Early Prediction of Visitor Engagement in Science Museums with Multimodal Learning Analytics

Andrew Emerson, Nathan Henderson, Jonathan Rowe, Wookhee Min, Seung Lee, James Minogue, and James Lester

North Carolina State University, Raleigh, NC 27695, USA
{ajemerso, nlhender, jprowe, wmin, sylee, james_minogue, lester}@ncsu.edu

ABSTRACT

Modeling visitor engagement is a key challenge in informal learning environments, such as museums and science centers. Devising predictive models of visitor engagement that accurately forecast salient features of visitor behavior, such as dwell time, holds significant potential for enabling adaptive learning environments and visitor analytics for museums and science centers. In this paper, we introduce a multimodal early prediction approach to modeling visitor engagement with interactive science museum exhibits. We utilize multimodal sensor data—including eye gaze, facial expression, posture, and interaction log data—captured during visitor interactions with an interactive museum exhibit for environmental science education, to induce predictive models of visitor dwell time. We investigate machine learning techniques (random forest, support vector machine, Lasso regression, gradient boosting trees, and multi-layer perceptron) to induce multimodal predictive models of visitor engagement with data from 85 museum visitors. Results from a series of ablation experiments suggest that incorporating additional modalities into predictive models of visitor engagement improves model accuracy. In addition, the models show improved predictive performance over time, demonstrating that increasingly accurate predictions of visitor dwell time can be achieved as more evidence becomes available from visitor interactions with interactive science museum exhibits. These findings highlight the efficacy of multimodal data for modeling museum exhibit visitor engagement.

CCS CONCEPTS

• Applied computing → Education;
• Computing methodologies → Machine learning

ICMI '20, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3418890

KEYWORDS

Museum-Based Learning; Visitor Modeling; Multimodal Learning Analytics; Early Prediction

ACM Reference format:

Andrew Emerson, Nathan Henderson, Jonathan Rowe, Wookhee Min, Seung Lee, James Minogue, and James Lester. 2020. Early Prediction of Visitor Engagement in Science Museums with Multimodal Learning Analytics. In *2020 International Conference on Multimodal Interaction (ICMI'20), October 25-29, 2020, Utrecht, The Netherlands.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3382507.3418890

1 Introduction

Visitor engagement plays a central role in learning in informal environments, such as museums and science centers [20, 24]. It affects how visitors navigate through a museum, form interests and attitudes, and acquire knowledge about concepts and ideas presented in exhibits. As museums strive to better understand visitor engagement, a key challenge is to devise computational models that predict how visitors interact with exhibits. Recent work in multimodal learning analytics has demonstrated the ability to model and represent learner engagement in many contexts [5, 35], which introduces the opportunity to model visitor engagement in museums [19]. Computational models of visitor engagement in museums that leverage multimodal learning analytics could inform better understandings of patterns in visitor engagement by incorporating multi-channel data streams (e.g., facial expression, eye gaze, posture, interaction logs) captured by both physical hardware sensors and exhibit-specific software. Several studies have demonstrated the efficacy of incorporating multi-channel data from students' learning interactions in multimodal sensor systems that model student knowledge [43-45, 49] and engagement [8, 13, 14, 46] in both classroom and laboratory settings. However, limited work has investigated multimodal learning analytics within informal contexts, such as science museums.

Inducing computational models of visitor engagement in museums poses significant challenges. Museum exhibits are frequently designed to foster brief interactions as visitors freely explore museums, resulting in very short dwell times [15, 28, 29].

Modeling visitor engagement using relatively short segments of data calls for rich, multimodal representations of visitor behavior. Specifically, it is important to predict engagement early in a visitor's interaction with an exhibit because it creates potential opportunities to intervene to support and prolong engagement. Another challenge is the inherent difficulty of measuring engagement in museums; administering surveys and tests, or performing field observations, can be disruptive to visitor experiences and disrupt the natural flow of learning in museums [6]. To address these challenges, it is critical to devise computational models that leverage available data at early points within visitor interactions and make accurate predictions of visitor engagement. These models should accurately predict visitor engagement utilizing data that are captured using physical sensors and software that are minimally disruptive to learning.

In this paper, we introduce a multimodal learning analytics approach for predicting visitor engagement with interactive exhibits in science museums. This analysis uses data collected from visitor interactions with a game-based interactive museum exhibit about environmental sustainability, FUTURE WORLDS. We captured visitors' posture, facial expression, interaction trace logs, and eye gaze to model salient features for predicting visitor engagement. Leveraging machine learning techniques (random forest, support vector machine, lasso regression, gradient boosting, and multi-layer perceptron), we investigate the accuracy of predictions made at early points within visitor interactions by segmenting visitors' multimodal data into thirty-second intervals. Additionally, we compare the accuracy of machine learning-based models of visitor engagement induced with several different combinations of input modalities. These ablation experiments were conducted to evaluate predictive models that use data collected from sensors that are less disruptive to learning. Results show that incorporating more modalities produces higher model accuracy by the end of visitor interactions. Additionally, the predictive models improved over time and converged to accurate predictions at points prior to the end of visitors' learning interactions, indicating that increasingly accurate predictions of visitor dwell time are possible as more data becomes available from visitor interactions with interactive science museum exhibits. This is the first work to use multimodal data to induce early prediction models of visitor dwell time for museum exhibits. In addition, the paper contributes findings on the influence of different combinations of data channels in predicting visitor dwell times by investigating the impact of different data channels through a series of ablation experiments.

2 Related Work

Promoting visitor engagement with exhibits is central to creating meaningful learning experiences in museums [15]. While there has been a significant amount of work modeling learner engagement in formal educational settings (e.g., classrooms) and laboratories [22], limited work has investigated this potential in museums. Visitor engagement in museums is manifested in several ways. Low levels of engagement may appear as shallow interactions (or no interaction at all) with interactive exhibits, while higher levels of engagement could be indicated by longer dwell times, thoughtful behaviors, and expressions stemming from visitors engaged with the exhibit. In our study, we focus on predicting visitor dwell time, a manifestation of visitors' behavioral engagement, which has been previously examined with museum exhibits [6, 19, 26].

Multimodal learning analytics has received significant attention for its potential to create a comprehensive view of learners' actions and their internal states [7, 33]. By taking advantage of trace data extracted from different modalities, multimodal learning analytics can assess learner's affective and cognitive behaviors [34, 49]. Prior work in learning analytics has explored the use of several modalities for modeling learning outcomes and components of learner engagement. For example, eye gaze has been used to determine a learner's level of attention in a discussion by tracking the gaze direction in classroom [37]; gesture has been used to provide alternative ways of emphasizing during presentation [17]; posture has been used to model learner attention in a classroom setting [38]; facial expression has been widely used to devise computational models for recognizing learner's affective states [8]; and speech has been used to evaluate question-answering interactions between instructor and learners in a lecture setting [12]. A broad range of multimodal learning analytics approaches have been investigated to measure or understand learning processes, ranging from detecting and identifying learners' collaborative behaviors in learning contexts [18, 25, 40], assisting teachers with classroom orchestration [1, 2, 36], and understanding visitor engagement in science museums [19]. A significant portion of this prior work has investigated modalities in a setting where sensors are used to monitor all aspects of student learning. These sensors are sometimes wearable by the student, and they often require calibration to accurately monitor student characteristics while engaging with learning content. However, little work has compared the performance of multimodal predictive models that incorporate sets of modalities that are less disruptive to learning, which can inform learning environments that are designed to be more naturalistic for the students using them.

A key challenge in modeling visitor engagement is developing the ability to make robust early predictions, i.e., making consistently accurate predictions as early as possible. This is important for run-time settings, as it sets the stage for the creation of adaptive exhibits that tailor content and feedback to promote engagement by intervening before visitors leave an exhibit. This characteristic has significant potential to improve visitor engagement with museum exhibits and to enhance learning in museums and science centers. However, little work has investigated predictive models that leverage multimodal data to make accurate predictions at early points in learners' interactions. Prior work on early prediction in museums has investigated a pair of visitors' social behavior types to provide socially aware, adaptive support that addresses visitors' diverse needs [16, 27]. Early prediction has been examined in the context of game-based learning environments centering on predicting middle-grade students' engagement [48] and in the context of recognizing dynamically changing learning goals [31]. When combined,

multimodal data streams have demonstrated significant promise for early prediction, such as predicting kindergarteners' reading skills in two years using electrophysiological and functional MRI data [3], enhancing the predictive performance of goal recognition with gaze data [31], and improving early prediction of academic failure with sentiment analysis of text-based self-evaluated comments [50]. Utilizing multimodal data for learning analytics problems has been studied extensively. Our work makes the following novel contributions: (1) we conduct a series of ablation experiments on multimodal data channels to identify modalities that achieve the highest predictive performance for visitor engagement, and (2) we investigate the early prediction capacity of visitor engagement models with the goal of providing adaptive support as early as possible. Unlike most previous early prediction work that has focused on classification settings, where performance was often measured with convergence-based metrics [30], our work demonstrates the effectiveness of predictive models using multimodal regression models and how these models' performances improve over time.

3 FUTURE WORLDS Testbed Exhibit

To investigate visitor engagement in science museums, we instrumented a game-based museum exhibit, FUTURE WORLDS, with several physical hardware sensors and data-logging software to capture museum visitor interactions. FUTURE WORLDS was developed with the Unity game engine and runs on an interactive surface display to enable touch-based interactions and explorations of environmental sustainability within a virtual environment [41]. As visitors interact with the game, they learn about environmental sustainability by solving problems through investigations of the impacts of environmental decisions on a 3D simulated environment (Figure 1). The goal of this game is for the visitor to improve aspects of the virtual environment (e.g., the renewable energy portfolio of a virtual location). Visitors can touch, swipe, and tap the screen to change the virtual environmental states and test hypotheses about the impacts of environmental decisions. Visitors learn through textual interfaces about the aspects of the virtual environment (e.g., the region's electricity portfolio or a farm's waste management practices) and how these are affected by environmental decisions. Changes made to the environment are rendered in real-time, and the virtual environment is colored to provide visual feedback based on the decisions the visitor makes. The primary target audience for FUTURE WORLDS is learners ages 10-11, and the educational content has been designed for this audience. In prior work, pilot testing with students from several elementary schools in a science museum revealed that visitors interacting with FUTURE WORLDS improved knowledge of sustainability content and showed promising levels of engagement measured by researcher observation [41].

4 Multimodal Data Collection

We captured visitors' facial expression, body movement, and eye gaze in conjunction with game interaction trace data as visitors

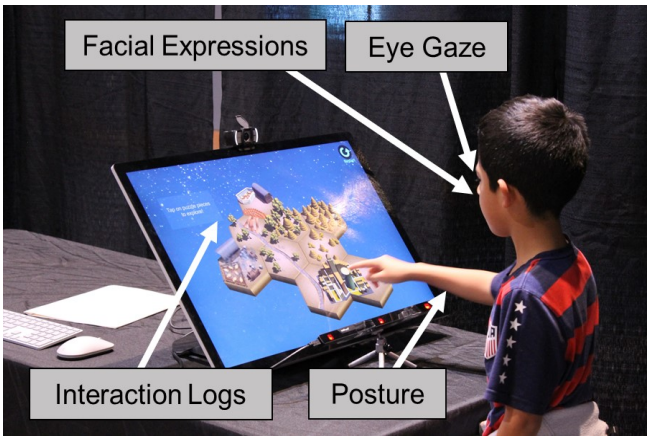

Figure 1: Museum visitor interacting with the FUTURE WORLDS exhibit

interacted with FUTURE WORLDS. The resulting data captured from these devices was subsequently analyzed to extract features for predicting visitor dwell time at the FUTURE WORLDS exhibit. As noted above, dwell time serves as the target variable for this prediction task.

4.1 Study Participants and Procedure

The predictive models are induced using data captured during three data collections with museum visitors interacting with the FUTURE WORLDS exhibit at the North Carolina Museum of Natural Sciences in Raleigh, North Carolina. The three groups each came from different socio-cultural backgrounds (e.g., race/ethnicity, urban vs. rural, language diversity). The schools each served student populations in which over 70% of students at each school are considered economically disadvantaged. The total number of participants in the study included 116 visitors between the ages of 10-11 (M = 10.4, SD = 0.57). Fourteen of the participants did not provide demographics data, leaving 47 female and 55 male participants. The racial makeup of the participants was 32.4% Hispanic or Latino, 21.6% African American, 11.8% American Indian, 8% Asian, 7.5% had mixed races, 3% Caucasian, and 15.7% preferred not to respond. Some participants were missing one or multiple modalities, resulting in the removal of several visitors' data from the final dataset for analyses. The final dataset consisted of multimodal data for 85 visitors, with demographics similar to the visitor population.

The exhibit setup and data collection procedure were as follows. Two instances of the exhibit were set up on tables in a temporary space divided by pipe and drape in a special exhibition room at the museum. Each station was instrumented with a tripod-mounted Kinect camera approximately 5-feet from the participant, a mounted eye tracker beneath the surface display, and an external Logitech webcam mounted on top of the display (Figure 1). Participants completed informed consent and pre-survey materials prior to the beginning of the study. Participants were introduced to the exhibit individually, and the multimodal sensors were calibrated at the beginning of the visitor's interaction with the exhibit with the assistance of a researcher. Visitors individually interacted with FUTURE WORLDS until they

completed the game or up to a maximum of approximately 12 minutes (M = 5.8, SD = 2.4, min = 1.8, max = 11.8). Visitor dwell times were captured by the game-based exhibit software. Participants were quietly observed by researchers, and after finishing with the game, moved to a different part of the special exhibition room to complete post-survey materials and participate in a short de-brief interview.

4.2 Multimodal Data Channels

We instrumented the FUTURE WORLDS exhibit with several sensors to track visitor engagement in real-time and to better understand visitors' interactions with the exhibit. The data streams collected by these sensors and logging software included posture, gesture, facial expression, eye gaze, and interaction trace logs, each of which is described in turn.

Body Movement. Utilizing physical behavior exhibited by learners has been demonstrated to be predictive of various states of affect within multimodal learning analytics [9, 23]. Visitors' posture and gesture movements were captured using the Microsoft Kinect for Windows v2. This motion-sensing camera tracks the movements and positions of 26 distinct vertices in 3D coordinate space, in addition to pixel data for both depth and camera sensors [10]. In this study, the Kinect was mounted to a tripod approximately 1.5 meters away from the exhibit and enabled posture tracking of visitors that characterizes different non-verbal behavioral signatures of visitor engagement.

Facial Expression. Facial expression provides an alternative perspective on learner emotion and engagement [32]. Facial movement data has been frequently used in multimodal learning analytics to produce models of learner affective states and learning [8]. The facial expression data collected in this study was from video recordings captured from an externally mounted Logitech C920 USB webcam. The captured video data is analyzed by OpenFace, an open-source toolkit to detect facial landmarks, estimate head pose, recognize facial action units (AUs), and estimate eye gaze [4]. OpenFace automatically detects and analyzes 17 distinct AUs for each participant face captured within the camera's field of view in real-time.

Interaction Trace Logs. The FUTURE WORLDS exhibit is equipped with software that enables the granular logging of learner interactions with the exhibit. These logs consist of sequential records (at the millisecond level) of taps and gestures on the multitouch surface, as well as learning events (e.g., requesting more information about a particular topic) and states of the underlying simulation, that arise during visitor experiences. This data can be utilized to investigate how learners explore and manipulate the underlying simulation provided by FUTURE WORLDS.

Eye Gaze. Gaze provides rich, task-based information that can inform models of learners' cognitive and affective states [9, 25]. Recent work has demonstrated the efficacy of using eye gaze for modeling learner interactions [43, 44]. We utilized a mounted Tobii EyeX eye-tracking sensor which uses near-infrared light to track eye movements and gaze points during visitor interactions with the interactive exhibit. We automatically identify in-game targets of visitor attention in FUTURE WORLDS using a gaze target-labeling module that processes eye tracking data using ray casting techniques. This module automatically tracks visitors' visual fixations on in-game objects and interface elements, yielding log events denoting the gaze target, timestamp, and duration of the fixation.

4.3 Multimodal Features

Using each of the modalities described above, we extracted features to serve as predictors of visitor dwell time. We engineered features for each modality that have proven to be valuable in prior multimodal learning analytics work [19].

To extract features that capture components of visitor body movement, we focused on four skeletal vertices tracked by the Microsoft Kinect motion sensor: *Head*, *SpineShoulder* (upper-back), *SpineMid* (mid-back), and *Neck*. These features were selected based on prior work on multimodal affect detection and engagement prediction with motion-tracking sensor data [9, 19, 21]. For each skeletal vertex, we calculated the minimum, maximum, and median position values, as well as the variance of each vertex. We used the four vertices to calculate two additional features used to represent changes in visitor posture. The first feature was the total posture change, which was generated by calculating the summative changes in each vertex's coordinates over a certain duration. The second feature was generated in a similar manner to the total posture movement but was instead calculated using the total change in the vertices' distance from the Kinect. In total, we distilled 18 posture-related features from the raw Kinect coordinate data.

To model visitor facial expressions, we processed facial AU data captured by OpenFace. We calculated the duration that each AU was exhibited throughout the visitor's interaction with FUTURE WORLDS. Each visitor's observed AU intensity values were standardized, and the duration of an AU was calculated during intervals where consecutive intensity values were at least one standard deviation greater than the mean of that particular visitor-specific AU feature. This is to ensure that only relative spikes of the intensity of each AU contributed towards the calculation of the total duration. Additionally, each duration was only recorded if it was present for longer than 0.5 consecutive seconds to avoid noise associated with facial micro expressions [42]. This process was performed for 18 AUs tracked by the software. Additional features were generated by calculating the percentage of a visitor's gameplay that contained the presence of the individual AU [10]. Using the same sequence, the standard deviation and maximum value of the AU values were calculated as well. In total, we distilled 72 facial expression-related features from video data processed by OpenFace.

Two interaction-based features were engineered based on the total number of times a visitor tapped on the multi-touch user interface of FUTURE WORLDS (Total Taps) and the total number of times the visitor tapped to display additional information about certain environmental sustainability elements of the game (Total Info Taps). Both of these interaction log features measure how actively participants interacted with the FUTURE WORLDS exhibit and its embedded environmental sustainability content.

A Tobii EyeX eye tracking device captured gaze data from each visitor and determined areas of interest (AOIs) within the context of the FUTURE WORLDS display. Instances of visitor gaze data that centered on in-game objects for a duration longer than 210 milliseconds were automatically identified and tracked, which is a fixation duration based on prior eye gaze research [39]. The gaze-based features that were distilled from this data include the proportion of time spent looking at four different types of in-game objects: virtual location (*AOI-Location*), environmental sustainability selection menus (*AOI-Menu*), environmental sustainability informational text and imagery (*AOI-Information*), and user interface elements (*AOI-Interface*). *AOI-Location* refers to fixations on any of nine distinct regions of the virtual environment. *AOI-Menu* refers to a menu that can be accessed by a visitor tapping on a particular location within the virtual environment, allowing the visitor to either learn more about that particular element or make changes within the virtual environment. *AOI-Information* represents fixations on textual labels pertaining to environmental sustainability concepts presented in FUTURE WORLDS, in addition to high-resolution images associated with the presented concepts. *AOI-Interface* represents fixations on UI-specific elements for navigating the FUTURE WORLDS software, such as the arrow buttons to change environmental conditions. These four categories of gaze-based features represent visitor fixation relative to different elements of the museum exhibit platform. Because these categories of durations can be summed to calculate dwell time directly, we scale each category to create a rate of fixation durations, which we describe in detail later.

5 Predictive Models of Visitor Dwell Time

We conduct two types of analysis: 1) early prediction of visitor dwell time and the improvement of these predictive models over time; and 2) ablation of specific modalities to determine how performance of early prediction models varies with different sets of modalities. Each of these is described below.

5.1 Early Prediction

A desirable characteristic of multimodal predictive models is that as a model observes more interaction data over time, its predictive accuracy increases. Rapidly converging toward accurate predictions provides adaptive learning environments information to proactively provide support or intervention. This could lead to more engaging interactive museum exhibits that can better capture the attention of visitors and lead to more effective learning experiences. To evaluate this characteristic, we trained machine learning-based regression models to predict visitor dwell time at various time points during the visitor's interaction. These models were trained on full sequences of visitor data and then tested on varying amounts of data stemming from different visitors.

To predict visitor dwell time throughout each visitor's interaction, we calculated the set of features described above using equal time intervals. Specifically, we split visitor interactions into 30-second cumulative segments and calculated the set of features from each modality encompassing data from the start of the interaction up until that 30-second interval. For example, if a visitor interacted with FUTURE WORLDS for two minutes and thirty seconds, he or she would have five time points where the multimodal features are calculated (at 30s, 1min, 1min 30s, 2min, 2min 30s). Each of these time points serves as a data point for our analysis. We used 30-second intervals to ensure sufficient changes in the features for each modality across time steps, as features from some modalities were sampled at different rates than others. Previous studies have successfully used smaller window sizes where data could be sampled at higher rates (e.g., video data), but this is not always possible [47].

After splitting all visitor data into time segments based on 30 second intervals, the resulting dataset consisted of 1,013 data points for the 85 visitors. Following this process, the features for each modality were scaled by dividing each feature by the total elapsed time at each point. This ensures that the models' performances are not artificially inflated due to monotonically increasing values within the features.

5.2 Ablation

To evaluate the performance of the early prediction models using different modalities, a set of ablation experiments was conducted. In practice, it may not always be possible to instrument a museum exhibit with multimodal sensors, as they could be obtrusive for exhibits on the museum floor or raise privacy concerns. For example, interaction logs can be collected from digital exhibits that have pre-existing logging features, or for which source code access is available, but this may not always be the case. Some data channels, such as visitor posture, can be collected unobtrusively from a physically distant camera, and can be configured to enhance privacy by performing run-time skeletal tracking without storing the raw video recordings. Similarly, facial expression poses privacy considerations, as personally identifying information from visitors is captured when facial video recordings are stored and analyzed. Other channels such as eye gaze require a cumbersome calibration process, and there are practical challenges with respect to adjusting the position and angle of the eye tracker based upon the height of the visitor. Data management, especially for data with personally identifiable information, is a critical consideration for multimodal instrumentation in museum contexts.

In the ablation experiments, we evaluate the impact of removing more intrusive sensors and data streams from the full set of multimodal data. First, we examine models induced with the full set of modalities: posture (P), facial expression (F), interaction logs (I), and eye gaze (E). Next, we removed the data captured by the eye tracker from the multimodal dataset, which resulted in three remaining modalities (PFI). This was done to simulate situations in which it may be infeasible to collect gaze data. The next ablation removed both the eye tracker data and the interaction logs, resulting in two modalities used (PF). Interaction logs require the ability to instrument exhibits with trace logging software, which sometimes is infeasible. The final ablation removed the facial expression data in addition to the eye gaze and interaction log data, leaving only posture data (P). This was done

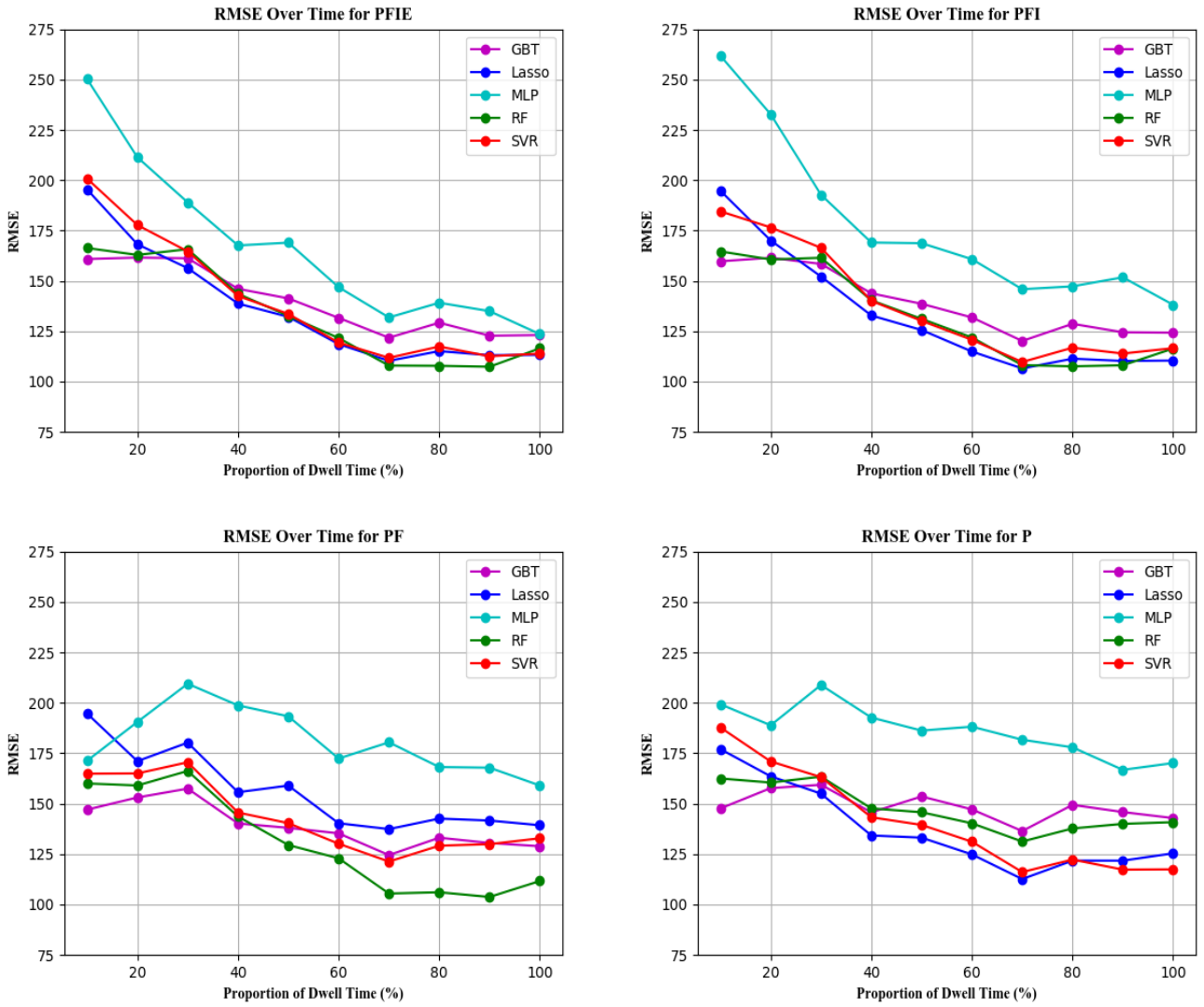

Figure 2: Early prediction performance using different sets of modalities (lower is better)

to simulate situations in which facial expression data is difficult (or impossible) to collect, either because of privacy concerns or risk of distracting visitors from the exhibit. In total, we compare four combinations of input modalities: the full set of modalities and three ablation conditions that remove one additional modality per condition.

6 Results

For evaluation, we trained five machine learning-based models: random forest (RF), support vector regression (SVR), Lasso regression (Lasso), gradient boosting trees regression (GBT), and multi-layer perceptron regression (MLP). We performed 10-fold cross-validation at the visitor level to ensure no data from a visitor was present in both the training and test partitions (i.e., 8-9 unique visitors per test set). For each machine learning model, we optimized the hyperparameters using grid search. The best set of parameters were determined in a nested 3-fold cross-validation within the training set in each fold of the outer cross-validation. For the RF models, we varied the number of estimators, the max depth allowed by the individual learned trees, and the minimum samples per leaf required to make a split on an internal node. For the SVR models, we varied the kernel and the regularization parameter. For Lasso regression, we varied the alpha regularization term. For the GBT models, we optimized the learning rate, the number of estimators, the max depth, and the minimum samples per leaf to make an internal node split. For the MLP modes, we used a two-layer model and varied the number of hidden nodes in both layers. All other hyperparameters were set to the default values offered by the machine learning library, scikit-learn. Once the best set of hyperparameters was found, the best performing parameters on the internal cross-validation were used to predict the dwell times of the data points in the outer

cross-validation test set, allowing for better generalization for each predictive modelling method.

Given the large number of available features when incorporating all sets of modalities, we performed feature selection to reduce the number of features that were utilized in the predictive models within each iteration of cross-validation. We performed univariate linear regression tests between the training set of features used in each experiment with dwell time and selected features that had p-values less than or equal to 0.15. Finally, all data was standardized within cross-validation splits using the training set's mean and standard deviation for each feature. We report the performance of early prediction models for each ablation condition in Figure 2.

In Figure 2, each graph shows the performance over proportions of visitor dwell times for each ablation condition for all predictive models. The x-axis denotes the percentage of the visitors' interaction that was used by the predictive models to predict dwell time. The y-axis shows the performance at each point in terms of root mean squared error (RMSE). We note that as the models see more data in each ablation, prediction performance tends to improve (i.e., the RMSE decreases). As modalities are removed, prediction performance tends to improve at a slower rate, and the predictions converge to a higher RMSE by the end of the visitors' interactions.

In addition to displaying early prediction performance, we also report the overall performance of the predictive models for each ablation. We report the R^2, RMSE, and mean absolute error (MAE) for each predictive model in each ablation condition using possible time segments of the visitors' interactions. This represents the average performance achieved by each model at all time points. These results are shown in Table 1. Each vertical section of the table displays the sets of modalities used, where "P" represents posture, "F" represents facial expression, "I" represents interaction logs, and "E" represents eye gaze. We bold the highest performing model for each set of modalities across the entire set of metrics. As displayed, the random forest (RF) and Lasso regression (Lasso) models tended to outperform competing methods across all time segments.

Table 1. Early prediction model performance on all time segments.

PFIE			
Model	R^2	*MAE*	*RMSE*
RF	**0.178**	**104.894**	**133.626**
SVR	0.105	109.906	139.442
Lasso	0.151	108.412	135.827
GBT	0.102	113.258	139.658
MLP	-0.271	127.448	166.140
PFI			
Model	R^2	*MAE*	*RMSE*
RF	**0.195**	**104.274**	**132.253**
SVR	0.127	110.053	137.721
Lasso	0.188	107.025	132.789
GBT	0.113	113.153	138.840
MLP	-0.433	139.786	176.453
PF			
Model	R^2	*MAE*	*RMSE*
RF	**0.207**	**102.005**	**131.258**
SVR	0.060	116.257	142.893
Lasso	-0.112	124.716	155.386
GBT	0.116	112.741	138.576
MLP	-0.515	150.354	181.423
P			
Model	R^2	*MAE*	*RMSE*
RF	0.010	120.477	146.646
SVR	0.092	110.978	140.417
Lasso	**0.142**	**109.816**	**136.548**
GBT	-0.017	120.857	148.609
MLP	-0.581	147.266	185.333

7 Discussion and Limitations

The evaluation revealed that predictive models of visitor dwell time can make improved predictions over time, and using additional modalities yields better performance as visitors near the end of their interactions. After evaluating five predictive models using four different combinations of modalities, the results revealed that random forest models outperformed competing models on three of the four modality combinations, and Lasso regression performed best on the unimodal configuration for predicting dwell time. These two models tend to perform well with limited data and relatively many features. The MLP model tended to perform worst on each set of modalities, likely suffering from overfitting and lack of data. Notably, overall predictive performance improves when removing eye gaze and interaction log modalities. However, removing facial expression results in a steep drop in performance.

One reason the eye gaze and interaction log modalities do not provide substantial boosts in performance is due to each of these features having few overall extracted features. We extracted four eye gaze features and two interaction log features, as opposed to 72 from facial expression and 18 from posture. Performance on earlier segments tends to improve faster when incorporating eye gaze and interaction logs, but the initial performance is worse.

The results in Table 1 indicate the overall trend for each predictive model by calculating the metrics using all time segments. This provides a snapshot of the performance, but it does not account for the final predictive performance achieved for each model. For example, the RMSE values for models in the PFIE condition are below 125.0, but the highest performing model overall (RF) yields an RMSE of 133.6. This tradeoff of evaluating models based on overall performance and performance when the models are given additional data should be considered further.

When incorporating more modalities, the predictive performance tended to converge to the highest achieved performance (RMSE: 100-125) for more models. Each predictive

model for each set of modalities tended to plateau around 70% of total dwell time, indicating that peak performance occurs before the end of visitor interactions. While this is the case for each set of modalities, the models that incorporate less modalities plateau much sooner. Additionally, the rate of improvement tended to be faster when the models were evaluated with more modalities.

There are a few potential causes of this pattern. First, when incorporating more modalities, the predictive models are better able to learn complex relationships between modalities, which can change over time. Uncovering these patterns can produce faster improvement when leveraging more data. Second, the models trained with more sets of modalities have access to more features, which introduces more opportunity to characterize visitor engagement. However, introducing many features also introduces the opportunity for additional noise when predicting dwell time. While feature selection attempts to minimize the effect of noisy features, there is still opportunity to have negative effects from combining modalities [11]. Additionally, multicollinearity and redundancy between features coming from different modalities can inhibit predictive performance when combining modalities.

The results from this analysis highlight the value of modeling museum visitor engagement with multiple modalities. Incorporating sensors that are relatively unobtrusive for exhibits can support accurate predictions of visitor engagement. Additional data channels such as eye gaze and interaction trace logs can provide meaningful data that can help early prediction models of visitor engagement generate accurate predictions efficiently. We anticipate that the methodological approach and feature representations presented in this work are generalizable to other exhibits and museums. Specifically, we hypothesize that the predictive performance of the models would be similar in other museums contexts, although the results may be influenced by factors such as visitor characteristics and exhibit design.

There are several limitations of the work that should be noted. A common challenge in multimodal learning analytics is determining which predictive features to extract from multimodal data streams. We distilled features based on the results of previous work investigating multimodal interaction in adaptive learning environments, but further investigation is needed to extract richer, more descriptive features in modalities where the number of features is scarce (i.e., eye gaze and interaction logs). A second limitation is the way that we represented visitor interactions. We chose to split visitor interactions into segments of thirty second cumulative intervals, making predictions after each thirty-second interval. Choosing intervals that are shorter (e.g., fifteen seconds) could allow for more predictions to be made at earlier points within a visitor's interaction, and it could increase the training sample size as well as afford the opportunity to provide more fine-grained adaptive support for visitors. We operationalize visitor engagement in terms of dwell time, but we do not utilize other channels of engagement (e.g., self-reports, field observations). Dwell time is a useful measure of behavioral engagement, but it does not provide a comprehensive picture of visitor experience. Another limitation of this work is related to quantifying early prediction convergence in regression-based models. Various metrics exist for early prediction in classification settings [30], but there are no analogous methods for regression models. As such, we measured the models' performance over time to visually display both convergence and improvement of prediction accuracy as the models were given additional data.

8 Conclusion

Visitor engagement is critically important for learning in informal learning environments such as museums. Multimodal learning analytics offers significant potential for modeling learner engagement in museums, which is challenging due to several factors, including very brief visitor dwell times and the multifaceted nature of how visitors manifest engagement (e.g., visual attention and body language). To address these challenges, we have introduced a machine learning-based approach to predicting visitor engagement with museum exhibits at early points in their interactions using multimodal sensor data. We found that by leveraging multimodal data collected from visitor interactions with an interactive game-based exhibit, early prediction models were able to predict visitor dwell times efficiently and accurately. Furthermore, the early prediction models that incorporated more multimodal channels reached accurate predictions at a faster rate. Results indicate that each multimodal channel has predictive value for modeling dwell time, highlighting the importance of modeling visitor dwell time with multimodal data. These results are in agreement with previous work investigating multimodal learning analytics for other learning settings, such as classrooms and laboratories. These findings also highlight the need to investigate predictive models that use unobtrusive sensors to model the museum visitor experience, as not all modalities provide equal predictive value.

There are several promising directions for future work. First, investigating sequential representations of visitor data could help reveal temporal patterns of visitor engagement. Sequential models (e.g., recurrent neural networks) that utilize fine-grained sequences of visitor data are promising techniques to investigate. Second, utilizing additional modalities will be an important next step as the results showed that more modalities provide predictive value. To this end, additional features from each modality should be distilled to better model how each modality bears on engagement. A third direction will be to incorporate this early prediction approach into museum exhibits to enable run-time interventions to support visitor engagement. This will introduce the opportunity to enrich understanding of how multimodal learning analytics can dynamically capture visitor engagement in museums and to enhance the learning experiences of visitors.

ACKNOWLEDGMENTS

The authors would like to thank the staff and visitors of the North Carolina Museum of Natural Sciences. This research was supported by the National Science Foundation under Grant DRL-1713545. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Pengcheng An, Kenneth Holstein, Bernice d'Anjou, Berry Eggen, and Saskia Bakker. 2020. The TA Framework: Designing real-time teaching augmentation for K-12 classrooms. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* 1-17.

[2] Sinem Aslan, Nese Alyuz, Cagri Tanriover, Sinem E. Mete, Eda Okur, Sidney K. D'Mello, and Asli Arslan Esme. 2019. Investigating the impact of a real-time, multimodal student engagement analytics technology in authentic classrooms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1-12.

[3] Silvia Bach, Ulla Richardson, Daniel Brandeis, Ernst Martin, and Silvia Brem. 2013. Print-specific multimodal brain activation in kindergarten improves prediction of reading skills in second grade. *Neuroimage* 82, (2013), 605-615.

[4] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1-10.

[5] Paulo Blikstein, and Marcelo Worsley. 2016. Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *J. of Learn. Analytics* 3, 2 (2016). 220-238.

[6] Florian Block, James Hammerman, Michael Horn, Amy Spiegel, Jonathan Christiansen, Brenda Phillips, Judy Diamond, E. Margaret Evans, and Chia Shen. 2015. Fluid grouping: Quantifying group engagement around interactive tabletop exhibits in the wild. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 867-876.

[7] Nigel Bosch, Huili Chen, Sidney D'Mello, Ryan Baker, and Valeria Shute (2015, November). Accuracy vs. availability heuristic in multimodal affect detection in the wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 267-274.

[8] Nigel Bosch, Sidney K. D'Mello, Ryan S. Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. 2016. Detecting student emotions in computer-enabled classrooms. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*. 4125-4129.

[9] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. 2017. The NoXi database: Multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 350-359.

[10] Cheng Chang, Cheng Zhang, Lei Chen, and Yang Liu. 2018. An ensemble model using face and body tracking for engagement detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 616-622.

[11] Sidney D'Mello and Arthur Graesser. 2010. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction* 20, 2 (2010), 147-187.

[12] Sidney D'Mello, Andrew Olney, Nathan Blanchard, Borhan Samei, Xiaoyi Sun, Brooke Ward, and Sean Kelly. 2015. Multimodal capture of teacher-student interactions for automated dialogic analysis in live classrooms. In *Proceedings of the 17th International Conference on Multimodal Interaction*. 557-566.

[13] Jeanine DeFalco, Jonathan Rowe, Luc Paquette, Vasiliki Georgoulas-Sherry, Keith Brawner, Bradford Mott, Ryan Baker, and James Lester. 2018. Detecting and addressing frustration in a serious game for military training. *International Journal of Artificial Intelligence in Education* 28, 2 (2018), 152-193.

[14] Abhinav Dhall, Amanjot Kaur, Roland Goecke, and Tom Gedeon. 2018. Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 653-656.

[15] Judy Diamond, Michael Horn, and David H. Uttal. 2016. Practical evaluation guide: Tools for museums and other informal educational settings. Rowman & Littlefield.

[16] Eyal Dim and Tsvi Kuflik. 2014. Automatic detection of social behavior of museum visitor pairs. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 4, 4 (2014). 1-30.

[17] Vanessa Echeverría, Allan Avendaño, Katherine Chiluiza, Aníbal Vásquez, and Xavier Ochoa. 2014. Presentation skills estimation based on video and Kinect data analysis. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*. 53-60.

[18] Lucca Eloy, Angela Stewart, Mary Jean Amon, Caroline Reinhardt, Amanda Michaels, Chen Sun, Valerie Shute, Nicholas Duran, and Sidney D'Mello. 2019. Modeling team-level multimodal dynamics during multiparty collaboration. In *Proceedings of the 21st International Conference on Multimodal Interaction*. 244-258.

[19] Andrew Emerson, Nathan Henderson, Jonathan Rowe, Wookhee Min, Seung Lee, James Minogue, and James Lester. 2020. Investigating visitor engagement in interactive science museum exhibits with multimodal Bayesian hierarchical models. In *Proceedings of the 21st International Conference on Artificial Intelligence in Education*. 165-176.

[20] John H. Falk and Lynn D. Dierking. 2018. *Learning from museums*. Rowman & Littlefield.

[21] Joseph Grafsgaard, Kristy Boyer, Eric Wiebe, and James Lester. 2012. Analyzing posture and affect in task-oriented tutoring. In *Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference*. 438-443.

[22] Lisa Halverson and Charles Graham. 2019. Learner engagement in blended learning environments: A conceptual framework. *Online Learning* 23, 2 (2019). 145-178.

[23] Nathan Henderson, Jonathan Rowe, Luc Paquette, Ryan Baker, and James Lester. 2020. Improving affect detection in game-based learning with multimodal data fusion. In *Proceedings of the 21st International Conference on Artificial Intelligence in Education*. 228-239.

[24] George Hein. 2009. Learning science in informal environments: People, places, and pursuits. *Museums and Social Issues* 4, 1 (2009). 113-124.

[25] Karina Huang, Tonya Bryant, and Bertrand Schneider. 2019. Identifying collaborative learning states using unsupervised machine learning on eye-tracking, physiological and motion sensor data. In *Proceedings of the 12th International Conference on Educational Data Mining*. 323-329.

[26] Karen Knutson, Mandela Lyon, Kevin Crowley, and Lauren Giarratani. 2016. Flexible interventions to increase family engagement at natural history museum dioramas. *Curator: The Museum Journal* 59, 4 (2016). 339-352.

[27] Tsvi Kuflik and Eyal Dim. 2013. Early detection of pairs of visitors by using a museum triage. In *Proceedings of the Annual Conference of Museums and the Web*. 113-124.

[28] Chad Lane, Dan Noren, Daniel Auerbach, Mike Birch, and William Swartout. 2011. Intelligent tutoring goes to the museum in the big city: A pedagogical agent for informal science education. In *Proceedings of the 15th International Conference on Artificial Intelligence in Education*. 155-162.

[29] Duri Long, Tom McKlin, Anna Weisling, William Martin, Hannah Guthrie, and Brian Magerko. 2019. Trajectories of physical engagement and expression in a co-creative museum installation. In *Proceedings of the 2019 ACM SIGCHI Conference on Creativity and Cognition*. 246-257.

[30] Wookhee Min, Alok Baikadi, Bradford Mott, Jonathan Rowe, Barry Liu, Eun Young Ha, and James Lester. 2016. A generalized multidimensional evaluation framework for player goal recognition. In *Proceedings of the 12th Annual AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. 197-203.

[31] Wookhee Min, Bradford Mott, Jonathan Rowe, Robert Taylor, Eric Wiebe, Kristy Boyer, and James Lester. 2017. Multimodal goal recognition in open-world digital games. In *Proceedings of the 13th Artificial Intelligence and Interactive Digital Entertainment Conference*. 80-86.

[32] Hamed Monkaresi, Nigel Bosch, Rafael Calvo, and Sidney D'Mello. 2016. Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Trans. on Affect. Comp.* 8, 1 (2016). 15-28.

[33] Xavier Ochoa. 2017. Multimodal learning analytics. *The Handbook of Learning Analytics* 1 (2017). 129-141.

[34] Sharon Oviatt. 2018. Ten opportunities and challenges for advancing student-centered multimodal learning analytics. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 87-94.

[35] Sharon Oviatt, Joseph Grafsgaard, Lei Chen, and Xavier Ochoa. 2018. Multimodal learning analytics: Assessing learners' mental state during the process of learning. *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition* 2, (2018). 331-374.

[36] Luis Prieto, Kshitij Sharma, Pierre Dillenbourg, and María Jesús. 2016. Teaching analytics: Towards automatic extraction of orchestration graphs using wearable sensors. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. 148-157.

[37] Mirko Raca, and Pierre Dillenbourg. 2013. System for assessing classroom attention. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*. 265-269.

[38] Mirko Raca, Roland Tormey, and Pierre Dillenbourg. 2014. Sleepers' lag-study on motion and attention. In *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge*. 36-43.

[39] Keith Rayner, Xingshan Li, Carrick C. Williams, Kyle R. Cave, and Arnold D. Well. 2007. Eye movements during information processing tasks: Individual differences and cultural effects. *Vision Research* 47, 21 (2007). 2714-2726.

[40] Joseph Reilly, Milan Ravenell, and Bertrand Schneider. 2018. Exploring collaboration using motion sensors and multi-modal learning analytics. In *Proceedings of the 11th International Conference on Educational Data Mining*. 333-339.

[41] Jonathan P. Rowe, Eleni V. Lobene, Bradford W. Mott, and James C. Lester. 2017. Play in the museum: Design and development of a game-based learning exhibit for informal science education. *International Journal of Gaming and Computer-Mediated Simulations* 9, 3 (2017). 96-113.

[42] Robert Sawyer, Andy Smith, Jonathan Rowe, Roger Azevedo, and James Lester. 2017. Enhancing student models in game-based learning with facial expression recognition. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*. 192-201.

[43] Kshitij Sharma, Zacharoula Papamitsiou, and Michail Giannakos. 2019. Building pipelines for educational data using AI and multimodal analytics: A "grey-box" approach. *British Journal of Educational Technology* 50, 6 (2019). 3004-3031.

[44] Kshitij Sharma, Zacharoula Papamitsiou, Jennifer K. Olsen, and Michail Giannakos. 2020. Predicting learners' effortful behaviour in adaptive assessment using multimodal data. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. 480-489.

[45] Michelle Taub, Robert Sawyer, Andy Smith, Jonathan Rowe, Roger Azevedo, and James Lester. 2020. The agency effect: The impact of student agency on learning, emotions, and problem-solving behaviors in a game-based learning environment. *Computers & Education* 147, (2020).

[46] Chinchu Thomas, Nitin Nair, and Dinesh Babu Jayagopi. 2018. Predicting engagement intensity in the wild using temporal convolutional network. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 604-610.

[47] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R. Movellan. 2014. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing* 5, 1 (2014). 86-98.

[48] Joseph Wiggins, Mayank Kulkarni, Wookhee Min, Bradford Mott, Kristy Boyer, Eric Wiebe, and James Lester. 2018. Affect-based early prediction of player mental demand and engagement for educational games. In *Proceedings of the 14th Artificial Intelligence and Interactive Digital Entertainment Conference*. 243-249.

[49] Marcelo Worsley, Stefan Scherer, Louis-Philippe Morency, and Paulo Blikstein. 2015. Exploring behavior representation for learning analytics. In *Proceedings of the 2015 ACM International Conference on Multimodal Interaction*. 251-258.

[50] Liang-Chih Yu, C. W. Lee, H. I. Pan, Chih-Yueh Chou, Po-Yao Chao, Z. H. Chen, S. F. Tseng, C. L. Chan, and K. Robert Lai. 2018. Improving early prediction of academic failure using sentiment analysis on self-evaluated comments. *Journal of Computer Assisted Learning* 34, 4 (2018). 358-365.