What's Fair is Fair: Detecting and Mitigating Encoded Bias in Multimodal Models of Museum Visitor Attention

Halim Acosta
Department of Computer Science, North Carolina State University, Raleigh NC, USA, hacosta@ncsu.edu

Nathan Henderson
Department of Computer Science, North Carolina State University, Raleigh NC, USA, nlhender@ncsu.edu

Jonathan Rowe
Department of Computer Science, North Carolina State University, Raleigh NC, USA, jprowe@ncsu.edu

Wookhee Min
Department of Computer Science, North Carolina State University, Raleigh NC, USA, wmin@ncsu.edu

James Minogue
Department of Teacher Education and Learning Sciences, North Carolina State University, Raleigh NC, USA, james_minogue@ncsu.edu

James Lester
Department of Computer Science, North Carolina State University, Raleigh NC, USA, lester@ncsu.edu

Recent years have seen growing interest in modeling visitor engagement in museums with multimodal learning analytics. In parallel, there has also been growing concern about issues of fairness and encoded bias in machine learning models. In this paper, we investigate bias detection and mitigation techniques to address issues of algorithmic fairness in multimodal models of museum visitor visual attention. We employ slicing analysis using the Absolute Between-ROC Area (ABROCA) statistic to detect encoded bias present in multimodal models of visitor visual attention trained with facial expression and posture data from visitor interactions with a game-based museum exhibit about environmental sustainability. We investigate instances of gender bias that arise between different combinations of modalities across several machine learning techniques. We also measure the effectiveness of two different debiasing strategies—learned fair representations and reweighing—when applied to the trained multimodal visitor attention models. Results indicate that patterns of bias can arise across different modality combinations for the different visitor visual attention models, and there is often an inherent tradeoff between predictive accuracy and ABROCA. Analyses suggest that debiasing strategies tend to be more effective on multimodal models of visitor visual attention than their unimodal counterparts

CCS CONCEPTS • Applied computing ~ Education • Computing methodologies ~ Machine Learning

Additional Keywords and Phrases: Multimodal learning analytics, algorithmic fairness, museum-based learning, visitor modeling

1 Introduction

Measuring visitor engagement in informal learning environments, such as museums and science centers, poses significant challenges. Visitor engagement is a core component of learning in museums [7]. Engagement influences how visitors interact with museum exhibits, form understanding, and follow up to learn more after museum visits [7]. Recent advances in multimodal learning analytics have been used to detect patterns of learner engagement by utilizing multiple sensor-based data streams such as facial expression, posture, eye gaze, and interaction log data [3, 14, 28, 39]. Although the benefits of incorporating multimodal data streams into models of learner engagement has been demonstrated in both classroom and laboratory settings [13, 27, 28, 31], investigating multimodal analytics in informal settings is still in its early stages [17].

Recent years have seen significant advances in utilizing multimodal machine learning for a wide range of tasks [3, 5]. In parallel, questions about algorithmic fairness for machine learning models have also been a topic of growing concern [1, 12, 19, 18, 21, 26, 34]. Conceptualizations of algorithmic fairness often range from individual and group fairness to quantitatively measurable "distance metrics" that can be minimized through a fairness optimization process [10]. Statistical formulations of fairness, such as equalized odds, demographic parity, and equal opportunity, also come into play [8, 4, 19]. In this work, we conceptualize algorithmic fairness in terms of encoded bias, which emphasizes the potential risk of machine learning models that have differential impact on different groups of individuals. Considerations of algorithmic fairness in machine learning-based models of museum visitor engagement are important because museums often have missions to serve learners from a broad range of socio-cultural backgrounds. As machine learning techniques are utilized to model visitor engagement in museums, there is a risk of inheriting implicit biases as well. Museums that utilize biased machine learning models to measure visitor engagement may unwittingly tailor their exhibits in favor of these biases, leading to different qualities of visitor experience across different populations. This process can further reinforce encoded biases during future data collection and model refinement phases, causing further downstream effects [9].

A central component of visitor engagement is visual attention. In this paper, we investigate automated detection and mitigation of encoded bias in multimodal models of visitor visual attention with an interactive science museum exhibit. We utilize data that was captured from visitor interactions with a museum exhibit about environmental sustainability, *Future Worlds*. We examine four standard machine learning methods (random forest, support vector machine, Naïve Bayes, decision tree) for predicting levels of visitor visual attention using posture and facial expression data. We investigate the predictive accuracy of multimodal variations of each model compared to unimodal baselines. We also measure encoded bias present within each model by performing slicing analysis across gender groups using the Absolute Between-ROC Area (ABROCA) statistic [19]. The ABROCA statistic measures the amount of bias present by looking for differential behavior in model performance between two sub-populations in the data. In combination with a modality-level ablation study, slicing analysis aids in identifying sources of bias that may necessitate corrective measures. Finally, we examine the impact of two debiasing techniques, reweighing (RW) [22] and learned fair representations (LFR) [43], on the accuracy and fairness of multimodal machine learning models of museum visitor attention.

2 Visitor Modeling in Museums

We draw on and extend the literatures on visitor modeling in museums, multimodal learning analytics, and algorithmic fairness. We discuss each of these in turn.

2.1 Museum Visitor Engagement

Prior work has explored modeling learner engagement in formal settings, such as schools and universities [11, 13, 14, 15, 42]. While formal and informal learning environments share many core objectives, museum-based learning presents distinctive challenges for measuring learner engagement, including short dwell times and an emphasis on free-choice learning. Well-designed, engaging exhibits frequently have very brief visitor interactions, and even

highly engaged visitors can have short dwell times [17]. Furthermore, museums often attract a diverse range of visitors in terms of age, gender, and socio-cultural background. A promising approach for measuring visitor engagement is to instrument an exhibit with physical sensors, including webcams, depth cameras, eye trackers, and microphones dependent upon the context, to capture rich multimodal data that can be modeled using multimodal learning analytic techniques.

2.2 Multimodal Learning Analytics

Multimodal learning analytics has been the subject of increasing attention in recent years and has shown significant promise for modeling learning and engagement across a range of educational contexts [2, 27, 30, 31, 35, 36]. For example, Sümer et al. examined learner engagement using pose estimation and facial expression data in school classrooms [35]. The authors generated feature representations from student interactions with their neighbors using the pose and motion data. The multimodal features were used to train deep learning models for creating separate feature embeddings for affect detection and attention detection, respectively. Sawyer et al. showed that student models enhanced with facial action unit data outperformed baseline unimodal models as well as models trained only on composite emotions for predicting student engagement [31]. Other studies have found that decision-level fusion with data from multiple modalities, including temporal posture information extracted from a Microsoft Kinect sensor, yields increased predictive performance over unimodal models for affect detection [28]. Eye gaze has also been found to be a useful modality for measuring engagement in classrooms [27, 30].

2.3 Algorithmic Fairness

To date, there has been limited work investigating algorithmic fairness in multimodal learning analytics [24, 25]. Solutions to algorithmic fairness often focus on statistical notions of fairness. Barrio et al. provide a review of common conceptualizations of fairness and other fair learning techniques [8]. The authors detail the mathematical frameworks underlying these definitions and propose a probabilistic framework to compare definitions of fairness with statistical independence. Mehrabi et al. present an extensive survey of types of biases, and they also evaluate a series of bias mitigation strategies [26]. Other recent work has focused on bridging the gap between statistical notions of fairness and individual fairness. Rich subgroup fairness, a formulation of fairness proposed by [23] and closely related to [20], extends prior fairness metrics to include the constraint that it must hold for all possible subgroups of the data. An application of fast subset scanning, which is an anomaly detection algorithm, has been applied to detect bias across all subgroups of the data in a black box fashion [44]. Slicing analysis using the ABROCA statistic has been proposed as a generalizable method for detecting bias over various thresholds, overcoming the limitations of many statistical definitions of fairness [19]. In this paper, we extend this line of investigation by using slicing analysis to examine the effectiveness of different de-biasing strategies (learning fair representation and reweighing) in multimodal machine learning-based models of visitor visual attention in science museums.

3 *Future Worlds* Testbed Exhibit

Future Worlds is a prototype game-based museum exhibit about environmental sustainability [17]. In *Future Worlds*, visitors interact with a touch-based display to learn about environmental sustainability by investigating the impacts of alternative decisions made within a 3D virtual environment. *Future Worlds* offers learners the ability to touch, swipe, and tap the screen while they improve aspects of the virtual environment. The exhibit's science content centers on themes of water, energy, and food. The primary objective of *Future Worlds* is to enable visitors to learn about sustainability in an engaging way by exploring alternative modifications to a simulated environment with the goal of improving its sustainability. Visitors' interactions enable them to test hypotheses about different environmental decisions by exploring cause-and-effect relationships between different components of the simulated environment and examining informational dialogs that impart knowledge about environmental elements (e.g., forests, rivers, solar power, industrial farms, etc.). The science content in *Future Worlds* is designed for learners ages 10-11. Prior studies have shown that learner interactions with *Future Worlds* yield enhanced

knowledge about environmental sustainability concepts and promising levels of observed visitor engagement [38]. Figure 1 shows an example of a visitor interacting with the *Future Worlds* game-based exhibit.

Figure 1: *Future Worlds* interactive museum exhibit on sustainability.

4 Multimodal Dataset

To investigate bias detection and mitigation in multimodal models of museum visitor attention, we utilized a multimodal dataset capturing visitor interactions with *Future Worlds* in a science museum. Learners from three different schools participated in the study. Each school served a student population in which over 70% of students came from economically disadvantaged homes. In total, 116 visitors between the ages of 10-11 participated. Each learner completed both pre- and post-surveys that captured visitor information such as demographics, science interest, sustainability content knowledge, and engagement. The participant sample was 32.4% Hispanic or Latino, 21.6% Black or African American, 11.8% Native American, 8% Asian, 3% Caucasian, and 7.5% mixed races. The remaining 15.7% of students indicated that they preferred not to respond. The gender makeup of the sample included 47 females and 55 males while the remaining students did not provide gender information. Data from 65 students were used to train the multimodal visitor attention models after removing students that were missing either survey data or multimodal data.

Prior to engaging with *Future Worlds*, visitors were introduced to the exhibit and the physical sensors were calibrated for each visitor. Visitors interacted with *Future Worlds* for a maximum of 10 minutes (M = 3.97, SD = 2.24). During the study, the *Future Worlds* exhibit was instrumented with physical sensors and logging software to capture real-time posture, facial expression, eye gaze, and interaction trace log data. Features extracted from these data channels were utilized to develop multimodal machine learning-based models of visitor visual attention.

Facial expression data was captured from an externally mounted Logitech C920 USB webcam, and the video recordings were analyzed in real-time using the OpenFace facial behavioral analysis toolkit [4]. Facial expression data has been widely used for modeling and predicting engagement [6, 27, 32, 33, 35, 37, 40]. OpenFace allows for the automated detection of 17 distinct facial action units (AU) for each face captured in the webcam's field of view as well as head pose and eye gaze estimation.

Using the Microsoft Kinect V2 for Windows, visitors' postural movements were tracked for 26 distinct vertices in 3D coordinate space along with RGB and depth channel representations. Posture has also been shown to be predictive of different affective states through bodily pattern mining [13]. Analyzing a learner's body position and movement has also been shown to be effective when combined with emotion templates in emotion recognition

tasks [28]. The Kinect sensor was positioned approximately five feet away from each visitor in a front-facing arrangement.

Timestamped records of visitors' interactions with the exhibit's multi-touch interface were captured through interaction trace log data. These interactions were recorded at the millisecond granularity and captured actions such as requesting more information about a particular concept or modifying the in-game virtual environment. A benefit of using an interaction-based modality to model visitor attention is that such sensor-free modalities are more robust against issues that frequently impact sensor-based modalities such as noise, mis-tracking, miscalibration, and hardware failure.

Visitors' eye gaze patterns were captured by employing an externally mounted Tobii EyeX eye tracking sensor. Prior work has shown that eye gaze can be used effectively in measuring engagement [27]. Ray casting techniques were utilized to automatically identify in-game targets of visitor's attention. This process yielded information including timestamps, eye gaze targets, and durations of visitors' fixations on regions of the exhibit's interface.

4.1 Multimodal Features

Several predictive features were extracted from the collected data based on prior work using multimodal learning analytics to predict museum visitor engagement [17]. Eight features were distilled from the interaction trace log data, including the total number of times that a visitor tapped on the interactive display and the total number of times a visitor tapped to examine informational texts about several in-game elements. Additional features distilled from the trace logs included whether a visitor solved the game's problem scenario and the total number of times a visitor interacted with the exhibit's interface through actions such as opening and swiping through different dialogs and modifying the virtual environment.

Facial expression features were extracted from the facial action unit data captured by the OpenFace software. To calculate the duration of each AU's presence during a visitor's interactions with the exhibit, the intensity of each AU was standardized and subsequently recorded only if the present intensity exceeded one standard deviation above the mean intensity. Each AU's calculated duration only includes intervals when the recorded intensity was prolonged for more than a half-second to avoid capturing noise due to micro expressions. The duration was calculated for 18 AUs, and 36 additional features were generated using the standard deviation and maximum values for each AU's intensity.

Posture-based features were extracted from four skeletal vertices tracked by the Microsoft Kinect: *head*, *upper back*, *mid back*, and *neck*. The minimum, maximum, median, and variance of the (x, y, z) coordinates for each vertex were used as features. Two additional posture-based features were also extracted. Total postural change was calculated from the summative change across all vertex coordinates. Total change in terms of the Euclidean distance from the Kinect sensor was calculated across all vertices as well. In total, 18 posture-based features were extracted from the raw Kinect data.

Eye tracking data captured from each visitor was mapped to predetermined areas of interest (AOIs) to quantify the duration a visitor was fixated on a particular region of the exhibit's interface. Informed by prior research, gaze fixations longer than 210 milliseconds were included as part of an AOI's total gaze duration. Gaze durations were calculated for four distinct AOI categories: virtual location (AOI-Location), environmental sustainability selection menus (AOI-Menu), environmental sustainability textual information and imagery (AOI-Information), and the navigational interface (AOI-Interface). This categorization was determined based upon the gaze targets' functional role in the game (e.g., imparting information about science content, navigating the interface, enacting a change to the simulated environment, etc.) AOI-Location refers to nine distinct regions within the virtual environment. AOI-Menu encompasses the in-game elements a visitor interacts with while modifying the virtual environment or querying a specific element for more information. AOI-Information represents the interface within *FUTURE WORLDS* that presents science content to the visitor, such as informational text or imagery pertaining to a particular aspect of the virtual environment (Figure 2). Finally, AOI-Interface represents elements within the navigational interface such as the arrows and buttons used to begin, pause, or exit the game. It should be noted that the four AOIs are disjoint groups: a visitor's eye gaze can only fall into a single AOI at a time, as AOIs do not overlap in this work.

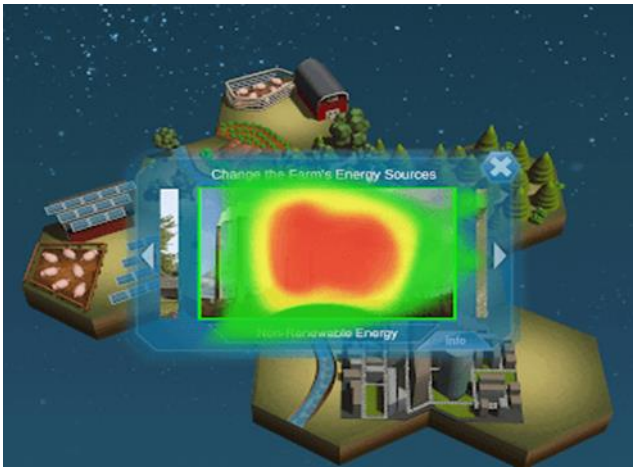

Figure 2: Visualization of visitor attention captured by AOI-Information.

To serve as a measure of visitor attention toward the exhibit's science content, we utilized the proportional time that visitors spent visually fixating on AOI-Information. Specifically, we calculated AOI-Information's proportional gaze fixation time by dividing the AOI's total duration by the total gameplay time for each visitor. The proportional fixation time of the AOI was categorized into "high" and "low" groups based on a median split (median = .031) across all visitors. This feature served as the target variable for our visitor attention models. This feature was chosen based upon the premise that the greater amount of time spent visually fixating on AOI-Information, the more visual attention was being dedicated toward the exhibit's science content. Attentional management has been shown to be correlated with high levels of engagement in classroom settings [11, 29, 41].

5 Method

We investigate encoded bias in multimodal models of visitor visual attention using four machine learning techniques: random forest (RF), support vector machine (SVM), Gaussian Naïve Bayes (NB), and decision trees (DT). We evaluate the predictive performance of these models for identifying visitors who spend a high amount of time fixating on science-related informational dialogs in the *FUTURE WORLDS* exhibit. We then performed a slicing analysis, in which we evaluated the predictive performance of the models on different "slices" of the data—across gender lines in our case—to detect the presence of bias using the ABROCA statistic. Finally, we investigate the use of the ABROCA statistic to measure the effectiveness of two debiasing techniques, learned fair representations and reweighing using AI Fairness 360 (AIF360). AIF360 is an extensible toolkit for detecting and mitigating bias in machine learning models [9]. The toolkit supports a broad range of debiasing strategies in the preprocessing, in-processing, and post-processing stages of the model training pipeline.

5.1 Visitor Visual Attention

During preprocessing, we removed the interaction-based modality because it was linearly related with the target variable. Interaction data included information such as whether a visitor completed the game, the total number of gameplay interactions within the environment, the duration of a visitor's interaction as well as any variable associated with the four main AOIs. Each visitor attention model was evaluated using nested cross-validation. Hyperparameter tuning was performed using 3-fold inner cross-validation within 5-fold outer cross-validation. For the random forest models, we optimized the number of features used as splits, the maximum tree depth, and the maximum number of trees used before majority voting. The hyperparameters tuned for the SVM models were the

margin width, the kernel type, and the gamma weighting parameter. The decision tree hyperparameters included the splitting criterion, the maximum number of features used, the maximum tree depth, and the minimum number of samples per split. The Gaussian Naïve Bayes model needed no hyperparameter tuning by design. Prior to training the models, the input features were normalized between 0 and 1, and univariate feature selection was performed using the chi-squared distribution, with the ten most predictive features for each modality being retained. Data normalization and feature selection were performed using the training set to protect against data leakage. Area Under Curve (AUC) was selected as the primary metric to assess model accuracy, and it was used to determine the optimal hyperparameter configurations for each model during the inner cross-validation step. The optimal hyperparameter values were then used to train and evaluate each model during the outer cross-validation step.

5.2 ABROCA

The models trained during this phase were also used to perform slicing analysis. To perform the slicing analysis, the data was split along gender lines to evaluate the model's performance in terms of ABROCA [19]. ABROCA examines a model's predictive performance across a range of classification confidence thresholds rather than restricting comparisons to fixed thresholds. The ABROCA metric is calculated by taking the absolute value of the difference between a model's AUC scores for different subpopulations of the data. ABROCA values closer to 0 show that a model's performance across subgroups is equal and therefore represents a lower amount of bias. The ABROCA statistic was chosen because it does not rely on a "similarity metric"; it is applicable across a range of confidence thresholds and can be empirically computed without requiring additional data collection. Each model's predictions from the cross-validation phase were used to generate an ROC curve to evaluate the model's predictive performance in terms of AUC and to generate the ABROCA score to quantify the bias present.

We evaluated the predictive performance and ABROCA score of the four machine learning models using multimodal data consisting of facial expression and posture combined using feature-level data fusion. The multimodal models were compared against unimodal models induced from each individual modality. We then performed a series of ablations to examine alternative combinations of modalities and machine learning techniques. The relationships and tradeoffs between the AUC and ABROCA metrics were examined to determine whether enhancing a model's fairness using debiasing techniques had a substantial impact on a model's predictive capacity.

5.3 Debiasing

Two debiasing strategies were evaluated: learned fair representation [43] and reweighing [22]. Learned fair representation (LFR) formulates fairness as an optimization problem of finding an intermediate representation of the data with opposing goals of encoding the data well while simultaneously attempting to obfuscate information related to protected attributes such as ethnicity and gender. The authors achieve this by mapping everyone, represented as a point in the input space, to a probability distribution in a new representation space. In this new space any protected information about an individual, such as ethnicity or gender, is lost while trying to maximally preserve information about the other attributes. A fair representation of the data is learned through an optimization process of mapping to a prototypical representation of the data that minimizes statistical parity. The intermediate representation can be used in fair transfer learning such that downstream models may benefit from less biased predictions.

Reweighing (RW) is a technique that attaches weights to each (group, label) combination in the dataset to ensure fair classification. It assigns tuple objects that belong to a particular group (e.g., gender) and contain a positive class label (e.g., high attention) a higher weight than objects in the same group with a negative class label. The weights are proportional to the expectation for group membership given class labels to the observed counts. Bias is defined by the difference in expected probabilities and the observed probabilities. If the expected probability is higher than the observed probability then the predictions are said to be biased. By assigning weights to each tuple according to its class label and group membership and then multiplying by its frequency it can be shown that the resulting dataset becomes unbiased and can further be used to train a bias-free classifier.

6 Results

The results of the experiments are summarized in Tables 1 and 2. Table 1 shows the AUC scores for each model with the top 20% best performing classification techniques shown in bold. Random forest tended to achieve higher performance on average across all modality combinations and debiasing strategies. The highest overall predictive accuracy was achieved by the multimodal random forest model (AUC = .832). Although Naïve Bayes had on average better scores than decision trees (Mean = .721, Mean = .627) across all modalities and debiasing strategies, the latter exhibited the next most accurate classification performance (AUC = .783). In contrast, SVM models achieved comparatively low AUC scores (Mean = .459). In general, the results indicate that random forest and Naïve Bayes models yielded the highest AUC scores on average across all experiments (Table 1). Further analyses indicate that Naïve Bayes showed the least variance from Table 1 (Var = .001) when compared to both random forest and SVM both containing a variance of .003, implying that Naïve Bayes may not be as strongly impacted by changes in modalities or debiasing strategy.

The ABROCA results are shown in Table 2 with the lowest ABROCA values in bold, signifying the models containing the least amount of bias. Bolded values denote the lowest 20% of ABROCA scores. The lowest mean ABROCA score was achieved by the Naïve Bayes models across all non-debiased and debiasing methods. The multimodal decision tree and the Kinect-only SVM with LFR were tied for the lowest ABROCA scores across all tests (ABROCA = .002). Random forest and SVM had similar mean ABROCA scores and were associated with some of the least biased scores after debiasing. The results show that both Naïve Bayes and Random Forest exhibit the lowest amount of variance of .001 and .002, respectively, across both debiasing techniques. Figures 3, 4 and 5 show the slice plots for the multimodal random forest model that achieved the highest AUC score.

Table 1: **AUC values for multimodal, Kinect, and OpenFace models across debiasing strategies.**

No Debiasing			
Model	*Multimodal*	*Kinect*	*OpenFace*
Random Forest	**0.832**	**0.811**	**0.807**
SVM	0.398	0.453	0.398
Decision Tree	0.724	0.599	0.616
Naïve Bayes	0.730	0.704	0.722
LFR			
Model	*Multimodal*	*Kinect*	*OpenFace*
Random Forest	0.777	0.659	0.664
SVM	0.538	0.530	0.531
Decision Tree	0.446	0.538	0.630
Naïve Bayes	0.694	0.754	0.685
RW			
Model	*Multimodal*	*Kinect*	*OpenFace*
Random Forest	**0.788**	**0.811**	**0.807**
SVM	0.402	0.456	0.425
Decision Tree	0.661	0.647	**0.783**
Naïve Bayes	0.777	0.704	0.722

Table 2: **ABROCA values for multimodal, Kinect, and OpenFace models across debiasing strategies.**

No Debiasing			
Model	*Multimodal*	*Kinect*	*OpenFace*
Random Forest	0.140	0.147	0.111
SVM	0.247	**0.040**	0.056
Decision Tree	**0.002**	0.098	0.054
Naïve Bayes	0.183	0.126	0.203
LFR			
Model	*Multimodal*	*Kinect*	*OpenFace*
Random Forest	0.050	0.119	**0.035**
SVM	0.084	**0.002**	**0.016**
Decision Tree	**0.017**	0.117	0.251
Naïve Bayes	0.133	0.174	0.089
RW			
Model	*Multimodal*	*Kinect*	*OpenFace*
Random Forest	**0.022**	0.147	0.111
SVM	0.196	0.121	0.123
Decision Tree	0.063	0.073	0.111
Naïve Bayes	0.106	0.126	0.203

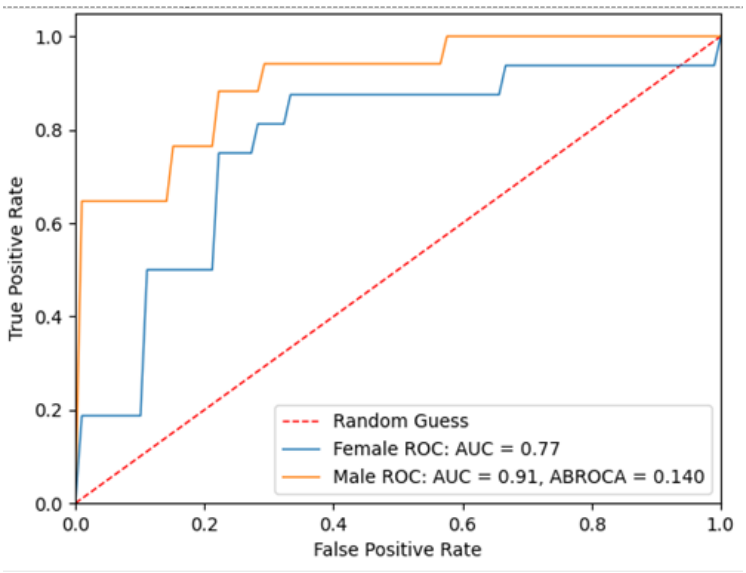

Figure 3: **Slice plot along gender showing bias in the multimodal random forest model.**

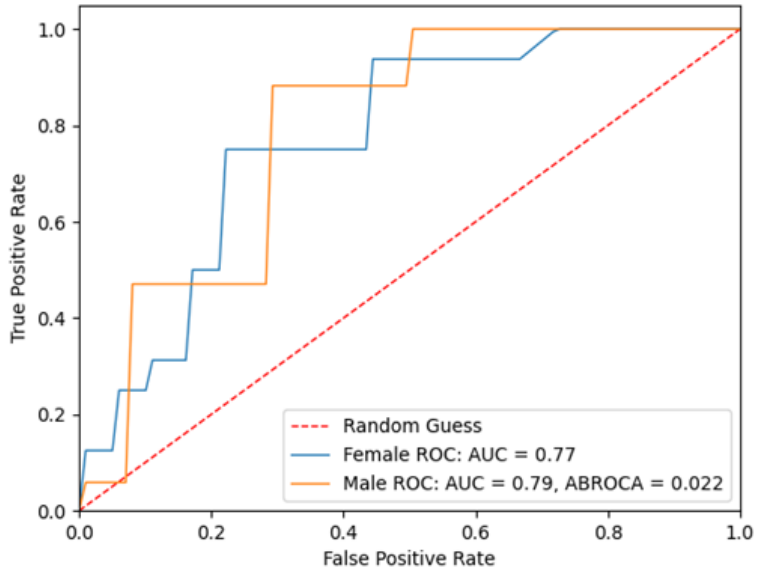

Figure 4: **Slice plot along gender showing the reduction in bias for the random forest model with reweighing.**

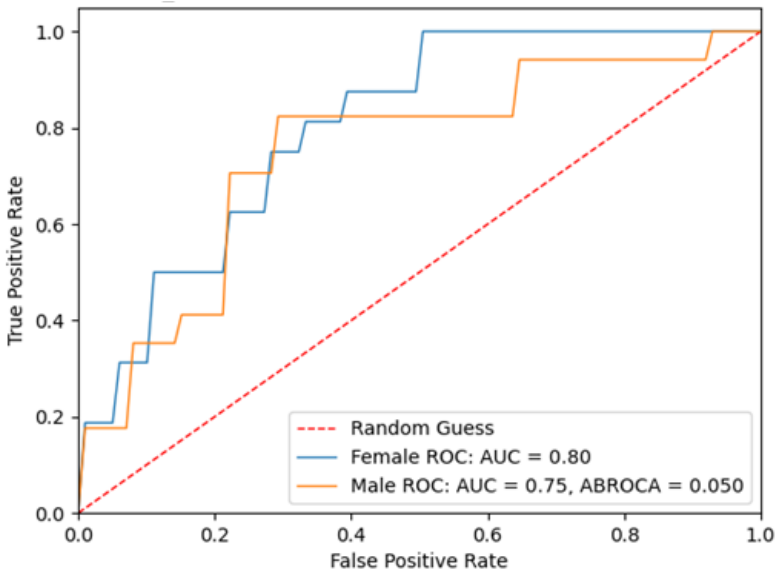

Figure 5: **Slice plot along gender showing the reduction in bias for the random forest model using learned fair representations.**

7 Discussion

With regard to predictive accuracy, random forest was the best performing model across all experiments, producing 6 of the 7 highest AUC scores. The multimodal and posture-only (Kinect) models yielded the highest ABROCA values before the application of debiasing techniques, suggesting a risk of encoded bias. After applying the LFR debiasing approach to random forest to generate a latent representation of the data, we saw a reduction in ABROCA values for each modality, with the facial expression modality showing the largest reduction in bias (0.111 to 0.035; 68%). Application of debiasing via reweighing for random forest reduced bias in the multimodal model by 84% (0.140 to 0.022) although it did not reduce bias for either of the unimodal models. These findings demonstrate that the two debiasing strategies are effective in mitigating bias present in the random forest models of visitor visual attention based upon using the ABROCA statistic for comparing models' bias. In addition, we observed that debiasing via reweighing for the multimodal data greatly reduces bias while having a minimal impact on the model's visual attention classification accuracy.

Visitor attention models trained on multimodal data without debiasing achieved the highest ABROCA scores, implying that the bias present in the underlying features from each modality may be associated with increasing complexity of the multimodal data. The effectiveness of our debiasing strategies for the multimodal feature set can be observed as both the LFR and RW debiasing strategies achieved greater decreases in ABROCA scores for the multimodal models than when applied to the unimodal baseline models. This illustrates that the tradeoff between the ABROCA and AUC scores was critical for the multimodal models, as there was a significant impact from the application of the debiasing strategies.

The results in Table 2 also indicate that each machine learning model responded differently to each debiasing technique. The RW method had little effect on the random forest and Naïve Bayes models in terms of ABROCA, and it did not significantly impact the predictive performance of the unimodal baselines. Both models are robust to noise and outliers and therefore may be minimally impacted by the weights of the RW method applied to the dataset. However, the additive bias from the combination of modalities was diminished by the RW technique, and in the case of random forest, it benefited by having one of the lowest ABROCA scores overall. Decision tree and SVM experienced an increase in classification performance, but each model had varied results in terms of the slicing analysis. The ABROCA scores for the facial expression-based models increased for each model after debiasing, and the posture modality showed an increase with the SVM as well. The increase in classification accuracy for the SVM may be caused by a balance in the separation between classes introduced by reweighing. The LFR debiasing technique decreased the bias present in at least one model for each modality, with the lone exception of Naïve Bayes. The latent representations of the features provided by LFR-based debiasing appear to be a more effective means of debiasing when compared to reweighing in general. RW was effective by demonstrating the lowest ABROCA values with low impact on AUC for multimodal random forest models (Table 2).

We can categorize each of these techniques by looking at to what extent the decrease in a model's classification accuracy affects the model's decrease in bias. LFR performs a more complex transformation on the training data compared to RW, and subsequently it has a larger impact on a model's predictive accuracy. This debiasing technique can also have a generally large, possibly adverse impact on bias reduction. The RW technique employs a relatively simple transformation on the data that has a smaller impact on predictive performance and ABROCA scores.

When determining which visitor visual attention models to deploy within a museum, considerations of the models' overall accuracy should be weighed against the risks associated with biased performance. It is possible that biased performance within certain subgroups has little tangible impact on visitors' learning experiences, thus imposing a negligible cost to the users of the model (e.g., visitors, exhibit designers, museum educators, museum researchers). Conversely, bias present in models of visitor attention may introduce unfair treatment between different groups, reinforcing existing biases and creating new sources of inequity. Most immediately, biased models of visitor attention could translate into the creation of exhibits that are less engaging for many learners, reducing overall engagement across the museum and visitor population. Thus, it is likely advisable to balance between model

accuracy and bias minimization. The issue of determining an "acceptable" amount of bias remains domain-specific and is a promising area of future research.

8 Conclusion

Multimodal learning analytic techniques hold significant promise for modeling visitor attention in museums by modeling multiple concurrent perspectives on visitors' behavioral cues. Many museums serve diverse learner populations with respect to age, gender, socio-economic status, and cultural background. Algorithmic fairness is a critically important issue in developing multimodal machine learning-based models of visitor attention that are accurate and free of encoded bias. We have presented a slicing analysis approach for identifying and mitigating encoded bias in multimodal models of visitor visual. This approach utilizes the ABROCA metric to quantify and evaluate bias within multimodal and unimodal visitor visual attention models, enabling analysis of different debiasing strategies in terms of predictive performance and performance differences between sub-groups. Results from a study using multimodal visitor interaction data from the *FUTURE WORLDS* game-based museum exhibit suggest that multimodal random forest models yield accurate predictions of visitor visual attention, but these models suffer from bias along gender lines. Debiasing via reweighing was found to be effective in mitigating bias from multimodal attention models while having low impact on predictive performance. In addition, we found that different combinations of machine learning techniques and modalities responded differently to applications of different debiasing methods.

The findings suggest several promising avenues for future work. Investigating encoded bias in multimodal visual attention models based upon alternative machine learning architectures, including deep neural networks, is an important direction for future investigation. Another promising direction is developing an algorithmic approach to debiasing that explicitly optimizes for the ABROCA statistic; neither reweighing nor learned fair representations utilize ABROCA to guide debiasing. Extending the debiasing strategies to optimize for ABROCA may enable them to encompass both group and individual fairness. A key attribute of the debiasing strategies used in this work is that they focus on preprocessing data to address sources of encoded bias. Further work should explore whether the findings observed in this study hold when using other debiasing techniques such as adversarial debiasing. Finally, it will be important to investigate the integration of debiased multimodal models of visitor visual attention into museum exhibits to enable run-time measurement and support of visitor experiences. This capability has promise for enhancing museum exhibits' capacity to create learning experiences that are effective and engaging for all learners.

ACKNOWLEDGMENTS

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

REFERENCES

<bib id="bib1"><number>[1]</number>Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, Hannah Wallach. 2018. A reductions approach to fair classification. Retrieved from https://arxiv.org/abs/1803.02453.</bib>
<bib id="bib2"><number>[2]</number>Alissa Antle, Allen Bevans, Theresa Tanenbaum, Katie Seaborn, and Sijie Wang. 2010. Futura. In Proceedings of the Fifth International Conference on Tangible, Embedded, and Embodied Interaction. 93–100. https://doi.org/10.1145/1935701.1935721.</bib>
<bib id="bib3"><number>[3]</number>Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 2 (2019), 423–443. https://doi.org/10.1109/tpami.2018.2798607 </bib>
<bib id="bib4"><number>[4]</number>Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. 2016. OpenFace: An open source facial behavior analysis toolkit. In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). https://doi.org/10.1109/wacv.2016.7477553 </bib>
<bib id="bib5"><number>[5]</number>Nigel Bosch, Chen Huili, Sidney D'Mello, Ryan Baker, and Valerie Shute. 2015. Accuracy vs. availability heuristic in multimodal affect detection in the wild. In Proceedings of the 2015 International Conference on Multimodal Interaction. 267-274.</bib>
<bib id="bib6"><number>[6]</number>Cheng Chang, Cheng Zhang, Lei Chen, and Yang Liu. 2018. An ensemble model using face and body tracking for engagement detection. In Proceedings of the 20th International Conference on Multimodal Interaction. 616-622.</bib>
<bib id="bib7"><number>[7]</number>Chantal Barriault and David Pearson. 2010. Assessing exhibits for learning in science centers: A practical tool. Visitor Studies 13, 1 (2010), 90–106. https://doi.org/10.1080/10645571003618824 </bib>
<bib id="bib8"><number>[8]</number>Eustasio del Barrio, Paula Gordaliza, Jean-Michel Loubes. 2020. Review of mathematical frameworks for fairness in machine learning. Retrieved from https://arxiv.org/abs/2005.13755.</bib>
<bib id="bib9"><number>[9]</number>Rachel Bellamy et al. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. Retrieved from https://arxiv.org/abs/1810.01943.</bib>
<bib id="bib10"><number>[10]</number>Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research 50, 1 (2018), 3–44. https://doi.org/10.1177/0049124118782533 </bib>
<bib id="bib11"><number>[11]</number>Elise Cappella, Ha Yeon Kim, Jennifer Neal, and Daisy Jackson. 2013. Classroom peer relationships and behavioral engagement in elementary school: The role of social network equity. American Journal of Community Psychology 52, 3–4, 367–379. https://doi.org/10.1007/s10464-013-9603-5 </bib>
<bib id="bib12"><number>[12]</number>Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. Communications of the ACM 63, 5, 82–89. https://doi.org/10.1145/3376898 </bib>
<bib id="bib13"><number>[13]</number>Sidney D'Mello and Arthur Graesser. 2010. Mining bodily patterns of affective experience during learning. In Proceedings of the 3rd International Conference on Educational Data Mining. 31–40.</bib>
<bib id="bib14"><number>[14]</number>Sidney D'Mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. ACM Computing Surveys 47, 3 (2015), 1–36. https://doi.org/10.1145/2682899</bib>
<bib id="bib15"><number>[15]</number>Soumia Dermouche and Catherine Pelachaud. Engagement modeling in dyadic interaction. 2019. In Proceedings of the International Conference on Multimodal Interaction. 440-445.</bib>
<bib id="bib16"><number>[16]</number>Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, Max Leiserson. 2017. Decoupled classifiers for fair and efficient machine learning. Retrieved from https://arxiv.org/abs/1707.06613.</bib>
<bib id="bib17"><number>[17]</number>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. https://doi.org/10.1145/3382507.3418890</bib>
<bib id="bib18"><number>[18]</number>Sorelle Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3287560.3287589 </bib>
<bib id="bib19"><number>[19]</number>Josh Gardner, Christopher Brooks, and Ryan Baker. 2019. Evaluating the fairness of predictive student models through slicing analysis. In Proceedings of the 9th International Conference on Learning Analytics & Knowledge. https://doi.org/10.1145/3303772.3303791 </bib>
<bib id="bib20"><number>[20]</number>Ursula Hébert-Johnson, Michael Kim, Omar Reingold, Guy Rothblum. 2017. Calibration for the (computationally-identifiable) masses. Retrieved from https://arxiv.org/abs/1711.08513.</bib>
<bib id="bib21"><number>[21]</number>Stephen Hutt, et al. 2019. Evaluating fairness and generalizability in models predicting on-time graduation from college applications. International Educational Data Mining Society (2019), 78-88.</bib>
<bib id="bib22"><number>[22]</number>Faisal Kamiran and Toon Calders. 2011. Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems 33, 1 (2011), 1–33. https://doi.org/10.1007/s10115-011-0463-8</bib>
<bib id="bib23"><number>[23]</number>Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. An empirical study of rich subgroup fairness for machine learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency. 100–109. https://doi.org/10.1145/3287560.3287592.</bib>
<bib id="bib24"><number>[24]</number>René Kizilcec and Hansol Lee. 2020. Algorithmic fairness in education. Retrieved from https://arxiv.org/abs/2007.05443 </bib>
<bib id="bib25"><number>[25]</number>David Madras, Elliot Creager, Tonian Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. In Proceedings of the 35th International Conference on Machine Learning. 80:3384-3393.</bib>
<bib id="bib26"><number>[26]</number>Ninareh Mehrabi, Fred Morstatter, Nrisputa Saxena, Kristina Lerman, Aram Galstyan. 2019. A survey on bias and fairness in machine learning. Retrieved from https://arxiv.org abs/1908.09635.</bib>
<bib id="bib27"><number>[27]</number>Yukiko Nakano and Ryo Ishii. 2010. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In Proceedings of the 15th International Conference on Intelligent User Interfaces. https://doi.org/10.1145/1719970.1719990 </bib>

<bib id="bib28"><number>[28]</number>Amol Patwardhan and Gerald Knapp. 2016. Multimodal affect recognition using Kinect. Retrieved from https://arxiv.org/abs/1607.02652 </bib>
<bib id="bib29"><number>[29]</number>Andrew Polaine. 2005. The flow principle in interactivity. In Proceedings of the Second Australasian Conference on Interactive Entertainment. Creativity & Cognition Studios Press, Sydney, AUS, 151–158.</bib>
<bib id="bib30"><number>[30]</number>Mirko Raca and Pierre Dillenbourg. 2013. System for assessing classroom attention. In Proceedings of the Third International Conference on Learning Analytics and Knowledge. https://doi.org/10.1145/2460296.2460351</bib>
<bib id="bib31"><number>[31]</number>Jonathan Rowe, Eleni Lobene, Bradford Mott, and James Lester. 2017. Play in the museum. International Journal of Gaming and Computer-Mediated Simulations 9, 3: 96–113. https://doi.org/10.4018/ijgcms.2017070104</bib>
<bib id="bib32"><number>[32]</number>Stefanie Rukavina, Sascha Gruss, Holger Hoffmann, and Harald Traue. 2016. Facial expression reactions to feedback in a human-computer interaction— Does gender matter? Psychology 07, 03 (2016), 356–367. https://doi.org/10.4236/psych.2016.73038</bib>
<bib id="bib33"><number>[33]</number>Robert Sawyer, Andy Smith, Jonathan Rowe, Roger Azevedo, and James Lester. 2017. Enhancing student models in game-based learning with facial expression recognition. In Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization. https://doi.org/10.1145/3079628.3079686</bib>
<bib id="bib34"><number>[34]</number>D. Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. 2018. Winner's curse? On pace, progress, and empirical rigor. In ICLR Workshops.</bib>
<bib id="bib35"><number>[35]</number>Ömer Sümer, Patricia Goldberg, Sidney D'Mello, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2021. Multimodal engagement analysis from facial videos in the classroom. Retrieved from https://arxiv.org/pdf/2101.04215.pdf </bib>
<bib id="bib36"><number>[36]</number>Chinchu Thomas, Nitin Nair, and Dinesh Babu Jayagopi. 2018. Predicting engagement intensity in the wild using temporal convolutional network. In Proceedings of the 20th ACM International Conference on Multimodal Interaction. 604-610.</bib>
<bib id="bib37"><number>[37]</number>Alexandria Vail, Kristy Boyer, Eric Wiebe, and James Lester. 2015. The mars and venus effect: The influence of user gender on the effectiveness of adaptive task support. In Lecture Notes in Computer Science. Springer International Publishing, 265–276. https://doi.org/10.1007/978-3-319- 20267-9_22 </bib>
<bib id="bib38"><number>[38]</number>Alexandria Vail, Joseph Grafsgaard, Kristy Boyer, Eric Wiebe, and James Lester. 2016. Gender differences in facial expressions of affect during learning. In Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization. 65-73. https://doi.org/10.1145/2930238.2930257 </bib>
<bib id="bib39"><number>[39]</number>Marcelo Worsley, Dor Abrahamson, Paulo Blikstein, Shuchi Grover, Bertrand Schneider, Mike Tissenbaum. 2016. Situating multimodal learning analytics. In Proceedings of the International Conference of the Learning Sciences. 1346-1349.</bib>
<bib id="bib40"><number>[40]</number>Suowei Wu, Zhengyin Du, Weixin Li, Di Huang, and Yunhong Wang. Continuous emotion recognition in videos by fusing facial expression, head pose and eye gaze. 2019. In Proceedings of the International Conference on Multimodal Interaction. 40-48.</bib>
<bib id="bib41"><number>[41]</number>Jianfei Yang, Kai Wang, Xiaojiang Peng, and Yu Qiao. 2018. Deep recurrent multi-instance learning with spatio-temporal features for engagement intensity prediction. In Proceedings of the 20th International Conference on Multimodal Interaction. 594-598.</bib>
<bib id="bib42"><number>[42]</number>Susan Yoon and Joyce Wang. 2013. Making the invisible visible in science museums through augmented reality devices. TechTrends 58, 1 (2013), 49– 55. https://doi.org/10.1007/s11528-013-0720-7 </bib>
<bib id="bib43"><number>[43]</number>Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In Proceedings of the 30th International Conference on International Conference on Machine Learning. 325-333.</bib>
<bib id="bib44"><number>[44]</number>Zhang, Zhe and Daniel Neill. 2016. Identifying significant predictive bias in classifiers. Retrieved from https://arxiv.org/abs/1611.08292.</bib>