want to use them to overcome the cold-start problem. As we still don't know what a "population" is [1], we focus our experiments in this study on one contextual dimension of urbanicity [cf. 22] to examine the use of past data from urban and suburban schools to warm start AL training in a school from the other context. Thus, our primary research questions in this study are -

- Does using past data help warm start the AL training process effectively?
- How does the urbanicity of the past data impact the effectiveness of the warm start process?
- How much past data from a different urbanicity is appropriate to use while warm starting the AL training process?

To this end, we use authentic affect labels collected through human observations in middle school mathematics classrooms to simulate the development of AL-based detectors of engaged concentration (described further in section 2), a common affective state among students. We experiment with two AL methods and one non-AL method (random sampling) for data selection. Our results suggest that using past data to warm start AL training could be effective for some methods. We also see that the urbanicity of the past data matters. We provide recommendations on the data selection method and the quantity of past data to use when warm starting AL training in the urban and suburban schools.

1.1 Contributions

Our primary contribution to AL research with education data is the critical analysis of the use of past data to overcome the cold start problem of AL training with complex constructs. More importantly, we show that mismatches in the urbanicity of the past data (and possibly other demographic dimensions) could be detrimental to effective model training in some cases.

2 BACKGROUND

The next subsections provide a brief introduction to AL methodology, its use in label data collection for affect detection, and the importance of studying differing contexts in this paradigm.

2.1 Active Learning Algorithms

Supervised learning is a machine learning task that involves learning a function that maps the input (a set of feature values for a data point) to a predefined output (a target label of the data point). A commonly used function is a classifier that maps the input to a set of categories (class labels). In the typical supervised learning setup, all labeled data is collected before model development starts and available at training time for the model to learn from. On the other hand, in an AL setting, data collection and model training occur concurrently. The label collection process is iterative since all or a relevant subset of the training data collected thus far is available in real-time to make a choice on which point will get labeled next. AL methods are used in a scenario where there is limited opportunity to obtain labeled data – typically, when one can only selectively label a small subset of an otherwise abundant unlabeled data. The goal of AL algorithms is to enable training a high-quality classifier with fewer data samples by selecting those that are the most informative to the classifier. Thus, as the training of a model progresses, the AL algorithm aims to select the next data point to obtain a label for, such that it will be the most informative for the current model and hopefully lead to the largest improvement in its predictive power [28]. Several metrics of informativeness have been explored in AL research such as entropy (or observation uncertainty) [19], expected error reduction [26], expected variance reduction [32], and model change [5]. A suite of algorithms has also shown promising results with a range of classifiers from logistic regression [30] to deep convolutional neural networks [27]. In this paper, we compare the following three approaches (two AL methods and one non-AL method) that have previously been applied to affect detection [31].

2.1.1 Uncertainty Sampling (UncS). Uncertainty sampling is the simplest and most commonly used AL method and has been shown to achieve comparable or even better performance than other more sophisticated AL methods on real-world data [30]. It uses the prediction entropy of the model's predictive distribution over each possible class label to quantize the informativeness of each data point. Therefore, in each iteration, it takes the current model and predict the label distribution of each unlabeled data point and selects the one that the model is the least certain of, i.e., the data point that has the highest predictive entropy under the current model.

2.1.2 The (L-MMSE)-based method. One limitation of the UncS method is that the accuracy of its notion of data informativeness, i.e., model uncertainty, is highly associated with the quality of the current model. Therefore, when the model only has access to limited data, this estimate of uncertainty may not be accurate. The Linear Minimum Mean Square Error (L-MMSE) Estimator, first proposed in [16, 17], provides a set of closed-form approximations of the

estimation error (a proxy of model uncertainty), which is shown to be highly accurate when the number of data points is small. Therefore, the L-MMSE-based AL method [31] selects the next data point as the one that leads to the maximum reduction of the MSE. Roughly speaking, it looks at how similar each unlabeled data point (behavior) is to previous labeled data points the model has seen, which means we want to label the next data point that looks the most like an outlier. It is shown to mostly outperform UncS for student affect detection, especially when the number of data points is small [31].

2.1.3 Random Sampling (Random). We also use a third, non-AL method for data selection, which is simply to randomly select an unlabeled data point from all possible points.

2.2 Student Affect Detection

Affective computing is an important area of interest in LA due to the close connection between a student's affect and their learning and experience. Affect has been shown to correlate with important educational constructs like self-efficacy [20], motivation [24], and learning [7]. Accordingly, affect-sensitive interventions have been designed in virtual learning environments to improve students' learning [8], and overall experience [10]. Thus, several research studies in the past decade have focused on building good quality affect detectors using physical and physiological sensors [21], and interaction log data [10] - the latter being the more affordable, less intrusive, and scalable option. Sensor-free affect detectors are classifiers categorizing a set of student interaction features into a predefined set of student affective states such as confusion, frustration, boredom, and engaged concentration. The features are distilled from the interaction log data which is easily available in most virtual learning environments. However, the affect labels required for supervised training involves a labor-intensive data collection process.

One commonly used approach to collect labels for affect is through field observations in a real classroom by certified expert coders. A frequently-used technique for classroom observations is the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP; [23]) - an affect coding protocol wherein students are observed by certified coders in a round-robin fashion. Each observation lasts up to 20 seconds. The affective state labels are boredom, confusion, frustration, and engaged concentration. During coding, some observations are labeled "NA", corresponding to the cases where i) the student could not be observed, ii) their affective state was unclear to the observer, or iii) they were in an affective state other than the states being coded. Following common practice in other detector development work [e.g. 22], we do not use these "NA" cases in our analysis.

2.3 Active Learning for Affect Detection

In a typical BROMP-based affect label collection, the students are observed in a pre-defined order. Thus, it is likely that the observers will miss opportunities to observe more informative cases. This could be an inefficient use of the limited time of the expert coders in an already short time window of fieldwork. Finding efficient ways to collect affect data is necessary as there are several constraints to conducting fieldwork in real classrooms (discussed in Section 1) along with the limited availability of the certified BROMP coders. AL provides an adaptive method to collect affect labels by directing the observers' attention to the more informative cases. For the first time in 2019, Yang and colleagues [31] investigated the use of AL to collect higher quality data for training affect detectors with fewer data samples. We will be adopting a similar experimental protocol as [31], which is the standard in most AL studies (elaborated in Section 4.1).

In addition to experimenting with the existing AL methods, Yang and colleagues [31] also proposed the new method of L-MMSE, which appears to be particularly suited for the affect data collection setting where the data is small and noisy. Their results suggest that, when compared to other AL methods, L-MMSE leads to efficient modeling i.e., high-quality sensor-free affect detectors with fewer labeled data. By letting the model pick the next observation to learn from, the AL models were able to reach a desirable performance with as little as 70 observations which would translate to around 20 minutes of field observations with BROMP. This could tremendously reduce the burden on human labeling. However, before we adopt this methodology in our data collection practices, we also need to critically examine any possible biases that the model could have picked while greedily choosing the next best data sample to get labeled.

2.4 Role of Student Population in Affect Detection

In the previous study, the empirical analysis was conducted on combined data from multiple schools in different urbanicities (urban, suburban, and rural). In the current study, we split the data based on the urbanicity to assess any potential discrepancies in using models trained on one population to test in another population. This is important because student demographics are known to influence several aspects of affect [13]. Differences in culture are known to influence variation in beliefs and personal dispositions towards emotional expression and moderation [29], and the frequency and emergence of certain affective states [15]. A recent study synthesizing results across multiple affect datasets showed that affective patterns seem to differ based on the country in which the data was collected (US versus Philippines; [13]). We chose to explore urbanicity as a contextual dimension in this study because past work suggests that affect detectors do not always transfer well between urbanicity categories [22]. In this study, we want to examine if this result holds true in the AL paradigm, especially when using data from different urbanicity to warm start the AL training.

3 DATA

We use a previously collected dataset from ASSISTments [4]– a computer based learning platform which allows teachers to assign content and monitor student performance while supplying students with immediate correctness feedback and on-demand supports in the form of hint messages and scaffolding [9]. The affect data was collected in middle school mathematics classrooms using BROMP (see Section 2.2). The dataset consists of 2511 affect observations for 367 students. For each observation, a set of 92 features is extracted from the log of the student's interactions with practice problems within ASSISTments. These features summarize student

within- and across-problem behaviors in the 20-second interval of the affect observation, such as the number of hints they seek, time spent on solving problems, the accuracy of responses, etc. In this paper, we will study affect detection for engaged concentration, a binary classification model (engaged concentration vs other affective states). Consistent with past studies in other learning systems [11], engaged concentration has the highest incidence among all the affective states in this dataset. However, the rate differs significantly between the two urbanicities - 93.85% among suburban students and 56.06% among urban students. Affect data were collected in schools located in northeastern US - 1772 observations from 153 students in three suburban schools and 755 observations from 222 students in one urban school [22]. All the schools are non-charter and non-magnet public schools. The original dataset also had three rural schools that we do not include in the current analysis due to data quality issues; past research with ASSISTments data in a similar context reported that affect detection models (without AL) generalize better between suburban and urban students than rural students [22]. In this paper, we would like to investigate if this property holds true between urban and suburban data when data from students in one urbanicity is used to warm start the AL-driven affect detector training process for students in the other urbanicity.

4 ANALYSIS AND RESULTS

In this section, we examine the effectiveness of using past data in the initial batch to warm start the AL training process. In addition, we investigate the impact of using a mismatching student population in the initial batch used for model development for differing initial batch sizes. We present the results for the experiments we conducted to answer the following research questions -

- Does using past data help warm start the AL training process effectively?
- How does the urbanicity of the past data impact the effectiveness of the warm start process?
- How much of the past data from a different urbanicity is appropriate to use while warm starting the AL training process?

4.1 Active Learning Experimental Design

As detailed in Section 2.1, we use three different approaches: 1) the linear minimum mean square error (L-MMSE)-based method [31], 2) uncertainty sampling (UncS), and 3) random sampling (Random). The first two are AL methods. We perform a train-validation-test split of the full dataset (70%-10%-20% ratio) at the student level i.e., the instances corresponding to an individual student are all in a single split. We use a simple logistic regression-based affect detector in all the experiments since it performs well and makes it possible to use all AL methods [31]; other more advanced affect detectors are not compatible with many AL methods. We use the standard area under the receiver operating characteristic curve (AUC) as the performance metric. The first step of the AL training is to select an initial batch from the training set. The initial batch size is a variable of interest in this research and we vary it based on each individual experiment (details in subsections below). Using the observations and affect labels in the initial batch, a base classifier is trained. We train our affect detectors using gradient descent and stop training as soon as performance on the validation set stops improving. Next, we select a data point for each AL method from the remaining training set based on its feature values. The model is re-trained with the selected data point, and the AUC is calculated using the test set. This process is repeated for 70 new observations. Each experiment is repeated 100 times by splitting the dataset randomly into train-validation-test sets and randomly selecting an initial batch from the training set each time. The plots presented in the results section contain the average AUC across the 100 random splits.

4.2 Baselines (Experiment Set #0)

We report baseline performances on two testing setups: i) test set drawn from only urban students, and ii) test set drawn from only suburban students, with three training setups each: i) training set drawn from only urban students, ii) training set drawn from only suburban students, and iii) training set drawn from both urban and suburban students) - leading to a total of six train-test setups. To ensure a fair comparison across the urbanicities, we match the randomly chosen test sets across the 100 runs for all three training setups. We ran the following two sets of baseline models -

- *Full data model (without AL)* - For each of the six train-test setups mentioned above, a logistic regression model is trained using all the data in the training set. This baseline represents the typical scenario where we collect the full data without using AL to optimize the label data collection process. It is the best-case scenario in terms of having all the data that can practically be collected given the resource constraints.
- *AL without warm start* - AL algorithms without a warm start. This baseline represents the scenario where we choose to disregard any past data we have collected for the same construct. Instead, we collect new data using AL. We run this for all the three approaches - L-MMSE, US, and random.

In Table 1, Figure 1, and Figure 2 we present the baseline performances (measured by average AUC) on the held-out test sets of a logistic regression model on full data (without AL) and AL models without a warm start.

Full data model (without AL). When compared to the within-urbanicity performance, the between-urbanicity transfer is relatively better for suburban -> urban (Table 1, exp# 5 vs exp# 4) than urban -> suburban (Table 1, exp# 2 vs exp #1). For testing with suburban data, the model trained on urban data (different urbanicity) has 0.045 AUC value less (0.583 vs 0.628) than the one trained on suburban data (same urbanicity). In contrast, for testing with urban data, the model trained on suburban data (different urbanicity) has 0.006 AUC value more (0.663 vs 0.657) than the one trained on urban data (same urbanicity). The better transferability of the model trained with suburban data to urban data could be due to the higher diversity within suburban data (from three different schools) as compared to urban data (from a single school). The three suburban schools may also vary in terms of the teacher practices and use of the system. The best performing model for the suburban data is the one trained on the combined dataset (urban+suburban). By contrast, the best performing model on the urban data is the one trained on the suburban dataset. We see that the models trained on the full data transfer well between urbanicities for testing on

Table 1: Baseline test performances (measured by mean AUC across 100 random splits) for the six train-test setups of the logistic regression model with full data (without AL) and AL training without warm start. We report the performance of the AL algorithms without a warm start for L-MMSE, US, and random at the last iteration of AL training for.

Exp#	Training Set	Test Set	Full data model (without AL)	AL Without Warm Start		
				L-MMSE	UncS	Random
Testing on Suburban Students						
1	Suburban	Suburban	0.628	0.617	0.607	0.578
2	Urban	Suburban	0.583	0.548	0.581	0.585
3	Urban + Suburban	Suburban	0.652	0.649	0.631	0.599
Testing on Urban Students						
4	Urban	Urban	0.657	0.617	0.645	0.652
5	Suburban	Urban	0.663	0.583	0.582	0.607
6	Urban + Suburban	Urban	0.638	0.626	0.639	0.638

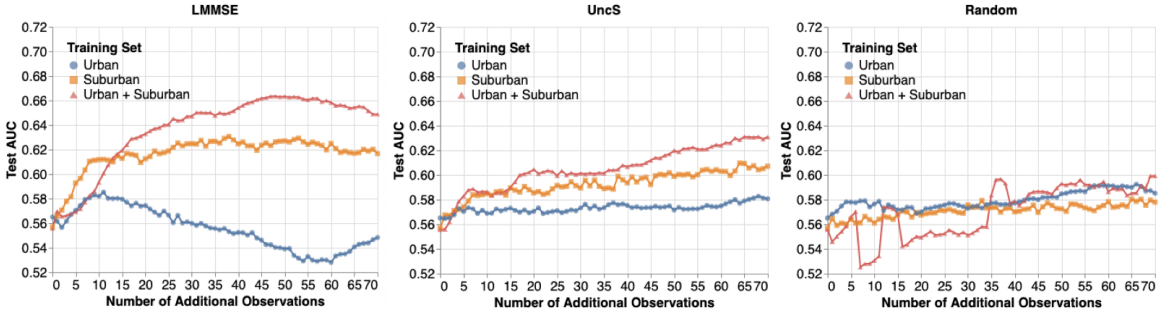

Figure 1: Comparing cross-validated performances of the L-MMSE, US, and random AL algorithms trained on different training sets without warm start and tested on suburban data.

urban students. The models tested on urban data have similar performances with a 0.025 difference in AUC value between the best (0.663) and worst-performing (0.638) models. The AUC value difference between the best (0.583) and worst-performing (0.652) models is higher for the suburban test data at 0.069.

AL Without Warm Start. The AL models have a relatively larger difference in performances across the six train-test setups. The urbanicity mismatch in training and testing sets hurts test performance in all the three approaches. A model trained using the data from a single urbanicity does not transfer well when tested on the other urbanicity. For testing with suburban data (Table 1, exp# 1-3), training with combined data leads to a better performance for all three approaches. This observation is consistent with the pattern in the models trained on the full data without AL. The best performing AL model (L-MMSE) has only 0.003 less AUC than the best model trained on full data without AL (0.649 vs 0.652). Note that random sampling does slightly better than the full data model when a model trained using urban data is used with a suburban test set (0.585 vs 0.583). One possible explanation is that random sampling learns from a smaller subset (50 samples) of training data from a mismatched urbanicity as compared to the model trained on the full urban data without AL (744 samples) – potentially reducing its generalizability to suburban students.

For testing with urban data (Table 1, exp# 4-6), a similar pattern of better performance with combined data is seen only for L-MMSE. For UncS and random, the best performing model is the one trained on urban data (same urbanicity). In contrast to the full data model without AL, the model trained on suburban data leads to a worse performance in urban data for all the three approaches. Among the three approaches, random sampling has the best performance at an AUC of 0.652 which is only 0.011 less than the full data model without AL.

4.3 Within-Urbanicity warm start (Experiment Set #1)

In this set of experiments, we warm start the AL training process with data based on schools from the same urbanicity as the test school. We train and test AL algorithms on a single school but take the initial batch data from other schools in the same urbanicity. Since our data is from one urban school and three suburban schools, we will run these experiments only with the suburban data. Also, we have 1598 observations from one suburban school and only 103 and 68 observations from the other two suburban schools, which is not sufficient to have diverse enough random splits between training and test sets across the 100 runs of AL training – leading to unreliable evaluation. Hence, we will be running the AL models only on the school with 1598 observations with initial batches drawn from the other two schools. For comparison, we also report results for the experiments where the initial batch is drawn from the same school (no warm start) for the same test sets. Since the total number of students in the smallest suburban school is 68, we limit the initial

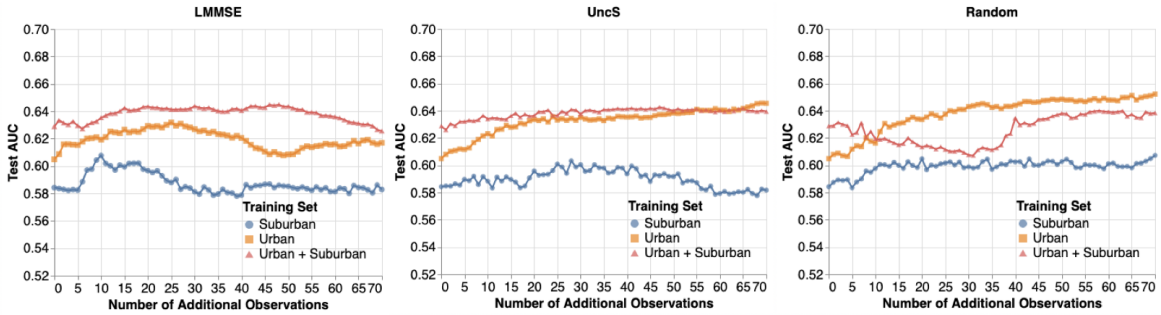

Figure 2: Comparing cross-validated performances of the L-MMSE, US, and random AL algorithms trained on different training sets without warm start and tested on urban data.

batch size to 68 across all experiments for consistency. In cases where the school has more than 68 samples, we randomly sample 68 observations for the initial batch. These experiments represent the scenario where we choose to use the past data collected in a different school(s) from the same urbanicity to warm start the AL training. We run these experiments to answer our first research question on the effectiveness of using past data as an initial batch to warm start the AL training. The results should help us decide if we want to use past data from a similar student population (schools in the same urbanicity) to warm start the AL training process.

The three plots in Figure 3 present the results for the three approaches after a within-urbanicity warm start for a single suburban school (described in Section 4.3). Each plot has one line for the same school warm start (blue circles - School A), two lines for warm start with two other suburban schools (orange squares - School B and red triangles- School C), and one line for warm start with the combined data from the two other suburban schools (green pluses - Schools A&C). As one would expect, the same school warm start (blue circles) generally has a better performance for all the three approaches - close to the full data model without AL for suburban data (Table 1, exp# 1). The AL training starts at a high AUC value, potentially because the initial batch data (with 68 data samples) are all from the same school. The AUC doesn't improve with training and stays the same throughout (almost a straight line at AUC value around 0.62). This is not surprising as the AL algorithms are expected to do well with less data and we had enough data points from the same population in the initial batch to start with.

Despite the other two schools (B and C) being in the same urbanicity (suburban) as school A, using the data from these two schools to warm start AL training in school A leads to strikingly different results. The initial batch from one of the two schools (School B; orange squares) leads to a model performance that is consistently better than the other school (School C; red triangles) for all the three approaches. For the UncS approach to AL, warm start with school B quickly improves the AUC value and surpasses the same-school warm start with only 5 additional training samples from the target school. With random sampling, the AUC improvement is relatively slower as compared to the UncS approach. Nevertheless, the steady improvement eventually converges with the same-school warm start at the end of AL training with only 50 samples from school A as compared to 118 samples from school A for same-school warm start (68 in the initial batch + 50 during AL training). With L-MMSE, however, the improvement to AUC value saturates after 5 additional samples from school A, starts to dip slowly with more samples leading to a close to chance AUC value (~0.50) at the end of AL training. This raises concerns on using past data from a different school to warm start L-MMSE model training, even when the school is from the same urbanicity.

The data from the other school (school C; red triangles) in the initial batch brings down the model performance severely (down by 0.30 AUC). In all the three approaches, the AUC value at the end of the training is below chance (<0.50), making the trained model inapplicable to the target population. The performance gradually improves as the AL training progresses for UncS and random sampling but fails to recover for L-MMSE. This observation raises questions on the robustness of the AL algorithms to out-of-context data during training - how quickly can L-MMSE recover when the new samples from the target population are introduced?

As one would expect, the initial batch with the combined data (green pluses) from the two other schools leads to a close-to-average performance compared to the two schools separately. With more observations, the combined data initial batch starts to improve steadily for the UncS and random (not for L-MMSE), and reaches a similar performance as the same-school warm start. The final performance for the UncS and random sampling is better than the AL without a warm start and is similar to the full data model without AL for suburban data (Table 1, exp# 1). Thus, there is some evidence supporting the use of combined data from multiple schools in the same urbanicity to warm start the AL training.

In summary, the within urbanicity warm start experiments suggest that not all schools in the same urbanicity have a similar effect when used to warm start the AL training process. Using a random sample from the combined data could be a better choice. More research on the similarities and differences between the three schools on other demographic variables is needed to better understand the warm start process's differing implications. The UncS approach to AL and random sampling are seen to be more robust than L-MMSE in improving the model performance when new data from the target population is introduced. Overall, we recommend using past data from the same urbanicity, preferably from multiple schools, in warm starting some data collection approaches (UncS and random sampling, not L-MMSE).

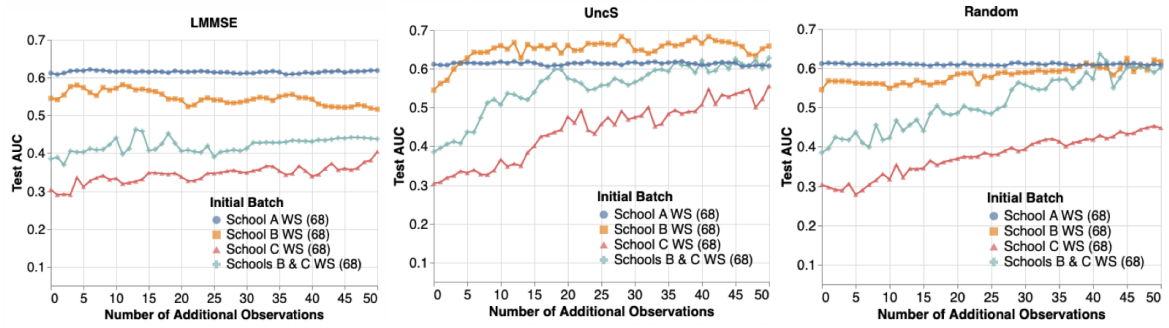

Figure 3: Comparing cross-validated performances of the L-MMSE, US, and random AL algorithms on a single suburban school (School A) with warm start data from the same and other suburban schools. In parenthesis in the legend are the initial batch sizes. WS = Warm Start.

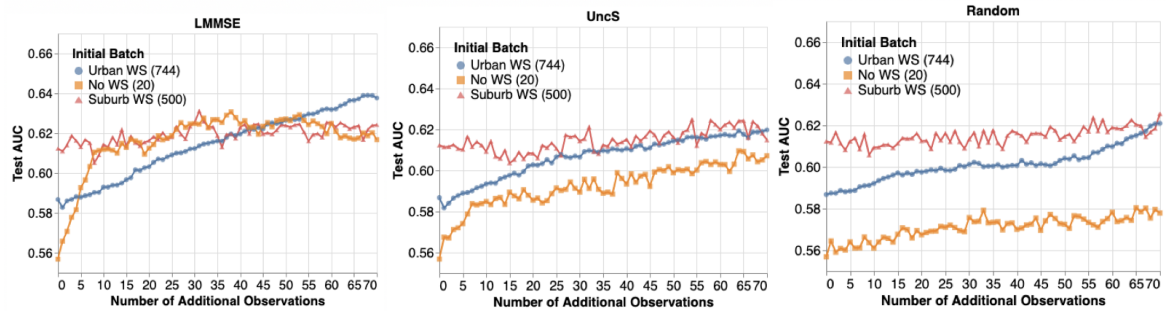

Figure 4: Comparing cross-validated performances of the L-MMSE, US, and random AL algorithms on suburban student data with no warm start, warm start with past urban student data, and warm start with past suburban student data. In parenthesis in the legend are the initial batch sizes. WS = Warm Start.

4.4 Between-urbanicity warm start (Experiment Set #2)

Our next set of experiments are similar to the experiment set #1, except the initial batch comes from schools in a different urbanicity. Specifically, we train and test AL algorithms on student data from urban schools but draw the initial batch from suburban schools. Likewise, we run the experiment with training and test sets drawn from suburban data and initial batch from urban data. In these experiments, we take all the past data of the chosen urbanicity in the initial batch (the size is varied in the next subsection). These experiments correspond to the scenario where we choose to use the past data collected in schools from a different urbanicity to warm start the AL training. For comparison, we also report results for the within urbanicity warm start for the same test sets. We run these experiments to answer our second research question on the impact of urbanicity in using past data to warm start the AL training. The results should help us decide if we want to use past data from a different student population (in this case urbanicity) to warm start the AL training process.

The results for the between-urbanicity warm start is presented for two cases - a) test on suburban data (Figure 4), and b) test on urban data (Figure 5). In the case of the test on suburban data (Figure 4), using past data from urban schools to warm start the AL training (blue circles) leads to a better model performance when compared to the model trained with no warm start (orange squares). This observation is true for all three approaches. In fact, all the three approaches start at an AUC greater than 0.58 (above the chance value) with an initial batch drawn from the past urban data. This is in contrast with like the last set of experiments where using past data from a different suburban school led to a performance below chance. The use of past data from urban schools gives a head-start of close to 0.02 AUC for the AL training with suburban data. Although the L-MMSE model without a warm start catches up after 20 new observations, the improvement in AUC saturates and stays at around 0.62, while the training with a warm start climbs up to 0.64. For UncS and random sampling, the progress in performance for both with and without a warm start is more gradual. Relative to the US, the gap widens further for random sampling as the training progresses and reaches 0.04 (0.58 without warm start vs. 0.62 with a warm start). As expected, a warm start using 500 samples from the same urbanicity (red triangles) leads to a peak performance right from the beginning of the AL training and shows little effect due to the new observations. This is not surprising because the AL algorithms are expected to do well with less data and we had enough data points (500) from the same population in the initial batch to start with. Unlike AL without a warm start, the between-urbanicity warm start catches up to this peak performance in all the three approaches - even exceeding it in the case of L-MMSE. L-MMSE

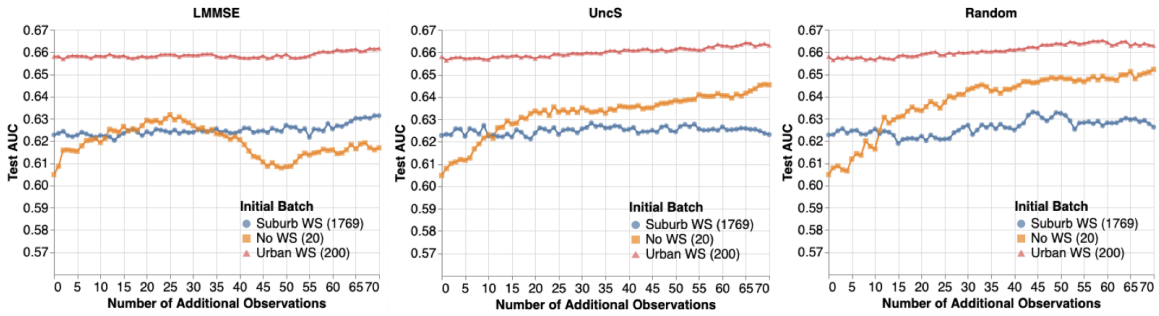

Figure 5: Comparing cross-validated performances of the L-MMSE, US, and random algorithms on urban student data with no warm start, warm start with past suburban student data, and warm start with past urban student data. In parenthesis in the legend are the initial batch sizes. WS = Warm Start.

with between-urbanicity warm start also manages to come close to the full data model without AL for suburban data, while random sampling only exceeds the AL without a warm start (Table 1, exp# 3). Overall, there is some evidence that using past data from urban schools effectively warm starts AL training in suburban schools.

The results are not as clear when testing on urban data (Figure 5). In comparison to AL training without a warm start (orange squares), the AL training with suburban data in the initial batch (blue circles) performs better for L-MMSE and not for the UncS and random sampling. In all the three cases, the AL training starts at a higher AUC value (0.625 vs 0.605) with the suburban data in the initial batch but remains constant as new data samples are introduced. One possible explanation is that the initial batch data from a different urbanicity (suburban) outnumbers the additional samples from the new population (urban) and the model fails to improve its predictive power on the target population. After around 12 new samples, the model without a warm start exceeds the warm start model in its performance. However, as the training progresses, the L-MMSE performance for no warm start starts to decline, while the performance of UncS and random sampling rises steadily. Both UncS and random sampling models without a warm start exceed the full data model without AL (Table 1, exp# 6). In contrast, the performance of all the three models with between-urbanicity warm start doesn't meet both the baselines (AL without warm start and full data model without AL). Within-urbanicity warm start with urban data consistently leads to peak performance in all three models (red triangles). Overall, there is some evidence that using past data from suburban schools could be detrimental in warm starting AL training in urban schools.

4.5 Vary the initial batch size for the warm start (Experiment set #3)

In the experiment set #2, we took all the data from a chosen urbanicity as the initial batch. In this experiment set, we try to answer our third research question on how much past data from different urbanicity is appropriate to warm start the AL algorithms. We repeat the same experiments as before, varying the initial batch size stepwise. These experiments represent the scenario where we may have a large amount of past data from a different urbanicity and need to find out how much data should be used to warm start the AL training process. The results should help us decide what amount of past data from the same or different student population (in this case urbanicity) is effective to warm start the AL training process.

5 DISCUSSION

AL is a promising subfield of ML that can be used in LA to increase the efficiency of the label collection process that is time-consuming and requires extensive human effort in many cases. Model training and label data collection go hand-in-hand in AL, with the model iteratively choosing the most informative next data point to get labeled by a human. One of the challenges in the adoption of AL for education data is the cold-start problem when it is hard to accurately estimate the informativeness of a data point due to the lack of data at the early stages of the AL process [31]. In this paper, we have explored the use of past data to overcome the cold-start problem seen for AL methods, along with the potential implications of differing student populations in the past data. Using an existing student affect dataset collected through human observations in middle school mathematics classrooms, we experimentally tested three training approaches (UncS, L-MMSE, Random) for the sensor-free detector of engaged concentration, studying performance within and across urbanicities (urban, suburban).

5.1 Summary of results

We conducted four sets of experiments to answer our research questions: how effective is AL i) without warm start, ii) using within-urbanicity warm start, iii) using between-urbanicity warm start, and iv) using varying batch sizes to warm start. Our results suggest the following. First, for all three approaches, training a model completely using data from a different urbanicity without warm start results in low detection accuracy in the target population [cf. 22]. Second, not all schools in the same urbanicity have a similar effect when used to warm start the AL training process. Using a random sample from the combined data (across suburban schools) could be a better choice when data from multiple schools are available. Third, using past data from urban schools effectively warm starts AL training in suburban schools. In contrast, using past data from suburban schools is detrimental to warm starting AL training in urban schools. One possible explanation is that the size of the suburban data is too large (1772 observations) compared to the urban

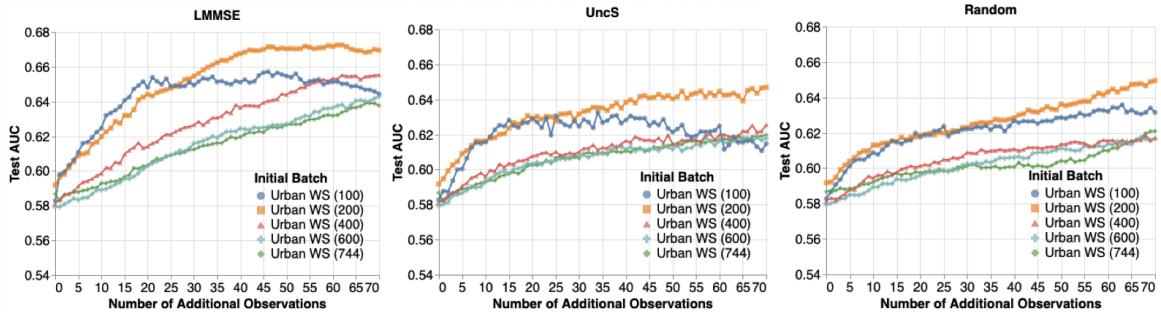

Figure 6: Comparing cross-validated performances of the L-MMSE, US, and random AL algorithms on suburban student data with varying amounts of past urban student data for the warm start. In parenthesis in the legend are the varying initial batch sizes. WS = Warm Start.

data (755 observations); a trained model on suburban data may overfit to suburban students and loses generalizability to a different student population. Lastly, using the right (often small) amount of past data can effectively warm start AL training if urbanicities do not match. When comparing the three approaches, we found that UncS and random sampling are more robust than L-MMSE in improving the model performance when new data from the target population is introduced during within-urbanicity warm start. With between-urbanicity warm start using a smaller batch size, it could help to switch from AL algorithms to random sampling after collecting some data points in the target population. In summary, with the experiments in this paper, we have shown that using past data to warm start some AL methods could be effective in training a good quality detector of engaged concentration (performance comparable to a detector trained using all data) with a fewer number of samples in some conditions and not so effective in others.

5.2 Implications for research

Our primary contribution to AL research with education data is the critical analysis of the use of past data to overcome the cold start problem of AL training with complex constructs. More importantly, we show that mismatches in the urbanicity of the past data (and possibly other demographic dimensions) could be detrimental to effective model training in some cases. As our AL modeling effort advances and finds innovative ways to improve model training with little data, it becomes essential to critically examine which data samples we are using. The need to consider human diversity in predictive modeling is becoming more apparent in LA research as the community moves to implement analytics solutions at a larger scale. If we aim to serve all students, we need to ensure the population validity of the models we build. This does not necessarily imply that all models must be within-population (and, indeed, we do not entirely know what a population is [cf. 20]) – our findings suggest that there are better and worse ways to use data from other populations when building a model.

5.3 Implications for practice

Yang and colleagues [31] discuss the implications for data collection procedures when using AL in real classrooms. They present a brief design of a three-component system – i) an interface to record human observations, ii) a training paradigm to build a detector, and iii) an active learning method that connects the labeling and model training processes (see [31] for more details). In addition, there should be a provision for the expert coders to ignore the suggestion made by AL and use their intuition to pick the most informative cases when necessary. This agency could be important, especially if the AL-based model is picking up some unknown biases and leading to a suboptimal model training. Differences between AL recommendations and expert choices could also be valuable in conducting a post-hoc analysis of an AL approach's functioning. Also, even within a single class, there could be student subgroups that may end being under-observed by the AL recommendation. It is important that we need to set up conditions in AL recommendation to pick samples that are representative of these subgroups. Partnering with teachers will be a useful direction for identifying important subgroups in a specific class or school. Such research-practice partnership can help mitigate potential biases in data selection.

Although this work focuses on classroom observations, we could extend it to other forms of label data collection such as self-reports, video coding, and text replay coding. With COVID-19 related school closures, we are currently exploring the use of AL in collecting labels through student self-reports. Since a student can be surveyed at any point in time, the observation window is not as strict as field observations. However, we must budget the surveys per student to not interfere with their learning or be too intrusive. Hence, our focus in using AL shifts from choosing which student to observe, to when and how often we survey each student. The data is likely to look different from classroom observations. It could have more missing data (e.g., student skips the survey) or be more noisy data (e.g., student responds incorrectly). In addition, the feature set for the AL algorithm will likely come from a longer time window when compared to being restricted to a single class period. Our next step is to collect some self-report data and conduct a similar analysis on warm starting AL training for the different student populations.

5.4 Limitations and future work

There are some limitations to our work presented in this paper. First, our experimental design does not consider the temporal nature of affect data collection in the real world. We choose the next most informative data point among all available data points in hindsight

after they are already collected, while in the actual observation session, only a subset of these students and only the temporally close data samples will be available for human observation. Second, we have only experimented with the detector for engaged concentration, which is the most common affective state in our dataset. This work needs to be replicated with other important but relatively rarer affective states like boredom, frustration, and confusion. Third, due to data quality issues, we could not include rural schools in this study. In general, further thought on categorizing urbanicity is warranted. Fourth, our mixed results on within-urbanicity warm start suggest that more research is need on the similarities and differences between the suburban schools on other demographic variables. Finally, we hope to explore more advanced AL methods to see if there are methods that respond better to the warm start condition than UncS and L-MMSE.

ACKNOWLEDGMENTS

Our thanks to the NSF (Division of Information & Intelligent Systems awards IIS-1917545 and IIS-1917713) for sponsoring this project, and our thanks to Brian Zylich for his support with the servers. Dr. Heffernan, Dr. Botelho and their team also want to thank multiple NSF grants (e.g., 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, 1535428, 1440753, 1316736, 1252297, 1109483, & DRL-1031398), as well as the US Department of Education for three different funding lines; a) the Institute for Education Sciences (e.g., IES R305A170137, R305A170243, R305A180401, R305A120125, R305A180401, & R305C100024), b) the Graduate Assistance in Areas of National Need program (e.g., P200A180088 & P200A150306), and c) the EIR. We also thank the Office of Naval Research (N00014-18-1-2768) and finally Schmidt Futures and another anonymous philanthropy.

REFERENCES

[1] Baker, R.S., Ogan, A.E., Madaio, M., Walker, E. (2019). Culture in Computer-Based Learning Systems: Challenges and Opportunities. *Computer-Based Learning in Context, 1* (1), 1-13.

[2] Blikstein, P., & Worsley, M. (2016). Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *Journal of Learning Analytics, 3* (2), 220-238.

[3] Bonwell. C. and Eison. J. (1991). *Active Learning: Creating Excitement in the Classroom.*Jossey-Bass.

[4] Botelho, A. F., Baker, R. S., and Heffernan, N. T. (2017). Improving sensor-free affect detection using deep learning. In Proc. *International Conference on Artificial Intelligence in Education*, pages 40–51.

[5] Cai, W., Zhang, Y., Zhang, Y., Zhou, S., Wang, W., Chen, Z., and Ding., C. (2017). Active learning for classification with maximum model change. *ACM Transactions on Information Systems*, 36(2):15.

[6] Childs, D. S. (2017). Effects of Math Identity and Learning Opportunities on Racial Differences in Math Engagement, Advanced Course-Taking, and STEM Aspiration. PhD Dissertation. Temple University.

[7] D'Mello, S., Person, N., Lehman, B. (2009). Antecedent-Consequent Relationships and Cyclical Patterns between Affective States and Problem Sol- ving Outcomes. In *International Conference on Artificial Intelligence in Education*, 57-64.

[8] DeFalco, J. A., Rowe, J. P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B. W., Baker, R. S., & Lester, J. C. (2018). Detecting and addressing frustration in a serious game for military training. *International Journal* of Artificial Intelligence in Education.

[9] Heffernan, N. T., & 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.*

[10] Karumbaiah, S., Lizarralde, R., Allessio, D., Woolf, B. P., Arroyo, I., & Wixon, N. (2017). Addressing Student Behavior and Affect with Empathy and Growth Mindset. *Proceedings of the 10th International Conference on Educational Data Mining.*

[11] Karumbaiah, S., Andres, J. M. A. L., Botelho, A. F., Baker, R. S., & Ocumpaugh, J. S. (2018). The Implications of a Subtle Difference in the Calculation of Affect Dynamics. *In 26th International Conference for Computers in Education.*

[12] Karumbaiah, S., Ocumpaugh, J., & Baker, R. S. (2019). The influence of school demographics on the relationship between students' help-seeking behavior and performance and motivational measures. *Educational Data Mining (EDM),* 4, 16.

[13] Karumbaiah, S., Baker, R. S., Ocumpaugh, J. & Andres, J. M. A. L. (2020). A Re-Analysis and Synthesis of Data on Affect Dynamics in Learning. Submitted.

[14] Kimble, G. A. (1987). The scientific value of undergraduate research participation. *American Psychologist,* 42(3), 267- 268.

[15] 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.

[16] Lan, A. S., Chiang, M., and Studer, C. (2018) An estimation and analysis framework for the Rasch model. In Proc. *International Conference on Machine Learning*, pages 2889–2897.

[17] Lan, A. S., Chiang, M., and Studer, C. (2018). Linearized binary regression. In *Proc. Conference on Information Sciences and Systems*, pages 1–6.

[18] Lang, C., Siemens, G., Wise, A., & Gasevic, D. (Eds.). (2017). *Handbook of learning analytics.* SOLAR, Society for Learning Analytics and Research.

[19] Lewis, D. D., and Gale, W. A. (1994) A sequential algorithm for training text classifiers. In Proc. *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12.

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

[21] Nye, B. D., Karumbaiah, S., Tokel, S. T., Core, M. G., Stratou, G., Auerbach, D., & Georgila, K. (2018). Engaging with the scenario: Affect and facial patterns from a scenario- based intelligent tutoring system. In *International Conference on Artificial Intelligence in Education* (pp. 352-366). Springer, Cham.

[22] Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for Educational Data Mining models: A case study in affect detection. *British Journal of Educational Technology*, 45(3), 487-501.

[23] Ocumpaugh, J. (2015). *Baker Rodrigo Ocumpaugh monitoring protocol (BROMP) 2.0 technical and training manual.*New York, NY and Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences, 60.

[24] Rodrigo, M. M. T., Anglo, E., Sugay, J., Baker, R. (2008). Use of unsupervised clustering to characterize learner behaviors and affective states while using an intelligent tutoring system. In *International Conference on Computers in Education*, 57-64.

[25] Rowe, E., Asbell-Clarke, J., Bardar, E., Almeda, M. V., Baker, R. S., Scruggs, R., & Gasca, S. (2020). Advancing Research in Game-Based Learning Assessment: Tools and Methods for Measuring Implicit Learning. In *Advancing Educational Research With Emerging Technology* (pp. 99-123). IGI Global.

[26] Roy, N., and McCallum, A. (2001). *Toward optimal active learning through Monte Carlo estimation of error reduction. In Proc.* International Conference on Machine Learning, pages 441–448.

[27] Sener, O., and Savarese, S. (2018). Active learning for convolutional neural networks: A core-set approach. In Proc. *International Conference on Learning Representations*, pages 1–13.

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

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

[30] Yang, Y., and Loog, M. (2016). A benchmark and comparison of active learning for logistic regression. *arXiv preprint* arXiv:1611.08618.

[31] Yang, T. Y., Baker, R. S., Studer, C., Heffernan, N., & Lan, A. S. (2019). Active Learning for Student Affect Detection. *International Educational Data Mining Society.*

[32] Yu, K., Bi, J., and Tresp, V. (2006) Active learning via transductive experimental design. In Proc. *International Conference on Machine Learning*, pages 1081–1088.