Early Prediction of Museum Visitor Engagement with Multimodal Adversarial Domain Adaptation

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

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

ABSTRACT

Recent years have seen significant interest in multimodal frameworks for modeling learner engagement in educational settings. Multimodal frameworks hold particular promise for predicting visitor engagement in interactive science museum exhibits. Multimodal models often utilize video data to capture learner behavior, but video cameras are not always feasible, or even desirable, to use in museums. To address this issue while still harnessing the predictive capacities of multimodal models, we investigate adversarial discriminative domain adaptation for generating modality-invariant representations of both unimodal and multimodal data captured from museum visitors as they engage with interactive science museum exhibits. This approach enables the use of pre-trained multimodal visitor engagement models in circumstances where multimodal instrumentation is not available. We evaluate the visitor engagement models in terms of early prediction performance using exhibit interaction and facial expression data captured during visitor interactions with a science museum exhibit for environmental sustainability. Through the use of modality-invariant data representations generated by the adversarial discriminative domain adaptation framework, we find that pre-trained multimodal models achieve competitive predictive performance on interaction-only data compared to models evaluated using complete multimodal data. The multimodal framework outperforms unimodal and non-adapted baseline approaches during early intervals of exhibit interactions as well as entire interaction sequences.

Keywords

Museum learning, visitor engagement, adversarial domain adaptation, early prediction, multimodal learning analytics.

1. INTRODUCTION

Visitor engagement is critical in museum learning [21]. Engagement defines how visitors experience museums, including how they move between exhibits, form and express interests, and acquire knowledge and understanding. Developing computational models of museum visitor engagement holds significant promise for identifying salient patterns of visitor behavior as well as providing insight into how specific exhibits can be designed to enhance engagement. For example, visitor analytics show potential for enabling adaptive learning experiences tailored to the preferences and tendencies of the visitors, leading to highly engaged interactions with the exhibit. Visitor interactions with museum exhibits are inherently *multimodal*. Visitor engagement manifests through a variety of behaviors such as facial expression, touch, eye gaze, and body posture. As such, multimodal learning analytics can model museum visitor engagement by capturing and analyzing visitor behavior from several different perspectives [2, 16]. Multimodal models of learner engagement have been shown to be effective in a range of environments, including laboratory [8, 22] and classroom settings [1, 6, 7]. More recently, multimodal learning analytics have been the subject of growing attention in informal education settings, such as museums [16, 20], but this line of investigation is still in its nascent stages.

Given the multimodal nature of visitor interactions in museums, the use of multichannel data provides important benefits for modeling visitor engagement. In particular, multimodal models can be used to predict visitor engagement early during a visitor's interaction with an exhibit [16]. This shows promise for enabling visitor-adaptive technologies that provide adaptive support for fostering engaged learning experiences with an exhibit or for notifying museum educators to inform decisions about staffing the museum floor. In predictive modeling, it is important that the multimodal visitor engagement models be evaluated in terms of both predictive accuracy and the minimum amount of time that the models require to achieve robust predictive performance.

Multimodal modeling of visitor engagement in museums also poses significant challenges. Interactions with exhibits are highly variable due to the free-choice nature of museum learning [12, 25, 28]. Additionally, multimodal frameworks often utilize physical sensors (e.g., video cameras, motion sensors, eye trackers), which introduce questions about scalability, privacy, and mistracking. Intrusiveness is also a concern, as suites of multimodal sensors may be impractical in some settings, or they may adversely affect the natural behavior of visitors [32].

Transfer learning presents itself as a natural solution to this issue, as the various modalities in a multimodal modeling framework share a common predictive task. In particular, recent years have seen an increased emphasis on domain adaptation, a type of transfer learning that investigates the predictive capacity of models that are pre-trained on one domain (*source* domain) and are subsequently reweighted to perform similarly on another domain with a different distribution (*target* domain) across a single common task [39]. A primary objective of domain adaptation is to obtain a domain-invariant representation of the salient features extracted from the two distinct data sources, where the shared feature space allows for improved predictive performance on data points from the target domain while still maintaining strong performance on data from the source domain. Examples of recent domain adaptation research include adapting across images extracted from different domains [34, 42] or across modalities captured from different data channels such as RGB-to-depth image translation [33, 42].

In this work, we investigate the use of domain adaptation as a method of translating unimodal, interaction-based data to a domain-invariant representation that can be used with predictive models previously trained on multimodal data. We demonstrate the effectiveness of a multimodal domain adaptation framework for making early predictions of visitor dwell times at an interactive museum exhibit. Our multimodal analytics framework is designed to operate in museum settings where sensor-based data capture may be restricted or otherwise impractical. We adopt an adversarial approach to generating domain-invariant representations of multimodal data (exhibit interactions and facial expression serving as the *source* domain) and unimodal data (exhibit interactions serving as the *target* domain) that are encoded using stacked denoising autoencoders. Empirical results of evaluations of the framework suggests that the use of adversarial discriminative domain adaptation allows for a unimodal target encoder to be trained to share a latent feature space with a multimodal source encoder [42]. The framework achieves higher performance than an interaction-only baseline model in terms of early prediction and visitor-level prediction of dwell time, a proxy indicator of visitor behavioral engagement with an exhibit. Dwell time has been frequently used to quantify visitor engagement in museum settings [5, 23]. The framework offers the potential to accurately predict visitor dwell time in museums, while also allowing for operation with reduced availability of physical sensor data, or even when no physical sensor data is available.

2. RELATED WORK

Visitor engagement is a critical aspect of learning in informal learning environments, such as science centers and museums [21]. Engagement shapes how visitors proceed throughout a museum, and interact with various exhibits [16]. There has been substantial work on modeling engagement in formal learning environments such as classrooms [19] and laboratories [8], and this focus has expanded in recent years to informal learning environments. This includes research efforts focused on analyzing engagement in groups of visitors around interactive tabletop exhibits [5], investigating the effectiveness of interventions for enhancing group engagement at different diorama exhibits [23], and predicting visitor dwell time [16]. However, devising computational models of museum visitor engagement remains a relatively unexplored area and presents distinctive challenges due to the free-choice nature of visitor learning in museums, creating a need for data-rich engagement modeling techniques.

Multimodal engagement modeling has shown significant promise as an engagement modeling approach due to its capacity to provide a data-rich multi-dimensional perspective on learner behavior [2]. In many cases, multimodal models lead to improved performance compared to models that utilize a single modality [19, 22, 32, 49]. Multimodal models have often utilized several diverse data channels when deployed in formal learning environments, including facial expression, posture, eye gaze, dialogue, and interaction trace data [40]. Facial expression data is commonly used in multimodal learner models of student affect [7] and performance [44]. Posture data has also been used for affect detection [22] as well as predicting learners' levels of engagement with Massive Online Open Courses (MOOCs) [9]. Eye gaze data has been combined with facial expression and head pose data to train models for continuous emotion prediction [48], while dialogue data has been utilized to predict dropout in online K-12 courses [26]. Finally, interaction trace logs and keystroke data have been used in conjunction with facial expression data to detect confusion in students engaging with an introductory computer science education learning environment to provide adaptive feedback and support dynamic adjustment of exercise difficulty levels [6]. While recent work has investigated multimodal approaches to modeling visitor engagement in museums [16], multimodal approaches to museum visitor modeling poses significant challenges, as these frameworks often necessitate physical, sensor-based data capture. This introduces various ethical and logistical concerns and may be impractical or prohibitive in certain informal learning environments.

Computational methods such as transfer learning, and particularly domain adaptation, provide a way to harness the predictive capacities of multimodal learning analytics while allowing visitor modeling frameworks to operationalize a reduced number of more intrusive modalities. Domain adaptation and transfer learning have shown significant potential in a variety of implementations, and have been utilized within educational contexts for tasks such as confusion detection in online forums for different online courses [50] and automated essay scoring across different prompts [35]. Additionally, domain adaptation has been investigated within multimodal contexts such as RGB and depth images [42], as well as video and audio modalities [36]. To our knowledge, adversarial domain adaptation has not been applied to unimodal and multimodal data to model learner engagement in museums.

Recent domain adaptation work has focused primarily on an unsupervised or semi-supervised variation of this problem, where deep learning models trained on a labeled source dataset are transferred to share latent representations alongside a target domain that may contain little or no previously labeled data. The issue of missing labels for the target domain data is addressed by obtaining a domain-invariant representation through minimizing the distance between the learned data representations between the two domains [17, 41, 42]. While prior efforts accomplish this task through statistical measures such as the Maximum Mean Discrepancy (MMD) [43] or the deep Correlation Alignment (CORAL) [39], other work has taken an adversarial approach, with the simultaneous goals of learning a data representation that is predictive of the source domain labels while also being indistinguishable to a domain discrimination model [27, 42]. One approach involves reversing the gradients of a domain discrimination model to maximize the model's loss and guide the learning to explore a domain-invariant representation [17]. Other approaches train a source encoder to reduce the source domain data to a latent representation and use a domain discriminator to adversarially train a target encoder to produce a latent representation of the target domain data that is indistinguishable to the discriminator [42]. The trained target encoder is subsequently used to process unlabeled data from the target domain to be classified by a model pre-trained on source data. Another approach is the Co-GAN approach, which involves two Generative Adversarial Networks (GANs) that generate source and target data, respectively [27]. The high layer parameters of the two GAN models are tied together, allowing the generators of the models to co-learn a shared latent representation while possibly sharing a common input noise vector.

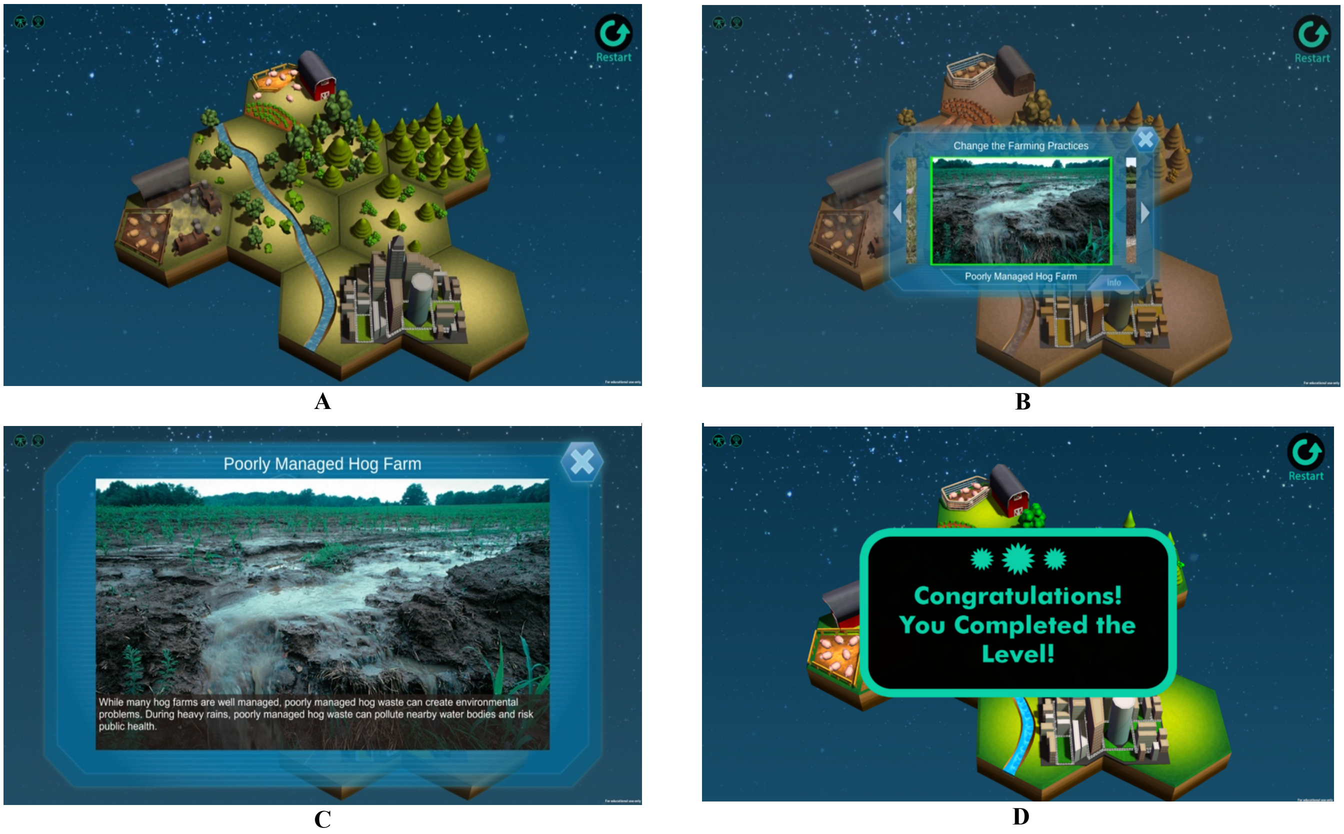
Early prediction is an important component of visitor modeling because it can drive run-time adaptive support to enhance visitor interest and engagement with interactive exhibits. A critical objective in early prediction is to reach a certain accuracy threshold in a timely manner. Early prediction has been investigated in the context of formal learning environments, such as predicting middle-grade learner engagement with a game-based learning environment [47], evolving learning goals throughout students' interaction trajectories [31], and student success in novice programming tasks [29]. Early prediction has also been the subject of prior work on museum learning, such as automatic detection of visitors' social behavioral patterns [13, 24] and multimodal regression-based modeling of visitor engagement in science museums [16].

The primary contributions of this work are as follows: (1) we demonstrate improved predictive performance of multimodal models of museum visitor dwell time using facial expression and interaction data compared to interaction-only baselines, (2) we evaluate the effectiveness of adversarial discriminative domain adaptation as a means of enabling the use of previously-trained multimodal models with unimodal data, and (3) we investigate the performance of each visitor engagement model using convergence-based early prediction metrics and standard predictive performance measures. Domain adaptation has been relatively underexplored with educational data, and this is especially true of data from informal learning environments such as museums. Furthermore, there has been limited work investigating domain adaptation in the context of early prediction of learner engagement. Our work shows that domain adaptation is effective at enhancing prediction of visitor dwell time by harnessing the capacities of multimodal visitor modeling, which leads to higher predictive accuracy when compared to unimodal models.

3. FUTURE WORLDS EXHIBIT

To investigate multimodal predictive models of museum visitor engagement, we use data collected from visitor interactions with a game-based museum exhibit, FUTURE WORLDS, which is designed to introduce visitors to concepts about environmental sustainability (Figure 1). FUTURE WORLDS runs on a multi-touch display, enabling visitors to interact with the virtual environment through touch and gestures on the screen. Visitors are faced with the challenge of improving the conditions of the virtual environment's biosphere through a series of changes such as farming practices and energy sources within the game. FUTURE WORLDS and its integrated educational content are targeted towards learners ages 10-11.

Visitors can tap or swipe on the screen to perform certain actions such as reading about a particular aspect of the virtual environment and its impact on sustainability or modifying an in-game element and observing the broader consequences of this decision on the environment. Upon making a change to the virtual environment, the visitor is given immediate feedback regarding the positive or negative impact of the gameplay action. A visitor can "win" by making the correct decisions to certain in-game elements that maximize the environmental sustainability of the virtual environment. Afterwards, the visitor is presented with the option to restart the game or continue interacting with the virtual environment in its completed state. Additionally, a visitor is able to leave the FUTURE WORLDS exhibit having not completed the game beforehand. Prior work with FUTURE WORLDS found that visitors improved their understanding of environmental sustainability


Restart
Change the Farming Practices
Poorly Managed Hog Farm
Info
A
B
Poorly Managed Hog Farm
C
Congratulations! You Completed the Level!
D

While many hog farms are well managed, poorly managed hog waste can create environmental problems. During heavy rains, poorly managed hog waste can pollute nearby water bodies and risk public health.

Figure 1. Gameplay of the FUTURE WORLDS interactive exhibit, including (A) 3D virtual environment, (B) selecting an element to modify, (C) viewing information about the selected element, and (D) correctly solving the in-game problem.

concepts, while also demonstrating high levels of engagement throughout their interactions with the exhibit [37].

4. MULTIMODAL DATA COLLECTION

To track visitor engagement and behavior with Future Worlds, the exhibit was instrumented with several sensors to collect the real-time behavior of visitors' interactions with the exhibit, as shown in Figure 2. We first describe the visitor population for study participants and then introduce the two modalities used for the domain adaptation approach (facial expression, exhibit interaction trace logs), and the features extracted from each input data channel.

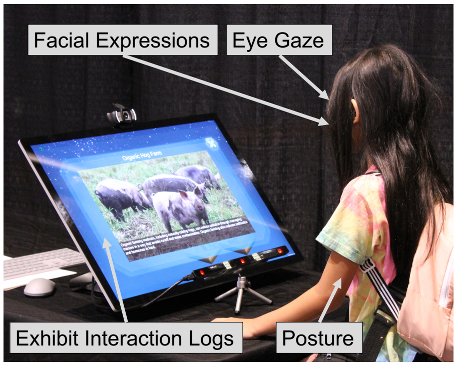


Figure 2. Visitor interacting with Future Worlds.

4.1 Study Participants and Procedure

We conducted a study of visitor interactions with the Future Worlds exhibit at the North Carolina Museum of Natural Sciences in Raleigh, North Carolina. The data were collected over a series of three sessions with different school groups of visitors aged 10-11 (*M*=10.4, *SD*=0.57). The school groups came from different socio-cultural backgrounds (e.g., race/ethnicity), and each school served student populations where 70% of the students are from low-income families. In total, 116 visitors interacted with Future Worlds. There were 47 female and 55 male participants, with 14 participants who did not provide data on their gender. The visitors were 32.4% Hispanic or Latino, 21.6% African American, 11.8% American Indian, 8% Asian, 7.5% mixed race, 3% Caucasian, and 15.7% preferred not to respond. Before interacting with the exhibit, visitors were asked to complete a series of surveys and questionnaires, including a demographic survey, sustainability content knowledge assessment, and the Fascination in Science scale [11]. Afterward, visitors interacted with the exhibit until they wanted to stop or after approximately 12 minutes had elapsed (*M*=5.8, *SD*=2.4, *Min*=1.8, *Max*=11.8). Visitor dwell times were captured by the game's internal logging functionalities. Once visitors finished their interaction with the exhibit, they were asked to complete a sustainability content knowledge assessment, engagement survey, and a short debrief interview. Several visitors were missing one or multiple data channels (e.g., facial mistracking), requiring the removal of their data from the final dataset for analysis. The final dataset that was used for the predictive models in this paper consisted of multimodal data from 79 visitors.

During the data collections, the visitors' body movement, eye gaze, facial expression, and interaction data from the exhibit were captured. For this study, we focus exclusively on the exhibit interaction data and the facial expression data. We selected the exhibit interaction data due to its unintrusive nature and its relative ease of data capture, as the trace data is captured in the background with the exhibit software and does not require any physical sensors or calibration. We selected facial expression data because of its predictive utility in previous work on unimodal and multimodal models of learner engagement [14, 15].

4.1.1 Facial Expression

Facial expression is an important indicator of learner emotion, and it has been widely used in previous studies on modeling learner engagement [46]. In this work, visitor facial expression was captured using video data from an externally mounted Logitech C920 USB webcam. In real time, the video data was processed by OpenFace, an open-source facial behavior analysis toolkit to detect facial landmarks, estimate head pose, recognize facial action units (AUs), and estimate eye gaze [3]. The OpenFace software automatically detects and analyzes 17 distinct AUs for each visitor's face captured within the camera's field of view.

4.1.2 Interaction Trace Logs

Future Worlds includes built-in logging functionalities to capture fine-grained logs of visitor interactions with the exhibit. The interaction trace logs consist of sequential records (at the millisecond level) of physical interactions with the multi-touch display (e.g., taps, swipes, and gestures), as well as specific in-game learning events (e.g., altering the virtual environment and accessing an embedded informational resources). The interaction trace logs are used to investigate how visitors interacted with the exhibit and progressed through the game.

4.2 Multimodal Features

Using both visitors' facial expression and exhibit interaction behavior, we distilled two sets of features to serve as predictors of visitor dwell time. Many of the extracted features for each modality were chosen based on their predictive performance in prior work on multimodal learning analytics [16].

4.2.1 Facial Expression

Using the processed AU data from OpenFace, we calculated the duration that each AU was exhibited throughout the visitor's interaction with Future Worlds. We first standardized each visitor's observed AU intensity values and then calculated the duration of each AU during time intervals where consecutive AU intensity values were at least one standard deviation greater than the mean of that particular visitor-specific AU feature. This filtering process ensured that only spikes relative to the specific visitor's AU values contributed towards the calculation of the total duration. To further filter the AU durations, we only recorded the duration if the AU was present for longer than 0.5 consecutive seconds. This avoided possible micro-expressions that could add noise to the overall data channel [38]. We performed this filtering process for all 17 AUs tracked by OpenFace. In addition, we generated the standard deviation and maximum AU values across the visitor's interactions up to the current timestamp. In total, we extracted and distilled 51 facial expression-related features.

4.2.2 Exhibit Trace Logs

We distilled eight features from the exhibit interaction data: (1) the total number of times a visitor tapped the Future Worlds multi-touch display, (2) the total number of times a visitor tapped informational tiles about environmental sustainability concepts, (3) the total duration of time an informational tile was open, (4) the total duration spent with labeled sustainability images displayed onscreen, (5) the total duration of time that a visitor spent directly interacting with the 3D simulated environment in Future Worlds, (6) the total number of times a visitor swiped the interface to explore alternative options for modifying the simulated environment, (7) the total number of times the simulated environment was modified, and (8) a binary feature that indicated whether a visitor had successfully solved the current environmental problem scenario in FUTURE WORLDS.

5. DOMAIN ADAPTATION

In this work, we present an unsupervised, adversarial discriminative domain adaptation approach that enables the use of multimodal visitor engagement models in settings where only unimodal data streams are available. In unsupervised domain adaptation, two datasets are extracted from two separate domains: (1) a source domain (s), from which data samples X_s and associated labels Y_s are drawn, and (2) a target domain (t), which contains unlabeled data samples X_t. It is also assumed that there exists a classifier C_s that has been previously trained on the source data X_s and source labels Y_s by learning a latent mapping M_s. The primary objective of the unsupervised domain adaptation approach is to learn a latent mapping M_t so that $M_t(X_t)$ can be correctly classified by C_s despite the absence of any associated labels for X_t.

The purpose of adversarial training within the domain adaptation framework is to learn a domain-invariant data representation that minimizes the distance between $M_t(X_t)$ and $M_s(X_s)$. This is accomplished through a separate binary discriminator, D, that is trained to distinguish between latent representations of the source domain and the target domain. The discriminator is optimized according to a standard cross-entropy loss function (Equation 1):

$$\begin{aligned}\mathcal{L}_{DISC}\,(X_s, X_t, M_s, M_t) \\ = -\mathbb{E}_{x_s \sim X_s}\left[\log D\left(M_s(X_s)\right)\right] \\ - \mathbb{E}_{x_t \sim X_t}\left[\log\left(1 - D\left(M_t(X_t)\right)\right)\right]\end{aligned} \qquad (1)$$

Adversarial domain adaptation focuses on two primary objectives implemented within a minmax framework: the discriminator attempts to accurately classify a latent data representation as either from the source domain or the target domain, while a target encoder attempts to learn a mapping $M_t(X_t)$ that deceives the discriminator, thus finding a latent representation that is domain-invariant but retains enough salient characteristics to provide predictive value to a source classifier C_s. To implement an adversarial loss function within the framework, a common practice is to simply invert the loss term when training the target encoder. This essentially reverses the gradients for the target encoder but can consequently lead to premature convergence and vanishing gradients [17]. A more stable training method is to invert the labels used to train the target encoder. This creates two distinct convergence objectives for the two elements of the adversarial framework [42]. The discriminator loss term remains the same as stated in Equation 1 above, while the loss term for the target encoder becomes:

$$\mathcal{L}_{TAR}\,(X_s, X_t, D) = -\mathbb{E}_{x_t \sim X_t}\left[\log D\left(M_t(X_t)\right)\right] \qquad (2)$$

This process is analogous to the process utilized by generative adversarial networks (GANs) [18]. A GAN attempts to emulate a fixed data distribution by adversarially training a discriminator to distinguish between "fake" data, which was produced by a generator that aims to generate data that is synthetic but realistic looking using a random noise vector, and "real" data that is extracted from the prior fixed data distribution. While GANs have been utilized in domain adaptation tasks [27], they are typically effective when the source and target domains are relatively similar. GANs have shown convergence issues in scenarios involving a high degree of domain shift [42]. As our work involves a domain shift from a multimodal source domain to a unimodal target domain, we opt to utilize a non-generative approach for this work and focus exclusively on discriminative adversarial methods. It is assumed that a pre-existing distribution of multimodal data (i.e., interaction trace logs + facial expression) is available to train the source encoder and the source classifier, while the target distribution consists of unlabeled unimodal data (i.e., interaction trace logs). This is intended to simulate scenarios where visitor engagement models have been previously trained on multimodal data but are deployed in situations where only interaction trace log data is available to generate new predictions of visitor engagement.

While much prior work in adversarial domain adaptation involves source and target domains of similar or identical dimensionality (e.g., image-to-image translation), the multimodal aspect of this work presents a distinct challenge, as the multimodal data in the source domain inherently contains more features than the unimodal target domain. To enable the pre-trained multimodal classifier to handle unimodal data as input, stacked denoising autoencoders [45] are used to reduce the multimodal and unimodal feature vectors to the same dimensionality. An autoencoder is an unsupervised method of using feedforward neural networks to reduce an input vector X to a latent data representation using an encoder that contains a mapping function M. The autoencoder then attempts to use a decoder that uses mapping function N to reconstruct $M(X)$ to its original input. The encoder and decoder components of the autoencoder are both optimized by minimizing the reconstruction loss between X and $N(M(X))$. A stacked autoencoder is a variation in which each component contains multiple hidden layers of autoencoders. A denoising autoencoder builds on the same concept but corrupts the input vectors using random noise injection, which allows effective model regularization [45]. In this work, we use a corruption level of 0.25 on each feature in each input vector, where a value is set to 0 when the input feature is corrupted. After input vector X undergoes random noise injection to produce X', the denoising autoencoder attempts to reconstruct X from $N(M(X'))$. This allows the autoencoder to become more robust against random noise within the input features while also preventing the autoencoder from overfitting or simply learning the identity function. Following the optimization of the autoencoder, the decoder component is discarded while the encoder component is retained for dimensionality reduction within our data processing pipeline. A denoising autoencoder is shown in Figure 3.

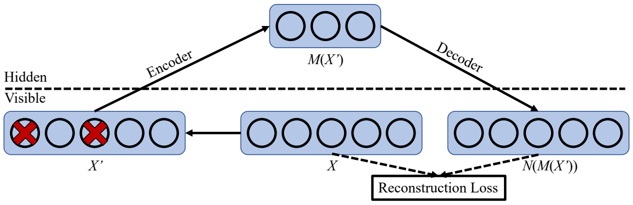


Figure 3. A denoising autoencoder.

Our adversarial domain adaptation process is shown in Figure 4. Figure 4A illustrates the initial training of the classifier and the source encoder. The features from the facial expression and interaction modalities are concatenated together and then used to train a stacked denoising autoencoder. Following this process, the trained source encoder is then used to reduce the multimodal input data to a latent representation that is then used to train a classifier. The classifier receives the latent data as input and is trained to predict the target variable, visitor dwell time. To enable the adversarial training of the target encoder and discriminator (Figure 4B), the weights of the pre-trained source encoder are fixed, and the target encoder weights are initialized using a pre-trained autoencoder optimized on the unlabeled, interaction-only data. An

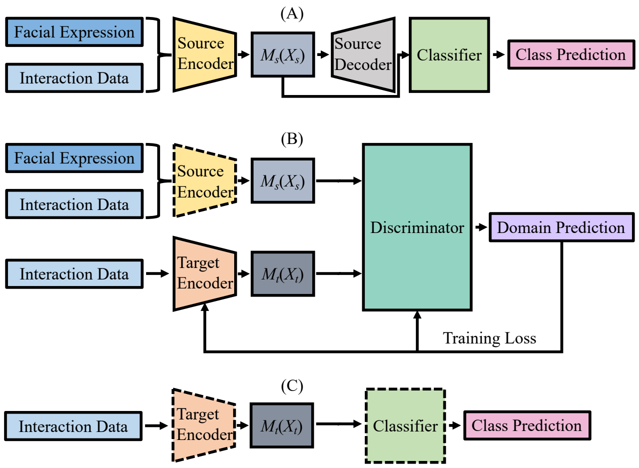


Figure 4. Domain adaptation process, including (A) the classifier and source encoder training, (B) adversarial training of the target encoder and discriminator, and (C) evaluation of the adapted target encoder on the classifier. Dashed lines indicate fixed model weights.

alternative approach is to initialize the target encoder weights from the source encoder. However, this can only be accomplished if the feature vectors extracted from the source domain are the same dimensions as the target domain. In our work, the multimodal feature vectors from the source domain have a higher dimensionality than the unimodal feature vectors from the target domain, since we remove the facial expression modality from the training data for the target encoder. The source and target encoders are used to produce latent representations of the multimodal and unimodal features, respectively. These representations are assigned labels of either 1 if the sample originated from the source domain, and 0 if the sample originated from the target domain. The data/label pairs are then used to train a feedforward network serving as the discriminator model. The discriminator is trained to distinguish between latent data from the source domain and from the target domain, while the target encoder is simultaneously trained to produce latent data from the target domain that consistently deceives the discriminator. To evaluate the target encoder (Figure 4C), unimodal data is passed through the encoder, and the resulting encoded data is then forward propagated through the trained classifier shown in Figure 4A. This procedure provides a way to evaluate the predictive performance of a multimodal classifier on unimodal data. It is important to note that some amount of multimodal data must be present prior to deploying our adversarial approach in order to train the multimodal classifier as well as the multimodal autoencoder.

6. METHODOLOGY

In multimodal models of learner engagement, some modalities that are highly predictive of engagement can also be impractical or undesirable in certain educational settings, such as sensors that require a cumbersome calibration process or expensive specialized equipment. Modalities that involve the capture of video data can raise concerns about privacy. However, eliminating physical sensors and exclusive reliance on sensor-free modalities may result in decreased performance on some tasks and settings. We propose a solution to this issue that (1) allows the predictive capacities of multimodal models to be retained, and (2) allows for the reduction in use of physical sensors. This work operates under the assumption that multimodal data is available in at least some capacity to facilitate the training of multimodal models prior to adversarial domain adaptation. As a result, the ideal setting for the proposed framework is after an initial multimodal data collection has taken place, enabling pre-trained multimodal models to be deployed. Below we describe the methods used to preprocess the multimodal and unimodal data, the feature selection process utilized to select the data used in the prediction and adversarial tasks, and the approach to training and validation of the visitor engagement models. Finally, we present the early prediction convergence metrics used to evaluate the final classification models and the domain encoders.

6.1 Data Preprocessing

6.1.1 Temporal Feature Engineering

To facilitate early prediction of visitor engagement, sequential representations were produced from the features engineered from the two modalities as described in Section 4.2. To accomplish this, feature vectors were engineered for every subsequent 10-second interval in a single visitor's interaction session with the exhibit. For each feature, the average or sum of all values from t=0 to t=10n seconds was calculated, where n is the number of 10-second intervals that have elapsed for that feature vector. For example, if a visitor engaged with the exhibit for one minute, then n=6, and the feature vectors are generated across time intervals of 10, 20, 30, 40, 50, and 60 seconds from the beginning of their session. This allows each feature vector to be a representation of a visitor's behavior over their entire interaction with an exhibit up to that point. Additionally, this approach solves the issue of the temporal alignment of the separate data channels caused by differing sampling rates of the facial expression modality and the interaction-based modality. As a result, the early prediction models are able to make predictions at a consistent frequency across every visitor's exhibit interaction trajectory (i.e., every 10 second). To ensure that the additive nature of the features does not contribute to artificially inflated model performance, each feature is scaled by the elapsed time up to the current timestamp. After this process is complete, 2,279 data samples were generated for 79 visitors.

6.1.2 Visitor Dwell Time

The beginning of a visitor's dwell time takes place after a calibration process with the eye gaze sensor is completed, and prior to when they are presented with an on-screen information dialogue box explaining the problem to be solved. The visitor's session can end one of three ways: (1) the visitor opts to end their session prior to completing the problem-solving task in Future Worlds, (2) the visitor solves the problem and chooses to end their session, or (3) the visitor solves the problem, opts to continue interacting with the virtual environment, and later chooses to end their session. Each visitor's dwell time was captured in total seconds (M=268.83, SD=137.48, Min=77.11, Max=657.48) and was recorded by the Future Worlds exhibit's built-in logging functionalities. For the purpose of this work, the dwell time prediction task was converted to a classification problem by splitting dwell time into three tertile groups and assigning approximately one-third of the visitors to each group. We use this classification approach instead of regression analysis due to the relatively low number of visitors in the dataset and to accommodate the use of early prediction convergence metrics. The visitors in the dataset were assigned to one of three possible groups according to their dwell time d: low (d <= 193.54, N=26), low (193.54 < d <= 318.82, N=27), and high (d > 318.82, N=26). We take this approach as a way to prevent a significant class imbalance while still retaining a higher level of granularity than a median split. The distribution of visitor dwell times, including the ternary groups, is shown in Figure 5.

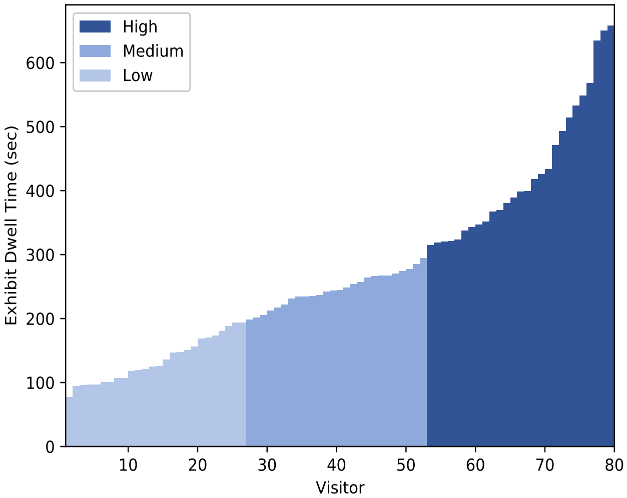


Figure 5. Distribution of visitor dwell times and ternary groups.

6.2 Feature Selection

Because of the large number of features in the multimodal data (51 facial expression features and 8 exhibit interaction features), we implemented forward feature selection to eliminate features with little or no predictive value and to reduce potential noise. Forward feature selection iterates through a list of features in a greedy manner, training a model on a single feature and continuing to add features if their inclusion increases the performance of the model on the target variable. This process continues until a predetermined number of features are selected or until all available features have been evaluated. This process has a few shortcomings. Due to the greedy nature of the algorithm, the features that are evaluated first have a higher chance of being selected. For example, the first feature that is evaluated is always retained, regardless of its true contribution to the predictive performance of the model. One approach to mitigating this issue is to perform forward feature selection for every possible combination of features, but this is often prohibitive as the number of combinations increases exponentially as the number of features increases, which imposes significant computational requirements. To mitigate the issue of bias in greedy feature selection while avoiding an exhaustive search across all feature combinations, we perform forward feature selection across a randomized ordering of all available features. We used a support vector machine (SVM) as the predictive model for each feature combination due to its effectiveness in high-dimensional spaces and relatively small computational overhead. This process was repeated for 100 separate iterations and randomizations to ensure that each feature had an equal probability of being placed at a specific point within each feature ordering. Following this process, the features were sorted according to the frequency that each feature was selected across all 100 iterations. To compensate for the difference in the number of features for each data channel, we performed forward feature selection on the facial expression modality and selected the ten most frequently selected features.

It should be noted that because we selected the ten most frequent features from the facial expression modality, and the interaction-based modality contained only 8 total features, each feature from the latter modality was included in the data modeling process. (We perform forward feature selection on the interaction-based features for analysis purposes only.) Because certain features such as AU durations and tile durations increase monotonically throughout a visitor's exhibit interaction trajectory and can lead to indirect data leakage with regard to the target variable (dwell time at the exhibit), the features were scaled by the total elapsed time up to the current timestamp, so these features were converted to proportional representations of the elapsed time at each time interval.

Table 1. Most frequent features from forward feature selection (interaction)

Feature	Frequency
Proportional Tile Duration	0.637
Proportional Open Tile Count	0.561
Proportional Info Duration	0.557
Proportional Info Taps	0.554
Proportional Taps	0.511
Proportional Swipe Tiles Count	0.416
Proportional Modify Tile Count	0.341
Beat Game	0.272

Table 2. Most frequent features from forward feature selection (facial expression)

Feature	Frequency
AU05 Max	0.317
AU10 Max	0.276
Proportional AU10 Duration	0.257
AU02 Max	0.237
Proportional AU01 Duration	0.227
AU26 Std	0.218
AU25 Max	0.214
Proportional AU17 Duration	0.208
Proportional AU45 Duration	0.206
Proportional AU26 Duration	0.196

This feature selection process took place within each cross-validation fold, and as a result, each fold produced a different combination of selected features. We calculated the frequency of the features across all cross-validation folds and present these in Table 1 and Table 2.

Based on the results in Table 1, features related to general interactions (proportional number of times any tile was opened, proportion of time any tile was open) were the most predictive interaction-based features. The features related to opening and viewing embedded graphical and textual science materials were also frequently selected features. The features representing the frequency a visitor modified the in-game virtual environment were less frequently selected as predictive features, as was the binary indicator of whether the visitor correctly solved the problem at that particular timestamp.

The most predictive features from the facial expression modality were primarily maximum values and proportional durations of certain AUs. AU05 (upper lid raiser) and AU10 (upper lip raiser) were the most predictive facial action units, followed by AU02 (outer brow raiser) and AU01 (inner brow raiser). AU25 (lips part) and AU26 (jaw drop) were moderately predictive, followed by AU17 (chin raise) and AU45 (blinking). Multiple representations of AU10 and AU26 were frequently selected during the feature selection process as well. It is notable that the overall frequency of the facial expression features is significantly lower than many interaction-based features. This is likely a result of the large number of facial expression features compared to the interaction-based features.

6.3 Model Evaluation

The models were evaluated using 10-fold cross-validation, with the splits for each fold occurring at the visitor level to ensure that a visitor's data was contained only within a single training, validation, or test set. The dataset was standardized within each cross-validation fold by dividing each feature by subtracting the feature's mean and dividing by the feature's standard deviation, as determined by the training data. This rescales the data to have a standard deviation of 1 (unit variance) while centering the mean to be 0. Following this process, class imbalances within the training data were resolved using Synthetic Minority Oversampling Technique (SMOTE) [10]. SMOTE is a common upsampling approach that resolves class imbalances through a randomized K-nearest neighbor approach, which brings the class balance to a uniform distribution while avoiding duplication of any data points. The upsampled, standardized training data is then used for forward feature selection as described in Section 6.2.

After feature selection, a classifier model was trained on the multimodal data and the visitor dwell time labels in each cross-validation fold to provide a comparison point for the domain-adapted models. The tertile labels for the target variable were encoded as one-hot vectors for each model output. We evaluated five different models: SVM, logistic regression, naïve Bayes, random forest, and a feedforward neural network. We performed hyperparameter tuning using a 3-fold nested cross-validation within the training set for each outer cross-validation fold. The hyperparameters that were varied for each model included the regularization parameter and kernel (SVM), regularization parameter (logistic regression), number of estimators (random forest), and number of layers and nodes (feedforward neural network). Additionally, the architecture of the autoencoder used to train the source encoder was evaluated during the hyperparameter tuning phase. The autoencoder was a feedforward neural network, and the hyperparameter values that were evaluated were the number of layers and nodes in the hidden layers within the encoder and decoder, as well as the number of latent dimensions. The feedforward neural network achieved the optimal performance as the classifier for visitor dwell time, using two hidden layers with 64 nodes each. The source encoder contained three hidden layers with 64, 32, and 16 nodes, respectively, with a latent output of 10. Additionally, all feedforward neural network models used a learning rate of 0.001, a dropout rate of 0.5 in the last hidden layer, and sigmoid activation functions. The loss function used for each model was categorical cross-entropy. Early stopping was implemented for each model using the validation data during the nested cross-validation to protect against overfitting. As a baseline, we follow the same process previously described, except using only the interaction modality. We evaluate both a unimodal and multimodal baseline in order to demonstrate the improved performance of the multimodal model of visitor dwell time as compared to the unimodal model, and to show the improved performance using the domain adaptation framework in situations where only unimodal data is available.

After the optimal classifier and source encoder for adversarial domain adaptation were trained for each cross-validation fold, the models' weights were fixed to evaluate the classifier performance on interaction-only data and to encode the multimodal data within the adversarial framework, respectively. The adversarial framework used a target encoder that is a feedforward neural network whose architecture and weights were pre-determined using the interaction-only baseline model. Although the source encoder and target encoder weights were not tied together as is common in other adversarial domain adaptation work [27], there was an imposed restriction that the latent dimensions be the same for both domains due to the fixed input size of the discriminator. The discriminator in the adversarial framework was a feedforward neural network with two layers of 64 nodes each. The learning rate of both the discriminator and the target encoder was 0.001, with a dropout rate of 0.05 in the last hidden layer and hyperbolic tangent activation functions. The loss functions for the discriminator and target encoder were based on binary cross-entropy as shown in Equations 1 and 2, respectively. The adversarial domain adaptation took place within each cross-validation fold to prevent data leakage from the test set.

To evaluate the predictive performance of the domain-adapted representations of the target data, the trained target encoder was used to encode the interaction-only data from the held-out test set within each cross-validation fold, and the encoded data was passed to the classifier model trained with the source data. The predictive performance of the classifier on this data was used to confirm that the use of multimodal data to train the classifier induces higher performance than if the facial expression data was removed from the dataset entirely. As an additional baseline, the target encoder trained on the interaction-only modality was used to pass the encoded data directly to the multimodal classifier without the domain adaptation procedure, following the source-only baseline approach of Tzeng et al. [42]. This illustrates that any improvement due to our method can be attributed to the adjusted weights through the adversarial adaptation process instead of just compressing the latent representation of the target domain data to the source domain's dimensionality. This specific baseline is called *target-only*.

6.4 Early Prediction

To quantify the models' ability to accurately predict a visitor's dwell time early and consistently, we utilize two metrics: *standardized convergence point* [30] and *convergence rate* [4]. The *standardized convergence point* calculates an average point of model convergence to the correct labels, while a particular visitor's sequence not converged to a correct prediction is penalized. This metric extends the conventional *convergence point* metric to account for sequences that are ultimately predicted incorrectly and fail to converge by instituting a penalty term [4]. In this instance, standardized convergence point is greater than one. In cases of convergence, a sequence's standardized convergence point falls within the range [0, 1]. Equation 3 displays the formula used to calculate the standardized convergence point across all sequences, where m is the number of sequences, and n_i is the number of data points in the i^{th} visitor's sequence. The value of k_i is the number of data points after which the model makes consistently accurate predictions, otherwise k_i equals n_i+p_i, where p_i is the penalty term for the i^{th} sequence [30]. (p_i is set to 1 for all sequences in this work following the original work.) A lower standardized convergence point indicates that the model's predictive accuracy tends to converge earlier in a visitor's interaction with the exhibit, indicating better early prediction performance.

$$Standardized\ convergence\ point = \sum_{i=1}^{m} \frac{\left(\frac{k_i}{n_i}\right)}{m} \qquad (3)$$

The second metric that we use to quantify a model's early prediction performance is the *convergence rate*. Convergence rate is the percentage of observed sequences in which the final prediction is accurate. Any sequence that contains an accurate dwell time prediction at the last data point is considered to have converged. Therefore, a higher convergence rate is indicative of better performance.

7. RESULTS AND DISCUSSION

The results for the unimodal and multimodal models as well as the unimodal latent representations (i.e., target-only encoding) and domain-adapted representations are shown in terms of early prediction and visitor-level predictive performance in Table 3. To measure visitor-level performance, a single point estimate of the predictive performance for each individual visitor is obtained by averaging across the predictions for all data points. The results for Table 3 are shown in terms of standardized convergence point (SCP) and convergence rate (CR) for early prediction, and area under curve (AUC), Cohen's Kappa, accuracy, and F1 score for visitor-level performance. Although AUC is commonly used for binary classification problems, we use this metric for a multi-class approach using a "one vs. rest" method which treats the correct class as the "positive" group and combines all other classes as a single "negative" group. The total AUC for a single model is calculated by using the unweighted mean of the AUC values across all three groups.

Based on the results in Table 3, the adversarial domain adaptation allows the multimodal classifier to outperform all baselines in terms of early prediction and across all sequences for each visitor. As expected, the complete multimodal model achieved the highest performance, achieving an AUC value of 0.660, while also outperforming the other models in all other evaluation metrics. The model achieved a standardized convergence point of 64.58%, indicating that the model achieved and maintained its optimal predictive performance approximately 64% into a visitor's total dwell time at the exhibit, while converging to the correct predictions more often than other baseline approaches. The interaction-only modality produced noticeably lower performance, achieving a convergence point of 75.95%, while also reaching a 0.574 AUC across all sequences. The adversarial domain adaptation allowed the classifier to achieve higher performance on the interaction-only data, with an early prediction performance of 67.42% and a visitor-level AUC of 0.585, similar to the full multimodal model while also outperforming the interaction-only baseline across all evaluation metrics.

The classifier's performance on the latent unimodal data (without domain adaptation) was notably poor, achieving an AUC that was slightly worse than random chance (0.500). This result is not surprising, as we are evaluating the model's performance using latent representations from a domain that has not been used to train the model beforehand. Although similar baseline approaches can achieve moderate performance in instances where the source and target domains are relatively similar, other work that investigates cross-modality adaptation or adaptation across dissimilar domains achieves much lower performance for this specific baseline [42].

While the adversarial domain adaptation proved more effective than the interaction-only and latent unimodal data baselines, the performance of our framework did not achieve the same performance as a framework that contained the full multimodal data. This could be attributed to the significant difference between the interaction and facial expression domains. The majority of the interaction-based modality is comprised of discrete, monotonically increasing features, which inherently are not as data-rich as the features from the facial expression modality. Because there are multiple features for each AU, this modality provides multifaceted perspectives on multiple AUs, leading to a relatively high number of continuous features. Adapting between two data channels with such a discrepancy in dimensionality may be a contributing factor to the framework's performance. Second, the relatively small number of visitors in the dataset may also be a contributing factor, as the performance of the models could be at risk for overfitting the classifier, source encoder, or target encoder. Contributing to this potential issue is the loss induced in the domain adaptation process. The size of the dataset may prevent the adversarial framework from reaching optimal convergence. Third, because there is no restriction regarding how long the visitors could remain at the exhibit, the target variable has a relatively wide range of values, approximately from one minute to more than ten minutes. Although this issue is addressed through the use of a tertile split, additional data could provide further evidence of behavioral patterns that are able to induce higher performance with more granular target variables.

Because timestamped interaction trace logs are the basis of one of the modalities used in this work, the design of the museum exhibit may play a role in the performance of the visitor models in terms of early prediction. During the early stages of Future Worlds, visitors are prompted to read an information dialog box explaining the premise of the game and a summary of the problem to be solved in the virtual environment. Because this event occurs at the beginning of every visitor's interaction sequence, it is likely that more indicative behaviors that allow the classifier to differentiate between groups occur at later stages of learner interactions with the exhibit. This is a potential explanation behind the early prediction performance of each model, as the standardized convergence point occurs after 60% of the overall exhibit interactions across all models.

To further investigate the impact that domain adaptation has on the predictive performance of the multimodal classifier, confusion matrices based on the target-only encoder and the adversarially-trained encoder are shown in Figure 6 as is the confusion matrix for the interaction-only classifier. The purpose of this analysis is to determine if adversarial domain adaptation results in any changes relative to the classifier's sensitivity to certain dwell time groups.

Table 3. Visitor-level predictive performance (all sequences)

	Early Prediction			Visitor-Level Prediction			
Encoding	Classifier	SCP	CR	AUC	Kappa	Accuracy	F1 Score
Interaction-Only	Unimodal	75.95%	34.18%	0.574	0.085	0.392	0.355
Multimodal	Multimodal	**64.58%**	**48.10%**	**0.660**	**0.278**	**0.519**	**0.511**
Target-Only	Multimodal	73.79%	34.18%	0.499	0.015	0.342	0.338
Domain Adaptation	Multimodal	**67.42%**	**43.04%**	**0.585**	**0.203**	**0.468**	**0.468**

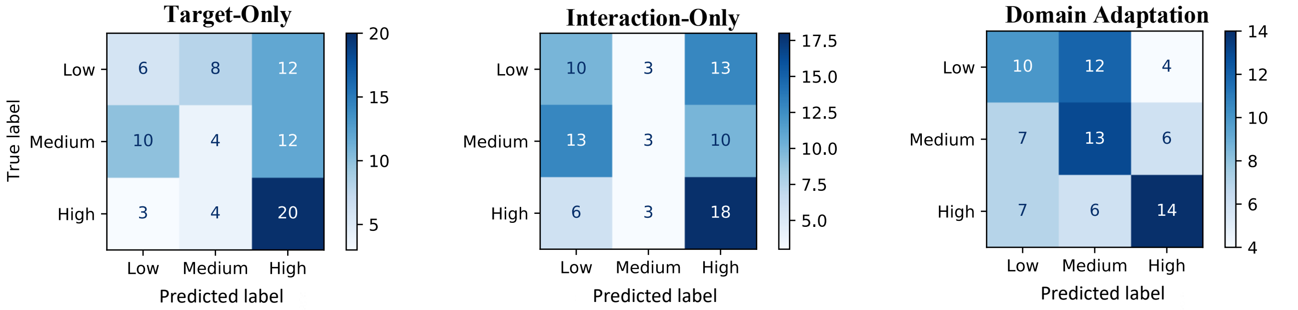


Figure 6. Confusion matrices for classifiers using target-only, interaction-only, and domain adaptation-based representations.

Based on the confusion matrix for the target-only classifier (i.e., the multimodal classifier evaluated on the interaction data without domain adaptation), the classifier appears to primarily predict high dwell times for a majority of visitors. The model also appears to frequently predict visitors with medium dwell time as having low dwell time. As this particular model performed similarly to a random chance classifier, it is likely that the interaction-only data representation was not easily identifiable to the classifier, leading it to primarily predict a single class and not classify the lower two groups accurately. The classifier that was trained and evaluated on interaction-only encodings performed slightly better and appears to become more accurate in cases of lower dwell time in visitors. However, it is notable that the model still does not appear to accurately predict instances of medium dwell time. This indicates that the interaction-based modality contains salient features indicative of noticeably low or high engagement but interactions from visitors with medium dwell time are not easily distinguishable to the model. Low dwell time may be characterized by a relatively low number of taps or interactions in the virtual environment, while high dwell time may be indicated by greater or more frequent tapping or interactions with the virtual environment. Additionally, visitors that have a higher dwell time are more likely to beat the game or read a higher number of information dialogs. However, this information may not be predictive enough with the ternary split, causing the interaction model to overfit to the two extremes.

The multimodal classifier that processes the modality-invariant data representations performs noticeably better for visitors with medium dwell time and continues to maintain fairly accurate performance on visitors with high dwell time. This may indicate that facial expression captures physical cues that allow the model to more easily distinguish between the medium group and the other groups, and the domain adaptation allows these features to be integrated into the interaction-only representations. By implementing this approach across the two modalities, it appears that the multimodal model retains its robustness to visitors with a medium dwell time in particular, while being able to achieve this performance using only features from the interaction data. This is significant because it appears that the interaction-only model does not appear to induce high performance on the medium dwell time visitors, so it remains important to utilize the multimodal data representations obtained through domain adaptation as pre-training for accurately predicting the visitor dwell time.

8. CONCLUSION

Modeling visitor engagement is an important task in museum-based learning. However, visitor engagement modeling presents significant challenges, as visitors' patterns of engagement with museum exhibits can vary widely. Multimodal frameworks show promise for the prediction of visitor engagement in museums because they capture information about visitor behavior that cannot otherwise be captured through interaction trace logs or similar unimodal data channels. Although multimodal sensor systems give rise to concerns about privacy, feasibility, and intrusiveness, the complete removal of sensor data from visitor engagement models may result in diminished predictive performance. To address this issue, we have introduced an adversarial domain adaptation approach to generating modality-invariant representations of interaction data and facial expression data from visitor interactions with the FUTURE WORLDS museum exhibit. The domain adaptation approach enables multimodal models to be induced in a pre-training phase while being deployed and evaluated with modality-invariant representations obtained using interaction-based data exclusively. We investigate the models' ability to predict visitor dwell time during the early stages of a visitor's interaction with the museum exhibit. Results indicate that the domain adaptation approach to modeling visitor engagement achieves higher performance than a visitor modeling approach using only a single modality. The domain adaptation approach also outperforms the unimodal baseline during early sequences of a visitor's interaction trajectory as well as across all sequences while demonstrating competitive performance compared to classifiers utilizing multimodal data.

There are several promising directions for future work. Alternative techniques for modeling visitor engagement should be evaluated, including sequential models like long short-term memory (LSTM) networks, to improve models' predictive accuracy and early prediction. Alternative approaches to the adversarial learning component of this framework include the use of generative models such as GANs or variational autoencoders. Attaining reliable training convergence continues to be a challenging problem within adversarial learning and investigating solutions to this issue may enhance the benefits of domain adaptation. The generalizability of the domain adaptation framework should be evaluated using larger and more diverse visitor populations on different exhibits and museum settings. Additionally, the domain adaptation framework should be evaluated using additional combinations of modalities (e.g., posture, gaze, speech), and extended to include three or more modalities simultaneously. Finally, this framework should be evaluated at run-time by integrating visitor engagement models into a museum exhibit to enable visitor-adaptive interventions to enrich visitor engagement and enhance museum-based learning experiences.

9. 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.

10. REFERENCES

[1] Aslan, S., Alyuz, N., Tanriover, C., Mete, S., Okur, E., D'Mello, S. and Arslan Esme, A. 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.

[2] Baltrušaitis, T., Ahuja, C. and Morency, L. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 41, 2 (Feb. 2019), 423–443.

[3] Baltrusaitis, T., Zadeh, A., Lim, Y.C. and Morency, L. 2018. OpenFace 2.0: Facial behavior analysis toolkit. In *Proceedings of the 13th IEEE International Conference on Automatic Face Gesture Recognition.* 59–66.

[4] Blaylock, N. and Allen, J. 2003. Corpus-based, statistical goal recognition. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*. 1303–1308.

[5] Block, F., Hammerman, J., Horn, M., Spiegel, A., Christiansen, J., Phillips, B., Diamond, J., Evans, E.M. and Shen, C. 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.

[6] Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J. and Shute, V. 2015. Temporal generalizability of face-based affect detection in noisy classroom environments. In *Proceedings of the International Conference on Artificial Intelligence in Education*. 44–53.

[7] Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M. and Zhao, W. 2016. Detecting student emotions in computer-enabled classrooms. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 4125–4129.

[8] Chan, M., Ochoa, X. and Clarke, D. 2020. *Multimodal Learning Analytics in a Laboratory Classroom*. Springer International Publishing.

[9] Chang, C., Zhang, C., Chen, L. and Liu, Y. 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.

[10] Chawla, N., Bowyer, K., Hall, L. and Kegelmeyer, W. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. 16, 321–357.

[11] Chung, J., Cannady, M., Schunn, C., Dorph, R. and Bathgate, M. 2016. Measures Technical Brief: Fascination in Science.

[12] Diamond, J., Horn, M. and Uttal, D. 2016. *Practical Evaluation Guide: Tools for Museums and Other Informal Educational Settings*. Rowman & Littlefield.

[13] Dim, E. and Kuflik, T. 2014. Automatic detection of social behavior of museum visitor pairs. *ACM Transactions on Interactive Intelligent Systems*. 4, 4 (Nov. 2014), 17:1-17:30.

[14] D'Mello, S. and Kory, J. 2012. Consistent but modest: A meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*. 31–38.

[15] D'Mello, S. and Kory, J. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys*. 47, 3, 43:1-43:36.

[16] Emerson, A., Henderson, N., Rowe, J., Min, W., Lee, S., Minogue, J. and Lester, J. 2020. Early prediction of visitor engagement in science museums with multimodal learning analytics. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 107–116.

[17] Ganin, Y. and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the International Conference on Machine Learning.* 1180–1189.

[18] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. 2014. Generative adversarial networks. *arXiv:1406.2661*. (Jun. 2014).

[19] Grafsgaard, J., Wiggins, J., Vail, A., Boyer, K., Wiebe, E. and Lester, J. 2014. The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. In *Proceedings of the 16th International Conference on Multimodal Interaction*. 42–49.

[20] Harrington, M. 2020. Connecting user experience to learning in an evaluation of an immersive, interactive, multimodal augmented reality virtual diorama in a natural history museum & the importance of the story. In *Proceedings of the 6th International Conference of the Immersive Learning Research Network.* 70–78.

[21] Hein, G. 2009. Learning science in informal environments: People, places, and pursuits. *Museums & Social Issues*. 4, 1 (2009), 113–124.

[22] Henderson, N., Rowe, J., Paquette, L., Baker, R. and Lester, J. 2020. Improving affect detection in game-based learning with multimodal data fusion. In *Proceedings of the International Conference on Artificial Intelligence in Education*. 228–239.

[23] Knutson, K., Lyon, M., Crowley, K. and Giarratani, L. 2016. Flexible interventions to increase family engagement at natural history museum dioramas. *Curator: The Museum Journal*. 59, 4 (2016), 339–352.

[24] Kuflik, T., Boger, Z. and Zancanaro, M. 2012. Analysis and prediction of museum visitors' behavioral pattern types. In *Ubiquitous Display Environments*. A. Krüger and T. Kuflik, Eds. Springer. 161–176.

[25] Lane, H., Noren, D., Auerbach, D., Birch, M. and Swartout, W. 2011. Intelligent tutoring goes to the museum in the big city: A pedagogical agent for informal science education. In *Proceedings of the International Conference on Artificial Intelligence in Education*. 155–162.

[26] Li, H., Ding, W., Yang, S. and Liu, Z. 2020. Identifying at-risk K-12 Students in multimodal online environments: A machine learning approach. In *Proceedings of the 13th International Conference on Educational Data Mining*. 137–147.

[27] Liu, M.-Y. and Tuzel, O. 2016. Coupled generative adversarial networks. *arXiv:1606.07536*. (Sep. 2016).

[28] Long, D., McKlin, T., Weisling, A., Martin, W., Guthrie, H. and Magerko, B. 2019. Trajectories of physical engagement and expression in a co-creative museum installation. In *Proceedings of the 2019 Conference on Creativity and Cognition*. 246–257.

[29] Mao, Y., Zhi, R., Khoshnevisan, F., Price, T., Barnes, T., and Chi, M. 2019. One minute is enough: Early prediction of student success and event-level difficulty during novice programming tasks. In *Proceedings of the 12th International Conference on Educational Data Mining.* 119-128.

[30] Min, W., Baikadi, A., Mott, B., Rowe, J., Liu, B., Ha, E.Y. and Lester, J. 2016. A generalized multidimensional evaluation framework for player goal recognition. In *Proceedings of the 12th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment.* 197-203.

[31] Min, W., Mott, B., Rowe, J., Taylor, R., Wiebe, E., Boyer, K. and Lester, J. 2017. Multimodal goal recognition in open-world digital games. In *Proceedings of the AAAI Conference*

on Artificial Intelligence and Interactive Digital Entertainment. 13, 1 (Sep. 2017).

[32] Müller, P.M., Amin, S., Verma, P., Andriluka, M. and Bulling, A. 2015. Emotion recognition from embedded bodily expressions and speech during dyadic interactions. In *Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction.* 663–669.

[33] Munro, J. and Damen, D. 2020. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 122–132.

[34] Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R. and Kim, K. 2018. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 4500–4509.

[35] Phandi, P., Chai, K. and Ng, H. 2015. Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 431–439.

[36] Qi, F., Yang, X. and Xu, C. 2018. A unified framework for multimodal domain adaptation. In *Proceedings of the 26th ACM International Conference on Multimedia*. 429–437.

[37] Rowe, J., Lobene, E., Mott, B. and Lester, J. 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.

[38] Sawyer, R., Smith, A., Rowe, J., Azevedo, R. and Lester, J. 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.

[39] Sun, B. and Saenko, K. 2016. Deep CORAL: Correlation alignment for deep domain adaptation. In *Proceedings of the ICCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision.* 443–450.

[40] Tiam-Lee, T.J. and Sumi, K. 2018. Adaptive feedback based on student emotion in a system for programming practice. In *Proceedings of the International Conference on Intelligent Tutoring Systems*. 243–255.

[41] Tzeng, E., Hoffman, J., Darrell, T. and Saenko, K. 2015. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision.* 4068–4076.

[42] Tzeng, E., Hoffman, J., Saenko, K. and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 7167–7176.

[43] Tzeng, E., Hoffman, J., Zhang, N., Saenko, K. and Darrell, T. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv:1412.3474*. (Dec. 2014).

[44] Vail, A., Grafsgaard, J., Boyer, K., Wiebe, E. and Lester, J. 2016. Predicting learning from student affective response to tutor questions. In *Proceedings of the International Conference on Intelligent Tutoring Systems*. 154–164.

[45] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. and Manzagol, P. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research.* 11, 12 (December 2010), 3371-3408.

[46] Whitehill, J., Serpell, Z., Lin, Y., Foster, A. and Movellan, J. 2014. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing*. 5, 1 (Jan. 2014), 86–98.

[47] Wiggins, J., Kulkarni, M., Min, W., Mott, B., Boyer, K., Wiebe, E. and Lester, J. 2018. Affect-based early prediction of player mental demand and engagement for educational games. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. 243-249.

[48] Wu, S., Du, Z., Li, W., Huang, D. and Wang, Y. 2019. Continuous emotion recognition in videos by fusing facial expression, head pose and eye gaze. In *Proceedings of the International Conference on Multimodal Interaction.* 40–48.

[49] Yang, J., Wang, K., Peng, X. and Qiao, Y. 2018. Deep recurrent multi-instance learning with spatio-temporal features for engagement intensity prediction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 594–598.

[50] Zeng, Z., Chaturvedi, S., Bhat, S. and Roth, D. 2019. DiAd: Domain adaptation for learning at scale. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. 185–194.