Multimodal Predictive Student Modeling with Multi-Task Transfer Learning

Andrew Emerson
Educational Testing Service
aemerson@ets.org

Wookhee Min
North Carolina State University
wmin@ncsu.edu

Jonathan Rowe
North Carolina State University
jprowe@ncsu.edu

Roger Azevedo
University of Central Florida
roger.azevedo@ucf.edu

James Lester
North Carolina State University
lester@ncsu.edu

ABSTRACT

Game-based learning environments have the distinctive capacity to promote learning experiences that are both engaging and effective. Recent advances in sensor-based technologies (e.g., facial expression analysis and eye gaze tracking) and natural language processing have introduced the opportunity to leverage multimodal data streams for learning analytics. Learning analytics and student modeling informed by multimodal data captured during students' interactions with game-based learning environments hold significant promise for designing effective learning environments that detect unproductive student behaviors and provide adaptive support for students during learning. Learning analytics frameworks that can accurately predict student learning outcomes early in students' interactions hold considerable promise for enabling environments to dynamically adapt to individual student needs. In this paper, we investigate a multimodal, multi-task predictive student modeling framework for game-based learning environments. The framework is evaluated on two datasets of game-based learning interactions from two student populations (*n*=61 and *n*=118) who interacted with two versions of a game-based learning environment for microbiology education. The framework leverages available multimodal data channels from the datasets to simultaneously predict student post-test performance and interest. In addition to inducing models for each dataset individually, this work investigates the ability to use information learned from one source dataset to improve models based on another target dataset (i.e., transfer learning using pre-trained models). Results from a series of ablation experiments indicate the differences in predictive capacity among a combination of modalities including gameplay, eye gaze, facial expressions, and reflection text for predicting the two target variables. In addition, multi-task models were able to improve predictive performance compared to single-task baselines for one target variable, but not both. Lastly, transfer learning showed promise in improving predictive capacity in both datasets.

CCS CONCEPTS

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

KEYWORDS

Multimodal Learning Analytics, Game-Based Learning, Predictive Student Modeling

ACM Reference Format:
Andrew Emerson, Wookhee Min, Jonathan Rowe, Roger Azevedo, and James Lester. 2023. Multimodal Predictive Student Modeling with Multi-Task Transfer Learning. In *LAK23: 13th International Learning Analytics and Knowledge Conference (LAK 2023), March 13–17, 2023, Arlington, TX, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3576050.3576101

1 INTRODUCTION

Student interest plays a critical role in learning [36]. Intelligent game-based learning environments are a form of learning environment that are designed to promote student interest while enhancing performance or knowledge on a particular subject through adaptive support in real-time [22, 50]. Automatically detecting specific student behaviors and modeling student knowledge and skills is a promising approach to drive adaptations in game-based learning environments [23, 47]. However, this approach poses significant challenges due to the complexity of student learning and interest. *Multimodal learning analytics* hold considerable promise for modeling complex student learning behaviors in game-based learning by leveraging advances in sensor-based technologies (e.g., facial expression toolkits, eye trackers) and natural language processing [8]. Predictive student models of complex learning phenomena that leverage multimodal learning analytics must also be able to make accurate predictions at early stages (i.e., as students are learning) in student gameplay for the system to appropriately make adaptations to support the student. Few studies have investigated the degree to which separate modalities impact the performance of predictive student models in game-based learning, and the capacity to use multimodal data to predict students' post-test performance and interest at early points during gameplay remains unexplored.

Inducing multimodal predictive student models of post-test performance and interest in game-based learning poses significant challenges and opportunities. First, the success of multimodal data for student modeling is not universal, as prior research has indicated that for certain learning-related constructs, additional modalities can provide additive or even super-additive effects on the accuracy of student models, but in other cases, more modalities can have

LAK 2023, March 13–17, 2023, Arlington, TX, USA

ACM ISBN 978-1-4503-9865-7/23/03…$15.00
https://doi.org/10.1145/3576050.3576101

inhibitory effects on model accuracy [6]. As a result, investigating the impact of separate modalities in the prediction of each learning outcome or motivational construct is critical. Second, these motivational constructs and learning outcomes also have complex relationships [16], which suggests that predictive student models may be able to leverage shared information learned by predicting each outcome simultaneously and improve predictive accuracy across multiple relevant modeling tasks. A potential solution to this problem is *multi-task learning*, wherein each learning outcome or motivational construct is a target variable in the same multimodal predictive student model. Third, different game-based learning environments and student populations may elicit different relationships between multimodal data, post-test performance, and interest, necessitating a separate model for each student population and game-based learning environment [36]. A key research question then is, how can multimodal predictive models be adapted across different student populations, game-based learning environments, and domains to improve the model's predictive performance? A promising approach for leveraging information across domains is the use of transfer learning (specifically, *domain adaptation*) [32]. Transfer learning in this context would include leveraging learned representations of overlapping, but not identical, multimodal data across populations in game-based learning environments for predictive student modeling tasks. However, research on leveraging transfer learning to create personalized predictive student models in game-based learning is still in early stages [44] and requires additional investigation.

In this paper, we introduce a unified multimodal predictive student modeling framework using two separate datasets collected using the Crystal Island game-based learning environment for microbiology education: Crystal Island – Sensor-Based and Crystal Island – Reflection. In Crystal Island – Sensor-Based, undergraduate students interacted with the game, and gameplay interaction logs, student eye gaze, and student facial expressions were collected. In Crystal Island – Reflection, a group of middle school students interacted with a different version of the game that did not include sensor data, where the system prompted the students for text-based reflections of their learning at various plot points during their interactions. For both datasets, we extracted the temporal features associated with the gameplay and additional available modalities and used them as input to multi-task early prediction models of post-test score and interest, which are the target variables in this work. We then investigated the use of transfer learning as a means to improve the performance of multimodal predictive models for each dataset by employing an unsupervised learning technique (i.e., an autoencoder) to learn representations of the shared modality (i.e., student gameplay) between domains. This is the first work to combine the three components of (1) multimodal learning analytics, (2) early prediction, and (3) multi-task learning into a single, unified predictive student modeling framework for game-based learning. Additionally, this paper contributes findings on the efficacy of transfer learning between multiple student populations and game-based learning environments, where there is a shared modality between the datasets. The resulting framework offers the capability of identifying students who are disinterested or struggling with learning at early points throughout their learning experience, which are two critical issues that can affect the learning and problem-solving experience. These identified instances can then guide a separate adaptive scaffolding system to give personalized support to the student and to keep track of the student's progress.

2 RELATED WORK

Multimodal learning analytics has been the subject of growing interest in recent years for its focus on creating a data-rich understanding of student learning and associated behaviors and constructs [3, 4, 30, 31]. Multimodal data have been used to model student engagement [18, 52] and knowledge [8, 24, 43], and have been utilized to support teachers in classrooms [1, 2, 38]. Modalities such as eye gaze [39], facial expressions [49], and student-generated text-based responses [21] have been incorporated into student models. In game-based learning environments, multimodal data show considerable promise for modeling student post-test performance and interest [8]. Previous studies have leveraged multimodal data in game-based learning environments [13], but few studies have observed the degree to which multimodal data can make accurate predictions of student learning-related constructs at early points in student gameplay (e.g., [9]).

Predictive student modeling holds significant potential for improving learning experiences and supporting struggling students. By observing students' learning behaviors over time, models could predict student knowledge and interest in the subject area, which could then be used in real-time to inform adaptive support. Using student gameplay interactions and other sources of data that provide evidence of student learning, this modeling approach aims to predict students' future competencies and characteristics using the data up to a specific moment [45]. This algorithmic technique is known as *early prediction*, which has seen significant success in student modeling applications [17, 28]. Prior research has investigated predictive student models of student knowledge in game-based learning [10], but limited work has investigated predictive student models of interest.

Due to the complex relationships between constructs such as interest and learning-related outcomes, modeling these variables by leveraging the related information could prove beneficial. Little work has explored the use of *multi-task learning* to account for multiple, related learning outcomes and constructs, especially in the context of game-based learning. Multi-task learning has been studied in several domains, such as computer vision [19] and natural language processing [46], but less so in student modeling. Prior work used student behaviors in an adaptive learning environment to simultaneously model hint-taking and knowledge [5]. Other work incorporated multimodal data into multi-task student affect models in game-based learning, finding that jointly predicting affective states yields improved results over single-task baselines [14]. To improve performance further, *transfer learning* between student populations or learning environments can provide the student models access to rich information from other datasets, but this family of algorithms has not been extensively studied in game-based learning. A common problem in student modeling is the *cold start* problem, where modeling new students or learning environments can be difficult without prior information [37]. Transfer learning has been

Figure 1: Crystal Island **game-based learning environment.**

adopted as one approach to mitigating this issue, including work in programming environments [25] and game-based learning environments [12, 44]. Building on this and other work in the context of game-based learning, we present a unified framework for multimodal learning analytics, predictive student modeling, multi-task learning, and transfer learning.

3 METHOD

To induce predictive student models of interest and post-test performance, we utilized two datasets stemming from the same game-based learning environment, Crystal Island. In the first, Crystal Island – Sensor-Based, we equipped the game-based learning environment with a suite of multimodal sensors to collect student gameplay, facial expression, and eye gaze data. In the second, Crystal Island – Reflection, we equipped the game-based learning environment with a reflection tool that allowed students to reflect on their progress in the game. The written (i.e., textual) reflections were collected in conjunction with student gameplay as the students interacted with Crystal Island. The specific datasets, study participants, and procedures, are described below.

3.1 Crystal Island Game-Based Learning Environment

Crystal Island is a game-based learning environment for microbiology education (Figure 1) [48]. This paper includes data collected from two versions of the game with two different student populations. In the game, students play the role of a medical detective whose goal is to uncover a mysterious disease outbreak on a remote island. Students collect information by reading virtual books, talking to non-player characters, and testing hypotheses using the in-game laboratory equipment. All student actions, movement, dialogue, and interactions with in-game objects are recorded and logged for further analysis. Once the student has gathered evidence, they are able to submit a diagnosis and treatment plan to the island's camp nurse.

3.2 Multimodal Data Collections

3.2.1 Crystal Island – Sensor-Based. In a study, sixty-five college student participants from a large North American university interacted with the Crystal Island game-based learning environment. Four students were removed from the initial dataset due to missing survey or sensor data. This resulted in a final dataset of 61 students (*M*=20.1 years old, *SD*=1.56) of which 42 (69%) were female. Each student played the game until correctly solving the science mystery or running out of allotted time (maximum of 3 hours). Gameplay durations ranged from 26.5 to 159.9 minutes (*M*=68.2, *SD*=22.7).

Prior to learning with Crystal Island, participating students completed a series of questionnaires and a 21-item, 4-option multiple-choice microbiology pre-test assessment to measure prior knowledge (M=11.84, SD=2.74). The researchers then calibrated an eye tracker and facial expression analysis software for the students. After calibration, students were instructed to begin playing Crystal Island, which started with a tutorial that introduced students to the overall objective of the learning session, which was to solve the mystery illness affecting the inhabitants on the island. Students then attempted to solve the science mystery. After interacting with Crystal Island, the Intrinsic Motivation Inventory (IMI) [41] was administered. For the purposes of this paper, the 7-point Interest-Enjoyment subscale (α=0.96; *M*=4.67, *SD*=1.37) was the primary subscale utilized to operationalize interest. Afterward, students completed a 21-item, 4-option multiple-choice post-test assessment similar to the pre-test assessment to measure acquired knowledge about microbiology (*M*=14.13, *SD*=2.85).

3.2.2 Crystal Island – Reflection. The second game-based learning environment used in this work is a different version of the Crystal Island environment. Students played a version of Crystal Island that prompted them periodically to reflect on what they had learned and their upcoming plans in the game (i.e., "In your own words, please describe the most important things that you've learned so far, and what is your plan moving forward?"). The embedded reflection prompts were designed to elicit reflective thinking, thus encouraging students to monitor and adapt their learning processes based on their game-based learning behaviors and problem-solving plans throughout the game thus far [26, 51]. The prompts were administered after key plot points and milestones in the game's science problem scenario.

Eighth-grade students from a middle school in the mid-Atlantic region were enrolled as part of two separate classroom studies. We combine the data collected across the 2018 (*n*=61) and 2019 (*n*=95) studies into a single dataset and analyze it in aggregate. After removing students with missing data (e.g., absences, failure to complete post-test survey), the final dataset was composed of data from 118 students. The average age of students was 13.6 (*SD*=0.5), with 55 male, 60 female, and 3 students responding as Other. There were 36% of students who identified as White, 27% as Black or African American, 18% as Hispanic or Latino, 2% as Asian, 1% as American Indian or Alaskan Native, and 15% as Other.

The pre-study survey included a 17-item multiple-choice content knowledge assessment on the student's microbiology content knowledge (*M*=6.78, *SD*=2.75). After students were introduced to the game and the purpose of the study, they began playing the tutorial phase of the game. Once finished with the tutorial, students had full agency to explore the island and investigate the mysterious illness. At several plot points in the game, students were asked to reflect on their learning by providing a free-response description of their game progress and upcoming problem-solving plans. Students were prompted up to five times at selected trigger points in the game chosen to minimize disruption to gameplay. After the students played Crystal Island, researchers administered a post-test that contained a separate set of 17-items addressing the same microbiology content knowledge (*M*=7.36, *SD*=3.36). All the information needed to answer the questions on both assessments could be found in Crystal Island, whether through books or articles, dialogue with in-game characters, or by viewing informational posters. After gameplay, the same Interest-Enjoyment subscale of the IMI survey was administered (α=0.90; *M*=5.10, *SD*=1.6). The average gameplay duration across all students was 76.3 minutes (*SD*=19.5). After all students had either solved the mystery (72%) or run out of time (28%), they were directed to the post-study survey.

3.3 Data Coding and Processing

3.3.1 Student Knowledge. To capture prior knowledge of microbiology, we first encoded the students' total correct answers on the pre-test for both Crystal Island – Sensor-Based and Crystal Island – Reflection using a binary representation for each question. Similarly, we counted the total number of correct responses on the post-test for both Crystal Island – Sensor-Based and Crystal Island – Reflection to operationalize student knowledge, and we included these data as target variables in our predictive models. We utilized the overlapping pre- and post-test questions from each dataset to use a fixed set of labels and input, resulting in 17 pre- and post-test questions each, which were positively correlated (*r*=0.51, *p*<0.05). For the post-test target variable for each dataset, we converted the predictive task into a classification task by splitting the post-test scores into two groups defined by a median split, where each group contained one-half of the sample scores. Next, we assigned participants to either a low (0) or high (1) post-test performance group. For Crystal Island – Sensor-Based, the low group was defined as a score of below 12.0 (27 students), and the high group was defined as a score above or equal to 12.0 (34 students). For Crystal Island – Reflection, the low group was defined as a score of below 7.0 (52 students), and the high group was defined as a score above or equal to 7.0 (66 students). We chose to split the data using this method over the continuous post-test values and more granular splits such as tertile splits because of the limited sample size. The sizeable score difference between groups occurred due to the different population ages (undergraduate vs. middle grade).

3.3.2 Student Interest. Using the Interest and Enjoyment subscale of the IMI as described in Sections 3.2.1 and 3.2.2, we encoded student interest for both Crystal Island – Sensor-Based and Crystal Island – Reflection. Similar to the median split process described for student knowledge, we split students into low (0) and high (1) interest groups based on these scores. For Crystal Island – Sensor-Based, the low group was defined as a score of below 4.86 (34 students), and the high group was defined as a score above or equal to 4.86 (27 students). For Crystal Island – Reflection, the low group was defined as a score of below 5.29 (56 students), and the high group was defined as a score above or equal to 5.29 (62 students).

3.3.3 Facial Expression Recognition and Feature Representation. We captured student facial expressions using the FACET facial expression toolkit [15], which extracts facial features for each video frame that correspond to the Facial Action Coding System (FACS) [7]. Each facial action unit (AU) was processed to derive a total duration each student spent exhibiting the given AU. In addition to the duration of each AU activation, we also computed the total number of times each AU was exhibited by each student. To compute the total number of AU occurrences, we counted the number of times each AU surpassed an intensity threshold of 0.5 for longer than a duration threshold of 0.5 seconds. This approach reduces the effect of noise in the sensor measurements [8]. There were 40 total features extracted from students' facial expression: a total duration and the total event count for each of the 20 available AUs. As this is an early prediction setting, each feature was computed cumulatively over time to enable model predictions at incremental moments in student gameplay. We describe this process in Section 5.

3.3.4 Gaze-Based Entity Tracking and Eye Tracking Feature Representation. Student eye gaze in Crystal Island – Sensor-Based was captured with the SMI RED 250 eye tracker using a 9-point calibration. During interactions with the game, the software responsible for logging student eye gaze data records fixations on each possible type of in-game object. In the game, there are 145 unique in-game objects. We used the fixations on each individual game object. In processing these fixations, we calculated the total duration that each student spent fixating upon each of the objects. We also computed the total number of fixation events (250 milliseconds or longer), a threshold used based on prior research on eye fixations during reading [40], for each student on each game object. Using both the counts and the fixation durations resulted in 290 total features for students' eye gaze. Each feature was computed cumulatively over time, enabling early prediction at incremental moments in gameplay.

3.3.5 Reflective Writing Processing. In Crystal Island – Reflection, students' written responses to the reflection prompts were encoded using word embedding techniques. Specifically, we compared two language models to embed each written reflection response: 300-dimensional GloVe embeddings [34] and 1024-dimensional ELMo embeddings [35]. A single embedding representation for each student's overall reflections at the time of prediction was computed to enable early prediction by averaging their mean reflection embeddings across their provided reflections up until the current time point. Each reflection response embedding was calculated by averaging word embeddings included in the response using a language model.

3.3.6 Gameplay Features. Both the Crystal Island – Sensor-Based and Crystal Island – Reflection datasets include gameplay data captured from students as they interact with the game environment. We encoded students' gameplay actions using several components: milestone completion, action type, action arguments, and action location. Specifically, each action is represented by a concatenation of one-hot encoded vectors for each action component at each time point. We then sum the sequence of these concatenated one-hot vectors up to the point of prediction. This yielded a count-based feature vector with 130 features, where the counts identify how many times each type of action and the corresponding components occurred in the sequence. Each feature vector also included the current time (seconds) elapsed from the start of the game.

4 PREDICTIVE STUDENT MODELS OF INTEREST AND KNOWLEDGE

To investigate the unified student modeling framework, we built predictive models of student knowledge and interest that were informed by the data described above. Using the available multimodal data from each dataset, we classified students into low and high groups for their post-test and interest scores. In the Crystal Island – Sensor-Based dataset, Pearson correlations indicated non-significant linear relationships between the post-test and interest scores (r=0.16, p=0.23). However, Pearson correlations did indicate significant linear relationships between post-test and interest scores in the Crystal Island – Reflection dataset (r=0.23, p<0.05). We compared different sets of multimodal predictive classifiers trained on the data from students who learned with Crystal Island – Sensor-Based (n=61) and Crystal Island – Reflection (n=118) separately. For Crystal Island – Sensor-Based, the total number of features includes 40 related to facial expression, 290 related to eye gaze, and 130 related to gameplay, for a total of 460 possible multimodal features per student. For Crystal Island – Reflection, the total number of features includes either 300 (GloVe) or 1,024 (ELMo) related to the reflection text and 130 related to gameplay, resulting in a total of 430 or 1,154 possible multimodal features per student. In addition to the multimodal features, each model also used data from the 17 questions from the pre-test. The training of each predictive model for each dataset was conducted using 5-fold cross-validation at the student-level, allowing all 61 or 118 students to be either used for training or testing. This means that 80% and 20% of the students were in the training and testing sets for each fold, respectively, and there was no overlap of students in each training and testing set to avoid data leakage.

To account for the high number of possible features in the models, we performed feature reduction with principal component analysis (PCA) in several ways to reduce the chance of overfitting. Within each cross-validation fold, we performed one of three possible PCA reductions: no PCA performed (i.e., no feature reduction), PCA performed on each input modality separately, and PCA on the input modalities jointly. When PCA was performed, the dimensions were reduced to either 32 or 64 considering our dataset size and findings from previous research that demonstrated high predictive performance with PCA applied to ELMo embeddings [29]. Specifically, to determine the optimal method of applying PCA across the multimodal dataset, we use two variations of feature-level data fusion. The first approach is performed by concatenating all features from each modality, and then performing PCA on this set to generate a final set of either 32 or 64 total features. In the second approach, PCA was performed on each modality's feature set separately, generating sets of 32 or 64 features per modality, which were then concatenated prior to training the student model. These varying data fusion techniques allow for the complex inter- and intra-modal relationships to be leveraged during model training and prediction, and they can be generalized further. When reporting the results of the multimodal predictive models for a specific combination of modalities, we will report the results from the data fusion technique (i.e., the PCA reduction details) that performed the best rather than reporting the performance of each possible data fusion technique due to the space constraints.

Student performance on the post-test and their interest score were predicted at two-minute intervals of gameplay as well as at the conclusion of the game. For each interval, we used a cumulative representation of the student data up to that point. In this analysis, we used the random forest (RF) classification algorithm, which supports both single-task learning and multi-task learning [33], for each predictive model. That is, we train an individual RF classifier to predict either post-test performance or interest, and we also use an individual RF classifier to predict both outcome variables. This allowed for easier comparison between the modality combinations and processing conditions using a single set of classification algorithms. The input features to each classifier were the pre-test items, the cumulative representations of the combination of modalities (i.e., a combination of facial expressions, eye gaze, gameplay, and reflection text depending on the experimental setting and modality availability), and elapsed game time. Each of the cumulative multimodal features were scaled first by elapsed game time (i.e., each feature sum was divided by the number of seconds elapsed at that time point). All features were then standardized within the 5-fold cross-validation, and then hyperparameter tuning of the classifiers occurred on an additional internal 3-fold cross-validation. The internal cross-validation split the training set into a training and validation set, which were iteratively used to compare a set range of model hyperparameters, including the minimum samples required per leaf node and the total number of trees.

4.1 Transfer Learning

To evaluate the effectiveness of transferring information from one source dataset to another target dataset in the multimodal predictive models, we constructed an unsupervised model of the shared modality between datasets: gameplay. Because both the datasets consisted of student gameplay with the same feature representation, it is possible to construct a model that learns a new, reduced feature set for this modality and apply it to a new target dataset. This idea is similar to the use of pre-trained language models (e.g., ELMo, BERT, T5), where we apply the pre-trained language model to domain-specific text and obtain a distributed vector representation of that text. Intuitively, a larger source dataset of gameplay data will allow for a more expressive unsupervised model to be trained, and this model can then be applied to new target dataset. We used a standard autoencoder, a type of neural network where the model attempts to reconstruct the input from a learned latent representation, to be used as the unsupervised learning method. In this analysis, we compared the use of an autoencoder on the gameplay data from each dataset separately. Specifically, we compared three conditions: (1) training an autoencoder from "scratch" on the current target dataset without leveraging any information from another source dataset (i.e., no transfer), (2) applying an autoencoder that was first pre-trained on the source dataset's gameplay data (similar to applying a pre-trained language model to a new text corpus), and (3) fine-tuning an autoencoder that was first pre-trained on the source dataset's gameplay data on the current target dataset's gameplay data (similar to fine-tuning a language model on a new domain-specific text corpus).

For Condition 1 (no transfer), the autoencoder is trained on the training data from the current target dataset. The encoder component of this autoencoder is then used to encode both the training and test data from the target dataset. Then, the RF classifier is trained on this encoded representation and the features from any additional modalities. For Condition 2 (transfer without fine-tuning), the autoencoder is trained on all the gameplay data from the source dataset, and the trained encoder component is used to encode both the training and testing gameplay data from the target dataset. The RF classifier is then trained on this representation as it was with Condition 1. For Condition 3 (transfer with fine-tuning), the autoencoder is trained on all the gameplay data from the source dataset, and then the autoencoder is further trained on the training data from the target dataset. The fine-tuned encoder is now used to encode both the training and testing gameplay data from the target dataset, and the RF classifier is trained and evaluated on this encoded data. Each autoencoder was constructed with a total of three hidden layers of dimensions 64, 32, and 64, meaning the innermost layer learned has a dimensionality of 32. The models were optimized with Adam [20], and training was terminated through early stopping with a patience of 5 or a maximum of 100 epochs. For each transfer learning experiment, we did not use any other feature reduction techniques in addition to the autoencoder (i.e., no PCA was performed in addition to the autoencoders for any modality). This was done to focus the comparisons on the effect of transfer learning. This work explores transfer learning across different versions of the same game-based learning environment. The student populations and types of data channels differed, making this approach fall under the framing of *covariate shift* and *domain adaptation*, both of which are included in the transfer learning paradigm [32].

5 RESULTS

To examine how well each combination of students' multimodal data classifies student post-test performance and interest in either a single-task or multi-task setting, we report the five-fold cross-validation results for each early prediction model. The critical metrics for this work include a metric from both early prediction (standardized convergence point, SCP) [27] and a standard classification metric (F1 score). SCP measures how early prediction models can consistently make accurate predictions while penalizing *non-converged* sequences (i.e., sequences whose last instance predictions are incorrect). For each model, we report the early prediction metrics of the best-performing model in terms of SCP across all possible data fusion and PCA reduction techniques and similarly report the classification results of the model with the highest F1 score across all possible data fusion techniques. Since SCP aims to identify model performance at early stages in the prediction sequence, lower values are better. F1 score aims to summarize how well the model is predicting the high groups (i.e., 1) by taking the harmonic mean of the precision and recall.

5.1 Crystal Island – Sensor-Based

We first report the performance of the predictive models with different combinations of modalities for classifying student post-test score performance and interest for Crystal Island – Sensor-Based in Tables 1 and 2, respectively. For each table, both the single-task performance and the multi-task performance are shown. For each modality configuration, we report the best performance as determined by SCP and F1 score, and the statistical significance for the metrics are based on these. The best performing data fusion techniques and PCA reduction dimensions varied across all modality configurations, so we do not report these aspects of the best performing modality configurations. Statistical tests were conducted to compare models to the gameplay-only baseline and to compare multi-task (MTL) and single-task (STL) models. We used the statistical tests to compare the best performing MTL models to the best performing STL models to reduce the overall number of statistical tests. All statistical tests were conducted using the one-sided Wilcoxon signed-rank test between the performance of the models in each cross-validation fold. This test is a non-parametric test because the results from each fold cannot be assumed to be normally distributed.

5.2 Crystal Island – Reflection

We now report the performance of the predictive models with different combinations of modalities for classifying student post-test score performance and interest for Crystal Island – Reflection in Tables 3 and 4, respectively. We again report both the single-task performance and the multi-task performance. For each modality configuration, we report the best performance as determined by SCP and F1 score, and the statistical significance for the metrics are based on these. The best performing data fusion techniques and PCA reduction dimensions again varied across all modality configurations.

5.3 Transfer Learning for Crystal Island – Sensor-Based

Next, we report the results of transfer learning for Crystal Island – Sensor-Based. Tables 5 and 6 illustrate the performance for post-test score and interest, respectively, but we now include instances where an autoencoder was used on the gameplay modality. For a fair comparison, we conducted transfer learning only on the modality combinations that include the gameplay modality. The non-transfer (NT) condition is compared to the transfer conditions of pre-trained (PT) and fine-tuned (FT).

5.4 Transfer Learning for Crystal Island – Reflection

Next, we report the results of transfer learning for Crystal Island – Reflection. Tables 7 and 8 illustrate the performance for post-test score and interest, respectively, but we now include instances where an autoencoder was used on the gameplay modality. For a fair comparison, we conducted transfer learning only on the modality combinations that include the gameplay modality.

6 DISCUSSION

We investigated the effectiveness of combining multimodal, multi-task, and early prediction components into a single, unified student

Table 1: Post-test score prediction results for Crystal Island – Sensor-Based. * indicates the model outperforms the gameplay-only model ($p < 0.05$), and $^\wedge$ indicates the MTL model outperforms the best STL model ($p < 0.05$). g, e, and f represent gameplay, eye gaze, and facial expressions, respectively. Bolded models outperform baselines in terms of SCP or F1 score.

			Single-Task					
Modalities	Maj. Class	{g}	**{e}**	{f}	{g, e}	{e, f}	**{g, f}**	{g, e, f}
SCP	0.531	0.554	**0.503***	0.593	0.541	0.655	**0.507***	0.590
F1	**0.694***	0.628	**0.682***	0.557	0.659	0.609	**0.715***	0.598
			Multi-Task					
Modalities	Maj. Class	{g}	**{e}**	{f}	**{g, e}**	{e, f}	**{g, f}**	{g, e, f}
SCP	0.531	0.557	**0.461***	0.556	**0.491***	0.572	**0.500***	0.583
F1	**0.694***	0.647	**0.699***	0.620	0.662	0.614	**0.693***	0.614

Table 2: Interest prediction results for Crystal Island – Sensor-Based. * indicates the model outperforms the gameplay-only model ($p < 0.05$), and $^\wedge$ indicates the MTL model outperforms the best STL model ($p < 0.05$). g, e, and f represent gameplay, eye gaze, and facial expressions, respectively. Bolded models outperform baselines in terms of SCP or F1 score.

			Single-Task					
Modalities	Maj. Class	{g}	{e}	{f}	**{g, e}**	{e, f}	**{g, f}**	**{g, e, f}**
SCP	0.579	0.600	0.698	0.593	**0.556***	0.595	**0.554***	**0.552***
F1	0.000	0.603	0.573	0.636	**0.717***	0.636	**0.702***	**0.717***
			Multi-Task					
Modalities	Maj. Class	{g}	{e}	{f}	**{g, e}**	{e, f}	**{g, f}**	**{g, e, f}**
SCP	0.579	0.575	0.714	0.551	**0.514***	0.580	**0.514***	**0.473***
F1	0.000	0.607	0.542	0.632	**0.692***	0.630	**0.698***	**0.769***

Table 3: Post-test score prediction results for Crystal Island – Reflection. * indicates the model outperforms the gameplay-only model ($p < 0.05$), and $^\wedge$ indicates the MTL model outperforms the best STL model ($p < 0.05$). g and t represent gameplay and reflection text, respectively. Bolded models outperform baselines in terms of SCP or F1 score.

		Single-Task		
Modalities	Maj. Class	{g}	**{t}**	**{g, t}**
SCP	0.474	0.525	**0.438***	**0.422***
F1	0.709	0.681	**0.756***	**0.726***
		Multi-Task		
Modalities	Maj. Class	{g}	**{t}**	**{g, t}**
SCP	0.474	0.524	**0.394***	**0.428***
F1	0.709	0.688	**0.740***	**0.749***

modeling framework. The goal of the framework is to classify students' post-test score performance and interest in the Crystal Island game-based learning environment. Using available multimodal data streams from the datasets, we investigated how effectively different combinations of modalities performed in classification models of the two target variables, both in a single-task and multi-task setting. Additionally, we evaluated the ability of each of these models to make accurate predictions at early points within students' gameplay. The results indicated that multimodal data can accurately predict both students' post-test scores and interest. Additionally, multi-task models were able to improve results in many cases when predicting student interest as compared to single-task models but did not markedly improve performance when predicting post-test scores. Transfer learning was able to improve results in cases where a larger amount of data was available in the source dataset and in cases where fewer modalities were available overall.

6.1 Performance of Multimodal Models of Post-Test Scores and Interest

For both predicting post-test scores and interest, we evaluated baseline models that only incorporated student gameplay data. The results in Table 1 for Crystal Island – Sensor-Based indicate

Table 4: Interest prediction results for Crystal Island – Reflection. * indicates the model outperforms the gameplay-only model ($p < 0.05$), and $^{\wedge}$ indicates the MTL model outperforms the best STL model ($p < 0.05$). g and t represent gameplay and reflection text, respectively. Bolded models outperform baselines in terms of SCP or F1 score.

		Single-Task		
Modalities	Maj. Class	{g}	**{t}**	**{g, t}**
SCP	0.596	0.669	**0.564***	**0.577***
F1	**0.673***	0.559	**0.627***	**0.634***
		Multi-Task		
Modalities	Maj. Class	{g}	**{t}**	**{g, t}**
SCP	0.596	0.658	**0.510***	**0.507***
F1	**0.673***	0.587	**0.649***	**0.695***

Table 5: Post-test score prediction results with transfer for Crystal Island – Sensor-Based. * indicates the transfer learning model outperforms the non-transfer (NT) model ($p < 0.05$), and $^{\wedge}$ indicates the MTL model outperforms the best STL model ($p < 0.05$). g, e, and f represent gameplay, eye gaze, and facial expressions, respectively. Bolded models outperform baselines in terms of SCP or F1 score.

						Single-Task						
Modalities	**{g}**	{g}	{g}	{g,e}	**{g,e}**	{g,e}	**{g,f}**	{g,f}	{g,f}	{g,e,f}	{g,e,f}	{g,e,f}
Transfer	**FT**	PT	NT	FT	**PT**	NT	**FT**	PT	NT	FT	PT	NT
SCP	**0.568***	0.597	0.608	0.702	**0.651***	0.715	**0.529***	0.540	0.584	0.663	0.659	0.669
F1	**0.620***	0.607	0.567	0.619	**0.648***	0.588	**0.663***	0.540	0.592	0.575	0.587	0.577
						Multi-Task						
Modalities	**{g}**	**{g}**	{g}	**{g,e}**	{g,e}	{g,e}	**{g,f}**	**{g,f}**	{g,f}	{g,e,f}	{g,e,f}	{g,e,f}
Transfer	**FT**	**PT**	NT	**FT**	PT	NT	**FT**	**PT**	NT	FT	PT	NT
SCP	**0.565***	**0.551***	0.605	**0.575***	0.622	0.656	**0.527***	**0.529***	0.585	0.649	0.647	0.630
F1	**0.669***	0.630	0.623	**0.643***	0.628	0.603	**0.664***	**0.666***	0.598	0.611	0.586	0.615

Table 6: Interest prediction results with transfer for Crystal Island – Sensor-Based. * indicates the transfer learning model outperforms the non-transfer (NT) model ($p < 0.05$), and $^{\wedge}$ indicates the MTL model outperforms the best STL model ($p < 0.05$). g, e, and f represent gameplay, eye gaze, and facial expressions, respectively. Bolded models outperform baselines in terms of SCP or F1 score.

						Single-Task						
Modalities	**{g}**	{g}	{g}	**{g,e}**	**{g,e}**	{g,e}	{g,f}	**{g,f}**	{g,f}	{g,e,f}	{g,e,f}	{g,e,f}
Transfer	**FT**	PT	NT	**FT**	**PT**	NT	FT	**PT**	NT	FT	PT	NT
SCP	**0.630***	0.676	0.698	0.768	**0.712***	0.783	0.679	**0.611***	0.673	0.669	0.616	0.643
F1	0.516	0.511	0.520	**0.599***	0.542	0.536	0.563	**0.608***	0.554	0.631	0.641	0.656
						Multi-Task						
Modalities	**{g}**	{g}	{g}	**{g,e}**	**{g,e}**	{g,e}	{g,f}	**{g,f}**	{g,f}	{g,e,f}	{g,e,f}	{g,e,f}
Transfer	**FT**	PT	NT	**FT**	**PT**	NT	FT	**PT**	NT	FT	PT	NT
SCP	**0.589**$^{*\wedge}$	0.697	0.709	**0.640***	**0.679***	0.741	0.667	**0.599***	0.653	0.673	0.700	0.646
F1	**0.573***	0.501	0.466	**0.641**$^{*\wedge}$	0.637	0.553	0.564	0.593	0.566	0.614	0.612	0.619

that in the single-task setting, both the unimodal model of eye gaze-only and the multimodal model of gameplay and facial expressions outperformed this gameplay-only baseline in both the SCP and F1 score metrics. It is notable that the combination of all three available modalities may be inhibitive for predicting post-test scores. For predicting interest, we found that several different combinations of modalities outperformed the gameplay-only baseline (Table 2). Specifically, the gameplay plus eye gaze, gameplay plus facial expressions, and gameplay plus eye gaze and facial expressions combinations all outperformed the gameplay-only single-task

Table 7: Post-test score prediction results with transfer for Crystal Island – Reflection. * indicates the transfer learning model outperforms the non-transfer (NT) model (p < 0.05), and ^ indicates the MTL model outperforms the best STL model (p < 0.05). g and t represent gameplay and reflection text, respectively. Bolded models outperform baselines in terms of SCP or F1 score.

			Single-Task			
Modalities	{g}	{g}	{g}	**{g,t}**	{g,t}	{g,t}
Transfer	FT	PT	NT	**FT**	PT	NT
SCP	0.473	0.477	0.476	**0.423***	0.464	0.473
F1	0.667	0.679	0.673	**0.752***	0.734	0.700
			Multi-Task			
Modalities	{g}	{g}	{g}	**{g,t}**	**{g,t}**	{g,t}
Transfer	FT	PT	NT	**FT**	**PT**	NT
SCP	0.451	0.457	0.479	**0.504***	**0.501***	0.553
F1	0.677	0.680	0.670	**0.731***	**0.731***	0.675

Table 8: Interest prediction results with transfer for Crystal Island – Reflection. * indicates the transfer learning model outperforms the non-transfer (NT) model (p < 0.05), and ^ indicates the MTL model outperforms the best STL model (p < 0.05). g and t represent gameplay and reflection text, respectively. Bolded models outperform baselines in terms of SCP or F1 score.

			Single-Task			
Modalities	{g}	{g}	{g}	**{g,t}**	**{g,t}**	{g,t}
Transfer	FT	PT	NT	**FT**	**PT**	NT
SCP	0.682	0.690	0.671	**0.620***	**0.613***	0.686
F1	0.508	0.520	0.550	**0.593***	**0.576***	0.522
			Multi-Task			
Modalities	{g}	{g}	{g}	**{g,t}**	**{g,t}**	{g,t}
Transfer	FT	PT	NT	**FT**	**PT**	NT
SCP	0.699	0.687	0.667	**0.579***	**0.590***	0.724
F1	0.491	0.505	0.500	**0.624***	**0.613***	0.474

model in terms of both SCP and F1 score. For Crystal Island – Reflection, the findings in Tables 3 and 4 reflect the same notion that by incorporating another modality in addition to gameplay, the early prediction models are better able to predict both post-test scores and interest, respectively. For this dataset, by adding reflection text as an additional modality to gameplay, the models are statistically better in terms of both SCP and F1 score.

We observed that reflection text alone in an early prediction model outperforms gameplay for predicting both post-test score performance and interest. It is possible that the models are better able to use direct feedback from students (e.g., topics they are struggling with) about their learning process (rather than students' in-game actions) as indicators of their knowledge and interest in the game. A game-based learning environment equipped with information about both students' gameplay and reflection text appears to be capable of making accurate predictions about students' post-test scores and interest. Additionally, we observed that the difference in performance between the post-test score models and interest models is much greater in Crystal Island – Reflection than it is for Crystal Island – Sensor-Based. Specifically, in Crystal Island – Reflection, the models of student post-test scores are slightly better than the models of student post-test scores in Crystal Island – Sensor-Based, and the opposite relationship is true for models of student interest. The key differences between the two datasets could influence this disparity. It is possible that in the case of predicting post-test performance, using student reflection text is more predictive than either eye gaze or facial expressions, when added to gameplay data. This is somewhat intuitive, because students are asked to reflect on their learning process and are directly providing information about what they know. This may not be the case for interest, however. Student facial expressions and eye gaze may be better at capturing affective states such as boredom or frustration, which may be more related to their interest levels in the game when compared to their reflection text.

6.2 Performance of Multi-Task Models of Post-Test Scores and Interest

For both predicting post-test scores and interest, we compared the performance of multi-task models of student post-test score and interest to their single-task counterparts as baselines. The results in Tables 1 and 2 for Crystal Island – Sensor-Based indicate that multi-task learning does not improve predictive performance significantly for post-test score prediction, but it does help significantly for interest prediction. In general, the same combination of modalities is predictive in the MTL setting compared to the STL setting for the respective target variables. However, the results indicate that predicting interest with MTL yields more improvement over STL models due to the joint prediction of post-test scores and interest. This finding highlights the connection between post-test scores and student interest in game-based learning, which aligns with previous theory [42]. For Crystal Island – Reflection, the findings in Tables 3 and 4 are also in support of multi-task models of student interest but not so for post-test scores. Specifically, there are no statistically significant differences between the best performing STL and MTL models for predicting post-test score performance in this dataset, but there are statistically significant differences in the models of student interest. This finding from both datasets could indicate that by also forecasting what a student knows, the models are better able to detect how interested they are in the game itself. This could be because students are more likely to be interested in their experience if they are learning about the subject matter. These findings have broad implications for predictive student modeling in game-based learning environments. While MTL was more useful in enhancing the predictions of student interest, prior work has found that as the number of tasks increases, predictive performance also increases for these tasks [11]. It is also possible to represent student knowledge by the student's mastery of individual concepts, such as by the student's predicted responses to individual post-test questions. This also would enable adaptive feedback mechanisms to pinpoint which specific areas the predictive model thinks the student needs support.

6.3 Performance of Transferred Models of Post-Test Scores and Interest

A final goal of this work was to investigate how the unified multimodal, multi-task, early prediction framework can be leveraged from one domain (i.e., the source dataset) and applied to another (i.e., the target dataset). To this end, we used the input information that was shared between the two domains: gameplay data. Because both datasets share the same core game-based learning environment, the gameplay logs from each dataset are very similar. To leverage information between the two datasets, we trained an autoencoder on the gameplay from a source dataset and applied the pre-trained model to the target dataset as a way of applying a feature reduction that had been previously trained. We compared the use of no transfer (NT) to transfer with two variations: 1) applying the pre-trained gameplay model to the current dataset's gameplay data as-is (PT), and 2) fine-tuning the pre-trained gameplay model to the current dataset's gameplay data in its training set (FT). Tables 5 and 6 illustrate the performance of these conditions for Crystal Island – Sensor-Based. The relationships between multimodal data versus unimodal data and MTL versus STL remain the same, but we note that by transferring the representation of the gameplay modality, we see a statistically significant increase in the predictive performance for both post-test scores and interest in many cases. In particular, we notice the biggest increase in performance when there are fewer modalities overall, meaning that gameplay is relied on more heavily for the prediction. For the results in these two tables for both transfer cases (PT and FT), the autoencoders were first trained on the gameplay data from Crystal Island – Reflection (*n*=118), which is a dataset of nearly double the size as Crystal Island – Sensor-Based (*n*=61). This increase in performance in this direction is a common characteristic of unsupervised machine learning models that were first trained on larger dataset. This is often seen in the field of natural language processing, where language models (e.g., ELMo, BERT, T5) are first trained on extremely large text corpora. The performance of the full set of modalities does not improve in the transferred setting, indicating that the models are more heavily relying on information from the sensor-based modalities compared to the gameplay modality. It is also noteworthy that both transfer conditions outperform the non-transfer condition at various points, with no clear best approach when multiple modalities are used. However, when gameplay is the only modality present, a transfer approach that involves fine-tuning appears to be superior in terms of the SCP and F1 score metrics. Tables 7 and 8 illustrate the performance of the transfer learning experiments for Crystal Island – Reflection. This means that for both transfer conditions (PT and FT), the autoencoders were trained first on gameplay data from Crystal Island – Sensor-Based and either applied as-is or fine-tuned on the new gameplay data, respectively. While it appears that models that transfer the gameplay representation outperform the non-transfer models for both post-test score and interest prediction for the multimodal combinations, we note that the gameplay-only models do not benefit from transfer. This finding is supported by the earlier point concerning the sizes of the two datasets. The Crystal Island – Sensor-Based dataset is much smaller, so a model first trained on this dataset is less likely to boost performance of a model that is trained and evaluated on a much larger dataset. It is notable, however, that the models incorporating both gameplay and reflection text are improved by leveraging gameplay data from the Crystal Island – Sensor-Based dataset. This is likely due to the models more heavily relying on text for these predictive tasks and benefiting from the richness of the previous dataset. As a general point, the transfer appears to be much more effective when first pre-training on the Crystal Island – Reflection dataset and applying the gameplay representation from the autoencoder to the Crystal Island – Sensor-Based dataset compared to the opposite direction. The transfer results highlight the promise of using similar data that was collected from a previous version of the game-based learning environment. More broadly, when there are limited modalities available, it is critical to leverage information from other source datasets to improve predictive model performance.

7 CONCLUSION

Predictive student models enable game-based learning environments to adapt to individual students' needs in real-time. Advances in sensor-based technologies and natural language processing introduce the opportunity to leverage multimodal data channels during game-based learning. We developed a unified student modeling framework consisting of multimodal learning analytics, multi-task learning, transfer learning, and early prediction. When combined, the unified framework was evaluated using two datasets collected from student interactions with the Crystal Island game-based learning environment. Approaches to improve the predictive performance of the framework by transferring the trained model from one dataset to another were also investigated. The framework was evaluated by early prediction convergence metrics as well as standard performance metrics on all predictions. Evaluations demonstrated that using an autoencoder to encode the gameplay data from one game-based learning dataset and applying the encoding to the other dataset improved predictive model results for predicting both post-test scores and interest. Multi-task learning was able to improve predictive results for interest, but not post-test scores. These findings, and the overall unified framework, advance the field of predictive student modeling by creating a unified framework to drive adaptive and personalized learning.

The findings presented here suggest several promising directions for future work. First, exploring more expressive and effective feature representations for each modality will be important. For all modalities, we used a static feature representation to both synchronize each input source and inspect the relationships between each modality with both student post-test performance and interest after game-based learning. A promising alternative could be a temporal representation that incorporates the dynamic nature of the data. A sequential model that leverages this representation could achieve more accurate, fine-grained early predictions of student post-test performance and interest. Second, it will be critical to investigate other modeling techniques (e.g., deep learning) to further improve predictions. With the insight of which features perform well in both the predictive tasks, more sophisticated modeling techniques may be able to achieve even higher accuracy. More advanced transfer learning approaches (e.g., adversarial-based domain adaptation) could further improve results. A final promising direction is to investigate how multimodal models can be incorporated into game-based learning environments to support real-time adaptive scaffolding. This will set the stage for empirically assessing the efficacy of early predictive models that integrate students' multimodal data to improve student learning outcomes.

ACKNOWLEDGMENTS

This study was supported by the funding from both the Social Sciences and Humanities Research Council of Canada (SSHRC 895-2011-1006) and the National Science Foundation (DRL-1661202). Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the SSHRC or NSF.

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. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3313831

[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 (CHI '19). Association for Computing Machinery, New York, NY, USA, Paper 304, 1–12. https://doi.org/10.1145/3290605.3300534

[3] Roger Azevedo and Dragan Gašević. 2019. Analyzing multimodal multichannel data about self-regulated learning with advanced learning technologies: Issues and challenges. Computers in Human Behavior, 96, 207–210. https://doi.org/10.1016/j.chb.2019.03.025

[4] Paulo Blikstein and Marcelo Worsley. 2016. Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. Journal of Learning Analytics, 3(2), 220–238. https://doi.org/10.18608/jla.2016.32.11

[5] Ritwick Chaudhry, Harvineet Singh, Pradeep Dogga, and Shiv Kumar Saini. 2018. Modeling Hint-Taking Behavior and Knowledge State of Students with Multi-Task Learning. In Proceedings of the 11th International Conference on Educational Data Mining, 21-31.

[6] Sidney K. 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, 147–187.

[7] Paul Ekman and Erika Rosenberg. 1997. What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS). New York, NY: Oxford University Press.

[8] Andrew Emerson, Elizabeth B. Cloude, Roger Azevedo, and James Lester. 2020. Multimodal learning analytics for game-based learning. British Journal of Educational Technology, 51(5), 1505-1526.

[9] 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 Proceedings of the 2020 International Conference on Multimodal Interaction. Association for Computing Machinery, New York, NY, USA, 107–116. https://doi.org/10.1145/3382507.3418890.

[10] Michael Geden, Andrew Emerson, Dan Carpenter, Jonathan Rowe, Roger Azevedo, and James Lester. 2021. Predictive student modeling in game-based learning environments with word embedding representations of reflection. International Journal of Artificial Intelligence in Education, 31, 1-23.

[11] Michael Geden, Andrew Emerson, Jonathan Rowe, Roger Azevedo, and James Lester. 2020. Predictive student modeling in educational games with multi-task learning. In Proceedings of the AAAI Conference on Artificial Intelligence, 34(01), 654-661. https://doi.org/10.1609/aaai.v34i01.5406

[12] Nathan Henderson, Wookhee Min, Andrew Emerson, Jonathan Rowe, Seung Lee, James Minogue, and James Lester. 2021. Early Prediction of Museum Visitor Engagement with Multimodal Adversarial Domain Adaptation. In Proceedings of the 14th International Conference on Educational Data Mining, 93-104.

[13] Nathan Henderson, Wookhee Min, Jonathan Rowe, and James Lester. 2020. Enhancing Affect Detection in Game-Based Learning Environments with Multimodal Conditional Generative Modeling. Proceedings of the 2020 International Conference on Multimodal Interaction. Association for Computing Machinery, New York, NY, USA, 134–143. https://doi.org/10.1145/3382507.3418892

[14] Nathan Henderson, Wookhee Min, Jonathan Rowe, and James Lester. 2021. Enhancing Multimodal Affect Recognition with Multi-Task Affective Dynamics Modeling. In Proceedings of the 9th International Conference on Affective Computing and Intelligent Interaction (ACII), 1-8. IEEE.

[15] iMotions. 2016. Attention Tool, 6.2. Boston, MA: iMotions Inc.

[16] G. Tanner Jackson and Danielle S. McNamara. 2013. Motivation and performance in a game-based intelligent tutoring system. Journal of Educational Psychology, 105(4), 1036-1049.

[17] Shamya Karumbaiah, Ryan Baker, and Valeria Shute. 2018. Predicting quitting in students playing a learning game. In Proceedings of the 12th International Conference on Educational Data Mining, 167–176.

[18] Angelika Kasparova, Oya Celiktutan, and Mutlu Cukurova. 2020. Inferring Student Engagement in Collaborative Problem Solving from Visual Cues. In Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion). Association for Computing Machinery, New York, NY, USA, 177–181. https://doi.org/10.1145/3395035.3425961

[19] Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7482–7491. IEEE Computer Society.

[20] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[21] Y. Alex Kolchinski, Sherry Ruan, Dan Schwartz, and Emma Brunskill. 2018. Adaptive natural-language targeting for student feedback. In Proceedings of the Fifth Annual ACM Conference on Learning at Scale (L@S '18). Association for Computing Machinery, New York, NY, USA, Article 26, 1–4. https://doi.org/10.1145/3231644.3231684

[22] Mei-Jen Kuo. 2007. How does an online game-based learning environment promote students' intrinsic motivation for learning natural science and how does it affect their learning outcomes? In 2007 First IEEE International Workshop on Digital Game and Intelligent Toy Enhanced Learning, 135-142. IEEE.

[23] James C. Lester, Eun Y. Ha, Seung Y. Lee, Bradford W. Mott, Jonathan P. Rowe, and Jennifer L. Sabourin. 2013. Serious games get smart: Intelligent game-based learning environments. AI Magazine, 34(4), 31-45. https://doi.org/10.1609/aimag.v34i4.2488

[24] Ran Liu, John C. Stamper, and Jodi Davenport. 2018. A novel method for the in-depth multimodal analysis of student learning trajectories in intelligent tutoring systems. Journal of Learning Analytics, 5(1), 41-54. https://doi.org/10.18608/jla.2018.51.4

[25] Ye Mao, Farzaneh Khoshnevisan, Thomas Price, Tiffany Barnes, and Min Chi. 2022. Cross-Lingual Adversarial Domain Adaptation for Novice Programming.

In Proceedings of the AAAI Conference on Artificial Intelligence.

[26] Lynn McAlpine, Cynthia Weston, Catherine Beauchamp, C. Wiseman, and Jacinthe Beauchamp. 1999. Building a metacognitive model of reflection. Higher education, 37(2), 105-131.

[27] 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 International Conference on Artificial Intelligence and Interactive Digital Entertainment.

[28] Wookhee Min, Megan Frankosky, Bradford W. Mott, Jonathan P. Rowe, Andy Smith, Eric Wiebe, Kristy Elizabeth Boyer, and James C. Lester. 2020. DeepStealth: Game-based learning stealth assessment with deep neural networks. IEEE Transactions on Learning Technologies, 13(2), 312-325.

[29] Wookhee Min, Randall Spain, Jason D. Saville, Bradford Mott, Keith Brawner, Joan Johnston, and James Lester. 2021. Multidimensional team communication modeling for adaptive team training: A hybrid deep learning and graphical modeling framework. In Proceedings of the Twenty-Second International Conference on Artificial Intelligence in Education (pp. 293-305). Springer, Cham.

[30] Xavier Ochoa and Marcelo Worsley. 2016. Augmenting learning analytics with multimodal sensory data. Journal of Learning Analytics, 3, 213–219.

[31] 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 - Volume 2. Association for Computing Machinery and Morgan & Claypool, 331–374. https://doi.org/10.1145/3107990.3108003

[32] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[34] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 1532– 1543.

[35] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2227–2237.

[36] Jan L. Plass, Richard E. Mayer, and Bruce D. Homer. 2020. Handbook of game-based learning. Cambridge, MA: MIT Press.

[37] Konstantinos Pliakos, Seang-Hwane Joo, Jung Yeon Park, Frederik Cornillie, Celine Vens, and Wilm Van den Noortgate. 2019. Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems. Computers & Education, 137, 91-103.

[38] Luis P. 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 (LAK '16). Association for Computing Machinery, New York, NY, USA, 148–157. https://doi.org/10.1145/2883851.2883927

[39] Ramkumar Rajendran, Anurag Kumar, Kelly E. Carter, Daniel T. Levin, and Gautam Biswas. 2018. Predicting Learning by Analyzing Eye-Gaze Data of Reading Behavior. In Proceedings of the 12th International Conference on Educational Data Mining, 455-461.

[40] Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin, 124(3), 372-422.

[41] Richard M. Ryan. 1982. Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. Journal of Personality and Social Psychology, 43, 450–461.

[42] Richard M. Ryan and Edward L. Deci. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist, 55, 68–78.

[43] 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 (LAK '20). Association for Computing Machinery, New York, NY, USA, 480–489. https://doi.org/10.1145/3375462.3375498

[44] Samuel Spaulding, Jocelyn Shen, Haewon Park, and Cynthia Breazeal. 2021. Towards Transferrable Personalized Student Models in Educational Games. In Proceedings of the 20th International Conference on Autonomous Agents and Multi Agent Systems, 1245-1253.

[45] Yu Su, Qingwen Liu, Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Chris Ding, Si Wei, and Guoping Hu. 2018. Exercise-enhanced sequential modeling for student performance prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, 32(01), 2435-2443.

[46] Chaohong Tan and Zhenhua Ling. 2019. Multi-Classification Model for Spoken Language Understanding. In 2019 International Conference on Multimodal Interaction (ICMI '19). Association for Computing Machinery, New York, NY, USA, 526–530. https://doi.org/10.1145/3340555.3356099

[47] Michelle Taub, Nicholas Mudrick, Amanda E. Bradbury, and Roger Azevedo. 2019. Self-regulation, self-explanation, and reflection in game-based learning. In J. Plass, B. Horner, & R. Mayer (Eds.), Handbook of game-based learning, 239-262. Boston, MA: MIT Press.

[48] 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, 103781.

[49] Güray Tonguç and Betul Ozaydın Ozkara. 2020. Automatic recognition of student emotions from facial expressions during a lecture. Computers & Education, 148, 103797.

[50] Christos Troussas, Akrivi Krouska, and Cleo Sgouropoulou. 2020. Collaboration and fuzzy-modeled personalization for mobile game-based learning in higher education. Computers & Education, 144, 103698.

[51] Gerard Van den Boom, Fred Paas, and Jeroen JG VanMerrienboer. 2007. Effects of elicited reflections combined with tutor or peer feedback on self-regulated learning and learning outcomes. Learning and Instruction, 17(5), 532–548.

[52] Marcelo Worsley. 2018. (Dis)engagement matters: identifying efficacious learning practices with multimodal learning analytics. In Proceedings of the 8th International Conference on Learning Analytics and Knowledge. Association for Computing Machinery, New York, NY, USA, 365–369.