Enhancing Affect Detection in Game-Based Learning Environments with Multimodal Conditional Generative Modeling

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

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

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

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

ABSTRACT

Accurately detecting and responding to student affect is a critical capability for adaptive learning environments. Recent years have seen growing interest in modeling student affect with multimodal sensor data. A key challenge in multimodal affect detection is dealing with data loss due to noisy, missing, or invalid multimodal features. Because multimodal affect detection often requires large quantities of data, data loss can have a strong, adverse impact on affect detector performance. To address this issue, we present a multimodal data imputation framework that utilizes conditional generative models to automatically impute posture and interaction log data from student interactions with a game-based learning environment for emergency medical training. We investigate two generative models, a Conditional Generative Adversarial Network (C-GAN) and a Conditional Variational Autoencoder (C-VAE), that are trained using a modality that has undergone varying levels of artificial data masking. The generative models are conditioned on the corresponding intact modality, enabling the data imputation process to capture the interaction between the concurrent modalities. We examine the effectiveness of the conditional generative models on imputation accuracy and its impact on the performance of affect detection. Each imputation model is evaluated using varying amounts of artificial data masking to determine how the data missingness impacts the performance of each imputation method. Results based on the modalities captured from students' interactions with the game-based learning environment indicate that deep conditional generative models within a multimodal data imputation framework yield significant benefits compared to baseline imputation techniques in terms of both imputation accuracy and affective detector performance.

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

ACM ISBN 978-1-4503-7581-8/20/10 $15.00
https://doi.org/10.1145/3382507.3418892

CCS CONCEPTS

• Computing Methodologies • Machine Learning • Machine Learning Approaches

KEYWORDS

Affective Modeling; Data Imputation; Game-Based Learning Environments; Generative Adversarial Networks; Variational Autoencoders

ACM Reference format:

Nathan Henderson, Wookhee Min, Jonathan Rowe, and James Lester. 2020. Enhancing Affect Detection in Game-Based Learning Environments with Multimodal Conditional Generative Modeling. In *2020 International Conference on Multimodal Interaction (ICMI '20), October 25-29, 2020, Utrecht, The Netherlands.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3382507.3418892

1 Introduction

Affect plays a critical role in student learning [11]. Recent years have seen a significant interest in using physical sensors to measure students' affect as they engage with adaptive learning environments. Sensors enable affective models that can generalize across learning environments by not relying upon environment-specific inputs. Sensor-based systems have also seen an increase in cost-effectiveness over time, leading to increased accessibility and scalability. Multimodal sensor systems employ a wide variety

of modalities, including eye gaze [31], posture [16], electrodermal activity (EDA) [33], and facial expression [8]. Interaction log-based modalities, such as keystroke [35] or game trace data [27], are also commonly employed within affective modeling due to their non-intrusive nature and ease of implementation, particularly when physical sensors are prohibitive or impractical [2,6]. The use of two or more modalities for affect detection has been shown to achieve notable improvement over unimodal models [15, 18] as it provides multiple concurrent perspectives on a student's behavior and learning [3].

Both sensor-based and interaction-based affect detection systems often encounter issues that distort or prohibit consistent data capture. Physical and physiological sensors can be impeded by noise [16], mistracking [4], and data storage constraints. Interaction log-based modalities also suffer from issues such as software or hardware failure, network connection problems, data logging and transfer problems [30], and incompleteness issues [25]. Data loss can also occur due to practical challenges that are common in educational settings, including schools' reliance on aging computers, accidental unplugging, and student mishandling of machines. Common approaches to addressing these challenges include discarding data samples with missing data and simple imputation methods such as mean imputation. However, discarding data significantly reduces the amount of training data available for machine learning models for affect detection.

In this paper, we investigate multimodal data imputation with deep conditional generative models to address data loss issues in student affect detection. We utilize deep conditional generative models because of their capability to effectively model complex relationships between multiple input variables and data channels. Generative neural models have seen increasing usage in areas such as synthetic image generation [38], facial expression recognition [20], data augmentation [7], translation [22], and data imputation [27]. The application of conditional generative models to multimodal data streams has received comparatively little attention, particularly in the context of modeling student affect in adaptive learning environments.

Specifically, we investigate two types of conditional generative models for multimodal data imputation: conditional GANs (C-GANs) and conditional VAEs (C-VAEs). The models are evaluated within a multimodal affect detection framework that tracks posture data and interaction trace data from students engaged with a game-based learning environment for emergency medical skills training. Affect detection models are induced to predict learning-centered affective states obtained from field observations of each student. The models are evaluated using varying levels of "missingness" to demonstrate the impact that intact data availability has on each generative model. The effectiveness of each imputation method is evaluated based on its impact on the predictive performance of multimodal affect detectors that are previously trained on all available multimodal data without any masking. Results indicate that the non-linear generative models based on deep neural networks show significant promise compared to several competing linear baseline approaches for data imputation.

2 Related Work

2.1 Multimodal Affect Recognition

Improvements in the accessibility and flexibility of multimodal sensor systems have fostered increased interest in multimodal affect recognition in adaptive learning environments. Grafsgaard et al. used multiple sensor- and interaction-based modalities such as posture, facial expression, dialogue, and task actions to predict engagement, frustration, and normalized learning gain in students [15]. Their results found a positive relationship between the number of modalities used to induce predictive affect models and the models' overall performance. Yang et al. explored the use of gaze and posture data in identifying instances of engagement elicited as viewers watched various educational videos [35]. Wu et al. utilized various multimodal data fusion techniques to perform continuous emotion recognition based on facial expression, pose, and eye gaze data, concluding that their multimodal model using all modalities achieved higher performance than other unimodal and bimodal baselines [34]. Bosch et al. utilized facial expression data to detect student emotion during game-based learning in naturalistic classroom settings. Due to various issues with the facial expression data capture, the authors also employed gameplay interaction data from instances where facial expression data was missing or corrupted; however, the gameplay data was not found to be as predictive as the facial expression data [5]. Henderson et al. investigated the predictive value of various features engineered from students' posture and gameplay trace log data [18]. Additionally, various methods of combining the modalities through multimodal data fusion were also evaluated. D'Mello and Kory provide an in-depth review of multimodal affect detection systems, drawing the conclusion that 85% of the observed multimodal systems were more effective than their unimodal counterparts [12]. However, in many multimodal affect recognition tasks, captured data that is noisy or otherwise invalid is often discarded completely, significantly reducing the amount of data that is available for inducing affect models and potentially adversely impacting the overall performance.

2.2 Multimodal Data Imputation

Noisy data and missing data are common challenges in multimodal systems, and multimodal data imputation techniques show significant promise for addressing these challenges. Jaques et al. used a multimodal autoencoder to reconstruct data given missing modalities and used the latent representation of the imputed data to train affect models [19]. Yang et al. also used an autoencoder-based approach to address block-wise missingness from three sensor-based and interaction log-based modalities. They used autoencoders to produce latent-space representations of single, pairwise, and entire modalities, reconstructing the original data based on the imputation using the pairwise mappings [36]. Liu followed a similar imputation approach using multimodal autoencoders while applying the latent representations of multiple modalities within a translation framework to allow the decoder to produce the missing modality

[21]. Thung et al. utilized a deep multitask model to impute missing data in separate modalities based on the pairwise combinations of missing and non-missing data [32]. Shang et al. introduced a multimodal data imputation method using generative adversarial networks [27]. By using the GAN model to learn cross-modality mappings, a multimodal denoising autoencoder was trained on augmented mapping data to reconstruct a missing modality.

The primary contribution of this work is demonstrating the improved performance of deep conditional generative models—specifically, a C-GAN and a C-VAE—relative to baseline approaches for multimodal data imputation in modeling student affect. We evaluate each model against multiple baseline approaches using three different levels of data "missingness" and two masked modalities for a comprehensive analysis of data imputation performance. We demonstrate the ability of conditional generative models to minimize the impact of missing data on multimodal affect detection models trained on all intact modalities.

Figure 1: TC3Sim game-based learning environment

3 Dataset

The dataset we use to evaluate generative data imputation models contains posture and gameplay data captured from students engaged with a game-based learning environment for emergency medical skills training, TC3Sim. The data was collected during a study involving 119 students (83% male, 17% female) at the United States Military Academy. Posture data from each student was captured using a front-facing Microsoft Kinect sensor, while the gameplay interaction logs were captured using GIFT, an open-source software framework for the development and deployment of adaptive learning environments [29]. As students engaged with TC3Sim, two researchers observed each of the students and recorded their perceived affective states according to the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) [24].

3.1 TC3Sim Game-Based Learning Environment

In TC3Sim, students assume the role of a combat medic within simulated military combat scenarios (Figure 1). Throughout gameplay, students complete a series of interactive medical scenarios within a 3D virtual environment, administering different forms of combat casualty care to injured computer-controlled teammates. Each student used TC3Sim individually, with each gameplay session lasting approximately one hour.

3.2 BROMP Protocol

Using the BROMP observation protocol, two trained observers walked around the perimeter of the classroom and discreetly annotated occurrences of different affective states using an affect recording application running on a small mobile device. Each student was observed at approximately 20-second intervals. Using observations collected at the beginning of the study, the two observers reached an inter-rater agreement in excess of 0.6 in terms of Cohen's Kappa [10]. For this study, six distinct learning-centered affective states were recorded: *boredom, confusion, engaged concentration, frustration, surprise*, and *anxiety*. A total of 755 BROMP observations were captured: 435 were labeled as *engaged concentration* (M = .576, SD = .239), 174 as *confusion* (M = .231, SD = .185), 73 as *boredom* (M = .097, SD = .161), 32 as *frustration* (M = .042, SD = .182), 29 as *surprise* (M = .038, SD = .045), and 12 as *anxiety* (M = .016, SD = .089). Due to the relatively low number of observations of *anxiety*, we exclude this affective state from our analysis.

4 Methodology

We seek to address issues of missing or noisy data in sensor- and interaction log-based data channels by investigating several data imputation methods based on a single artificially masked modality. We investigate multimodal data imputation with conditional generative modeling as follows. Feature engineering is performed on both modalities based on the data corresponding with each BROMP observation. Following this process, baseline affect detectors are built using features from both modalities. To evaluate the data imputation performance of the generative models, one of the modalities is masked by setting either 25%, 50%, or 75% of each student's data to be missing. Following this process, generative models are trained using the non-missing data for each masked modality and are conditioned using the corresponding features from the intact modality. The trained generative models then impute the masked data, which are compared to the unmasked values which serve as ground truth for validating the imputation models. Each model is evaluated in terms of root mean squared error (RMSE) to determine the optimal generative model configuration. The data imputed by each optimal generative model is then used to train new affect detection models for each affective state following the same configuration as the baseline models. Finally, we compare the deviation of the models' performance to the baseline models' performance on datasets where no data is masked.

4.1 Posture-Based Feature Engineering

The Kinect sensor captured 3D coordinate data for 91 distinct vertices. Based on prior work related to posture-based affect detection, we selected three vertices from which 73 posture-based features were distilled, *top_skull*, *center_shoulder*, and *head* [14]. Each of the features was computed based upon the students' posture and movement prior to the given BROMP observation. Each vertex produced 18 statistical features, including features such as most recent observed distance, most recent Z-coordinate value, minimum and maximum observed distance, median observed distance, and variance in the observed distances. Distance was defined as the Euclidean distance between the vertex and the Kinect sensor. Additionally, summative features were calculated for each vertex using the minimum, maximum, median, and variance in distance observed across the preceding 5, 10, and 20 seconds prior to each BROMP observation. In addition to these 54 features, several features were generated to provide the total change in position and distance from the Kinect sensor over the prior 3 and 20 seconds. Finally, features were engineered to indicate whether the student was leaning forward, backward, or sitting upright based upon the median distance of the *head* vertex for each individual workstation and the current position of the *head* vertex. These three features were calculated across time windows of 5, 10, and 20 seconds, as well as the entire gameplay session up to the current BROMP observation.

4.2 Interaction Log-Based Feature Engineering

The interaction log-based features are extracted from students' interaction (i.e., gameplay) trace logs from TC3Sim. These features represent the students' actions within the game as well as information about the virtual patients that received treatment during each training scenario. Features representing the states of the virtual patients include changes in systolic blood pressure and heart rate, exposed wounds, and lung volume. Features were also generated based on students' gameplay actions such as checking a patient's vital signs or requesting a medical evacuation. Each of the features was calculated cumulatively over the preceding 20 seconds prior to a BROMP observation and reported in terms of the sum or current count of a certain action. Additionally, measures such as the virtual patient's blood pressure were reported using the standard deviation or average. This process produced 39 distinct interaction log-based features.

4.3 Affect Model Evaluation

Based on the BROMP observations, separate datasets were created for each of the five affective states with a binary class label indicating whether the recorded BROMP observation corresponded to a positive instance of the target emotion class (e.g. *bored, confused, engaged concentration, frustrated, surprised*). Following this process, the datasets were divided into separate training and test sets, which were split along a student-level to avoid data leakage from a single student's data during model training and evaluation. Additionally, stratified sampling was used to ensure a relatively similar class distribution between the training and test sets. Approximately 80% of the data was used as training data, and the remaining 20% was used as a held-out test set. To resolve class imbalances within each dataset, the Synthetic Minority Oversampling Technique (SMOTE) was applied to each training fold [9]. This process randomly selects a positive instance of the minority class and generates synthetic data based on linear interpolation between the selected point and another positive instance chosen using randomized K-nearest neighbor selection.

Automated feature selection was performed based on the training set by identifying the features with the highest chi-squared correlation with the binary emotion label. To combine the multimodal feature data, 15 features were selected from each modality and concatenated at a feature-level to train each affect model [17]. Following the feature selection process, five classification techniques were evaluated for each affective state: Support Vector Machine (SVM), Logistic Regression (LR), Gaussian Naïve Bayes (NB), Random Forest (RF), and Multi-Layer Perceptron (MLP). The models were trained using 4-fold cross-validation while performing iterative grid search on the model hyperparameters to determine the optimal model, with the best model's performance reported using the held-out test set. During the cross-validation process, synthetic data due to upsampling was removed from each fold used as a validation set to avoid artificially inflated performance.

4.4 Conditional Generative Adversarial Networks

C-GANs are an extension of generative adversarial networks (GANs), which consist of two deep neural networks, a *generator* and a *discriminator*, that "compete" against one another other in an adversarial fashion [13]. Using a noise vector as input, the generator attempts to produce synthetic data that will deceive the discriminator, which subsequently attempts to determine whether its input is synthetic (i.e., "fake" data) or sampled from the actual data (i.e., "real" data). The loss of the discriminator is backpropagated through both components of the GAN, with the goal of teaching the generator to produce increasingly realistic augmented data, while the discriminator also learns to accurately distinguish between the real and fake data samples. GAN convergence is theoretically achieved when the model achieves a Nash equilibrium [1]. Conditional GANs (C-GANs) expand upon the traditional GAN model by training the discriminator on additional data, or "conditions", associated with the input feature vector, such as a class label [23]. Likewise, the generator is trained on the same additional conditions alongside the input noise vector. This allows for the generator and discriminator to be guided by the conditional input, so the data augmentation process is not completely stochastic. In our work, we seek to use the intact (non-masked) modality (i.e., posture or interaction log data) to be the conditional input to the generator, so the generator imputes the missing modality based on the associated, non-missing data.

4.5 Conditional Variational Autoencoders

C-VAEs are similar to C-GANs with regard to the conditioning of a generative model [28]. The standard VAE contains two neural network models, an encoder and a decoder. The encoder learns to

model latent variable representations of the input data, while the decoder reconstructs the original input based upon the generated latent representation. The VAE model constrains the latent representation to follow a specified probability distribution, typically a Gaussian distribution. Thus, the loss function of the VAE typically consists of two terms: one based on the reconstruction error and the other based on the Kullback-Leibler divergence between the two relevant distributions (i.e., the latent representation distribution and the Gaussian distribution). In the C-VAE implementation, the input of the encoder as well as the decoder are both conditioned using the same corresponding condition vector as the C-GAN. In this work, we use the same intact modality as the condition to the C-VAE model.

4.6 Generative Model Training

Each generative model is evaluated by masking either the interaction log modality or the posture modality. To evaluate the performance of the generative models for varying levels of missing data, each modality is masked by selecting 25%, 50%, or 75% of the data points (BROMP observations) for each student. The posture data is masked intermittently throughout the student's session, with each data point having an equal probability of being masked. This is equivalent to masking posture data in 20-second intervals, as each BROMP observation coincides with a roll up of the prior 20 seconds of student behavior. By masking 20-second intervals, we effectively simulate sporadic data loss, such as that caused by mistracking, a student exiting the sensor's field of view, or intermittent sensor error, where posture data is missing for consecutive readings. The interaction log data is masked by masking the last 25%, 50%, or 75% of the student's data. This resembles real-world situations where an adaptive learning environment crashes or fails resulting in data loss for the remainder of the student's session. The data is masked by setting all features in the masked modality to be missing for a selected data sample or sequence. The original values are stored as ground-truth data for evaluation of the data imputation methods. This masking process is performed on the training data described earlier for inducing affect detection models.

Following this phase, fully connected, generative models are trained on the non-masked data. The C-VAE is trained using non-masked features from the masked modality as input, and the features from the intact modality are used as conditions for the model. The C-GAN also uses features from the intact modality as the conditions, but the generator takes a Gaussian noise vector of size 32 as input. The C-GAN's discriminator takes an input of either "fake" data produced by the generator or "real" data consisting of non-missing data samples from the masked modality. The generative models are trained for 1,000 epochs each with hyperparameter tuning performed on the number of layers in the generator and discriminator (C-GAN), the number of layers in the encoder and decoder (C-VAE), and learning rate. Each generative model was optimized using the ADAM optimization algorithm and binary cross entropy as the loss function. Additionally, each model utilized a hyperbolic tangent activation function in the hidden layers, necessitating a normalization of the data to be within the range of -1 and 1.

The optimal model was determined by calculating the RMSE between the imputed values and the original, ground-truth values. The selected model was then used to impute the missing values in the training set which was subsequently used to train affect detection models as described in Section 4.3. The changes in the affect models' predictive accuracy demonstrate how different data imputation methods approximate masked or removed predictive features or trends in the affect training data. In this way, the effectiveness of the generative data imputation methods is evaluated in two different thrusts. For comparison, we evaluate the generative models against two baseline methods: mean imputation and probabilistic matrix factorization (PMF) [26]. Mean imputation was implemented by taking the mean value of each feature within a student's session data. PMF has become relatively common within recommender systems, where data imputation is a frequently encountered task. This process factors a sparsely populated matrix into two distinct lower-rank matrices, the multiplication of which approximates the original data. This process is repeated iteratively to minimize the reconstruction loss of the imputed data using expectation maximization.

5 Results

To determine the impact that missing data and the resulting imputation have on affect detection, baseline models induced using multimodal data that is completely unmasked are evaluated. The ideal performance of the generative model would result in the imputed data matching the previously masked data samples perfectly, which would result in no deviation from the performance of the models trained on non-masked data. This allows the impact of the data imputation to be evaluated relative to the original, non-masked data, as well as to the performance of the affect models trained on masked and non-masked data.

5.1 Affect Detection Results

We examine the predictive performance of the affect detection models trained with complete data using area under curve (AUC), accuracy, precision, recall, and F1 score. The best performing classifier, hyperparameter configuration, and selected features based on non-masked data are then preserved for the data imputation phase. This allows for any deviation in the model's performance to be attributed to the data imputation process rather than other factors. For the purpose of this work, the affect detection models' predictive performance serves as a "gold standard" for examining how different generative imputation methods impact the performance of the affect models trained with missing data.

Table 1 shows that multi-layer perceptron was the optimal classification model for each of the affective states. All of the affect models achieved an AUC greater than random chance (0.500) with the exception of confused. Frustrated and surprised had relatively low precision and recall values, which could be attributed to class imbalances, as positive instances of both classes individually comprised less than 5% of the total dataset.

Table 1: Performance of optimal affect models.

Emotion	Model	AUC	Accuracy	Precision	Recall	F1 Score
Boredom	MLP	0.837	0.840	0.395	0.833	0.536
Confused	MLP	0.462	0.463	0.202	0.459	0.281
Engaged Concentration	MLP	0.620	0.636	0.628	0.807	0.706
Frustrated	MLP	0.638	0.759	0.128	0.500	0.204
Surprised	MLP	0.594	0.877	0.118	0.286	0.167

5.2 Interaction Log Data Imputation

The four imputation models were evaluated across the five affective states, each using three possible levels of missing data. The interaction log modality was masked using the end of each sequence as the masking location to simulate the loss of all interaction log data following a software failure in the middle of a student's interaction with the game-based learning environment. The results of data imputation on the affect model training data are shown in terms of RMSE in Figure 2. The impact of data imputation on affect models using the optimal hyperparameter configuration from each model in Table 1 is shown in Figure 3. The impact on the affect models' predictive performance is measured in terms of the absolute difference between the original affect model's AUC and the AUC of the same model when trained on the imputed data and evaluated on the same held-out test set.

As shown in Figure 2, deep generative models yielded the best data imputation performance for 4 of the 5 affective states when 25% of the interaction log modality was masked, 2 of the 5 when 50% was masked, and 3 of the 5 when 75% was masked. Specifically, the C-VAE was the best performing model in 2 cases, while the C-GAN was the most effective imputation model in 7 cases. However, C-GAN imputation appeared to have a less adverse impact on the affect detection models (Figure 3). From this perspective, the C-GAN was the best performing imputation approach for 12 of the 15 total evaluations, including all of the 25% missingness level. The C-VAE was the best model for only 2 of the 15 evaluations in terms of adverse impact on the affect models, both occurring with the *surprised* affective state. In total, data imputation with generative modeling had the least adverse impact on the affect detectors' performance in 13 of the 15 cases.

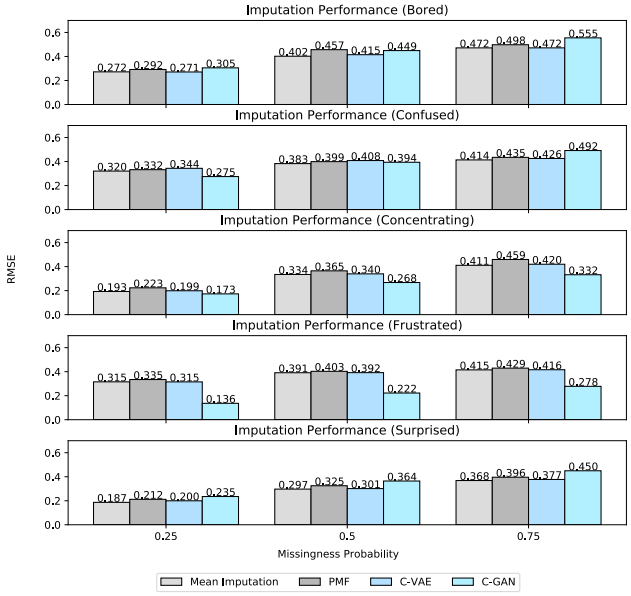

Figure 2: Imputation performance for the interaction log modality (Lower is better)

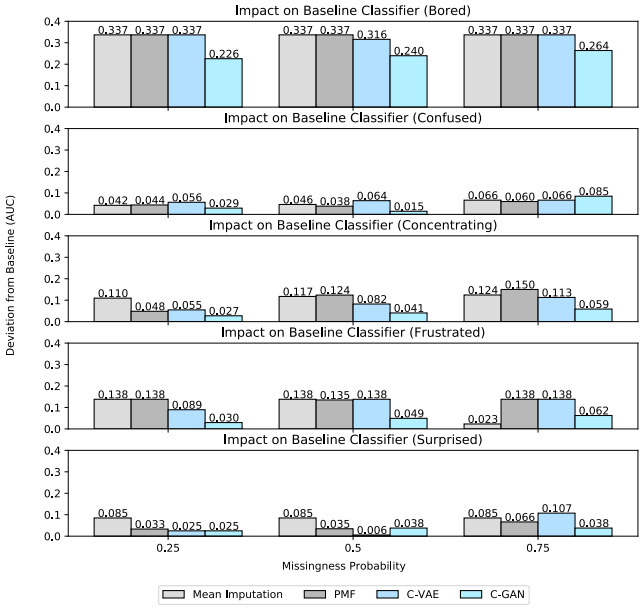

Figure 3: Impact of data imputation on affect models for interaction log modality (Lower is better)

5.3 Posture Data Imputation

In a similar manner, the same four imputation models were evaluated using the posture data as the masked modality, while the corresponding interaction log data was used as the conditional modality for the generative models. The primary difference from the gameplay modality masking is that the posture data was masked using a uniform probability across the entirety of each student's sequence. For example, 25% of the posture data samples were selected to be masked, but that selection occurred with equal probability across all data samples. This was done to ensure realistic masking of the sensor-based modality by simulating intermittent issues that may occur throughout student interactions with an adaptive learning environment, such as sensor mistracking, noise, or reliability issues. The results of these

evaluations are based on the missing posture data, evaluated for the same 25%, 50%, and 75% levels of missingness, and they are shown in terms of imputation performance (Figure 4) and impact on affect detector performance (Figure 5).

Generative models outperformed the two baselines in terms of imputation RMSE for 4 of 5 cases with 25% and 50% masking, and all 5 cases with 75% masking. The C-VAE was the optimal data imputation method in terms of RMSE for 10 of the 15 evaluations, compared to only 3 for the C-GAN model. In terms of adverse impact on the affect models, the C-VAE was the optimal method in 6 cases, while the C-GAN was the optimal method for 7 cases. However, in several cases the best-performing generative model was matched by one or both of the baseline models, such as when 50% of the *surprised* data was masked (Figure 5).

6 Discussion

The findings suggest that multimodal conditional generative models outperform the two baseline data imputation methods in 60% (9 out of 15) of the interaction log masking evaluations across the three data missingness levels and five affective states with respect to data imputation RMSE. This is compared to generative modeling outperforming the baseline data imputation methods in 86.7% (13 out of 15) when using the posture data as the masked modality. However, in terms of mitigating the adverse impact of missing data on affect detection models' performance, generative models performed optimally for 86.7% (13 out of 15) of the total evaluations for interaction log masking, while posture masking resulted in generative models outperforming the baselines in 80% (12 out of 15) of the evaluations, indicating consistent performance across the two modalities. Although different imputation methods (e.g., C-VAE, C-GAN) yielded the best performance for different affective states and modalities, it should be noted that different generative models could be utilized depending on the affective state and modality in a run-time setting. For this reason, we focus on the performance of deep conditional generative models as a family rather than individual models.

The generative models offer several benefits that contribute toward their higher performance than baseline data imputation techniques. Because the generative models are conditioned on separate, concurrent modalities, it is possible to maintain a multimodal perspective during the modeling process, allowing the imputation to be based on both non-missing data from the masked modality as well as the other intact modality. Mean imputation only takes into account a current student's single feature, and the PMF model only focuses on a single modality during its imputation. Additionally, because the C-GAN and C-VAE are based on deep neural networks, they are well suited to extract and model complex patterns between the multimodal data that may otherwise be ignored or removed. By using a deep learning-based imputation approach, these underlying features are able to be partially or fully reproduced within the masked modality, which can prove beneficial to the predictive performance of affect detection models. This is a possible explanation as to why the

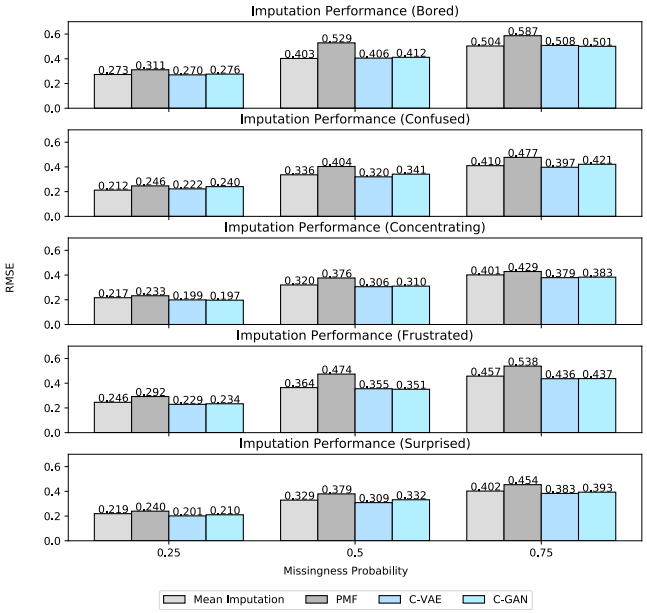

Figure 4: Imputation performance for the posture modality (Lower is better)

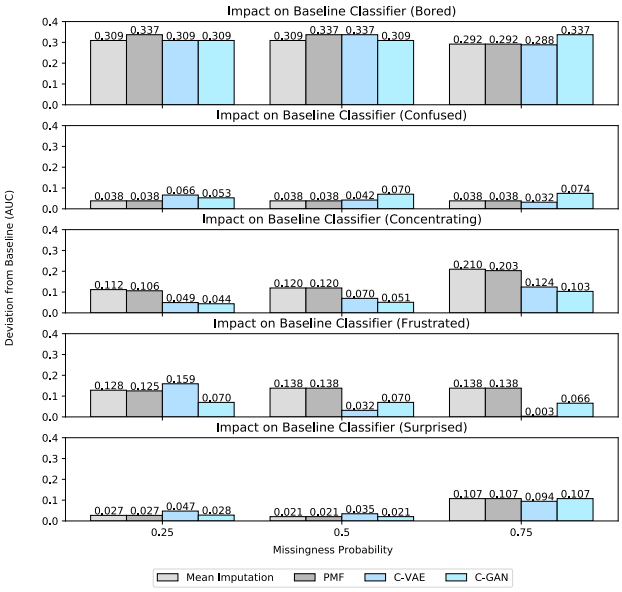

Figure 5: Impact on affect model performance for posture modality (Lower is better)

conditional generative models are the optimal imputation method for at least 80% of the affect models examined.

It is notable that mean imputation appears to produce similar RMSE values compared to the C-VAE and C-GAN models for imputing missing data. Many of the features of the interaction log data were reported using either a standard deviation or average number of certain gameplay actions. Because of the length of gameplay sessions (~1 hour each), these attributes may contain less variance than the posture-based data. Due to the averaging of

these features within mean imputation, it is possible that trends in the interaction log data that may be of use to the affect models are smoothed during the imputation stage. This may explain why the mean imputation produced a similar RMSE to the generative models when the interaction log modality is masked.

It is notable that the C-GAN was the most frequent optimal imputation method in terms of RMSE for the interaction log data masking, while the C-VAE was the most frequent for the posture data. We observe that the average difference between the C-GAN and C-VAE's RMSE is 0.076 for masked interaction logs, and 0.010 for masked posture data, indicating that while the C-VAE may have outperformed the C-GAN more frequently for the posture-based evaluations, the margin between the two generative models was extremely slim. However, this does not appear to be the case with the interaction log masking, where the C-GAN outperformed the C-VAE (and baselines) by considerable margins.

While the two deep conditional generative methods showed improved data imputation performance in most cases, it should be noted that the two modeling techniques are inherently different: GANs are constructed for generative tasks through the adversarial setup of their architecture, whereas VAEs are primary intended for latent representation modeling by minimizing the loss defined with the Kullback–Leibler divergence and the reconstruction error. It is often a challenge to accurately contextualize or quantify GAN convergence or performance as a whole due to the competing situation between the generator and the discriminator. This motivates the need to extend the evaluation of data imputation techniques to consider adverse impacts on the affect models, which appears to provide additional support for the use of C-GANs as a multimodal data imputation method.

The performance of data imputation techniques appeared to be related to specific affective states. For example, during interaction log masking, the C-GAN produced the lowest RMSE for *frustrated* in each of the missingness levels but produced the highest RMSE for *surprised* in each of the missingness levels. This behavior was also observed when evaluating the variance of the affect models. This can be attributed to a number of factors, such as inherent data imbalances (particularly for *frustrated* and *surprised*), physical behavioral cues that are distinct for each affective state, and differently predictive features for each binary class label.

A decline in imputation performance is expected as the missingness level increases [19, 37], particularly for deep learning models that suffer from significantly reduced training data. While this behavior does occur for each of the affective states and modalities, the decline is not as drastic as might be expected given the size of the initial training data, step size of the masking (25% increments), and depth of the deep learning architectures, particularly the C-GAN. As the amount of intact data decreases for a certain modality, the generative models risk overfitting to the training data or generating random noise as output. However, this issue is partially mitigated through the conditional input to each of the generative models. The inclusion of the condition as input allows the discriminator to be provided with additional training data, consisting of 1) real, non-masked data with corresponding conditional data, and 2) fake, synthetic data generated with randomly selected conditional data. This allows for the generator to be further refined during the training phase and provides further support for the use of C-GANs within our multimodal data imputation framework.

7 Conclusion

Multimodal student affect detection has shown significant promise for adaptive learning environments. However, both sensor-based and interaction-based data streams are often plagued by noisy or missing data, which significantly impedes the performance of affect detection models due to insufficient training data. To address this issue, we present a multimodal data imputation framework based on deep conditional generative models to support student affect detection in adaptive learning environments. Results indicate that conditional variational autoencoders (C-VAEs) and conditional generative adversarial networks (C-GANs) show significant promise for imputing sensor- and interaction log-based modalities for affect modeling. The generative models were compared against two common imputation methods, mean imputation and probabilistic matrix factorization (PMF), and outperformed these baseline models across a range of learning-centered affective states and levels of data missingness. Overall, deep conditional generative models show significant promise for their capacity to model patterns of data that are important for multimodal affect detection in adaptive learning environments and serve as an improved alternative to mean imputation and PMF. Conditional generative models impute data using input from separate, intact modalities, which allows the imputation process to maintain a multimodal perspective during data imputation in contrast to the baseline methods.

There are several promising directions for future work. Investigating additional types of data corruption, including artificial noise injection, within the multimodal data imputation framework is an important future extension to this work. Examining feature-level masking, data sample-level masking, or entirely missing modalities should also be considered. Conditional generative models for multimodal data imputation should also be investigated with additional sensor-based modalities that are commonly used in adaptive learning environments, such as facial expression, eye tracking, and physiological data. One area of interest is the use of multiple modalities as the conditioning factors for each generative model, and conversely, the imputation of multiple modalities using a single conditioning modality. Finally, additional generative model architectures should be explored, including auxiliary classifier GANs (AC-GANs), stacked VAEs, and Wasserstein-based GANs (W-GANs) to investigate their impact on improving student affect detection.

ACKNOWLEDGMENTS

The research was supported by the U.S. Army Research Laboratory under cooperative agreement #W911NF-13-2-0008. Any opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the U.S. Army.

REFERENCES

[1] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. 2017. Generalization and equilibrium in generative adversarial nets (GANs). In Proceedings of the 34th International Conference on Machine Learning (ICML). 322–349.

[2] Ryan S. Baker, Jaclyn Ocumpaugh, and Rafael Calvo. 2015. Interaction-Based Affect Detection in Educational Software. In The Oxford Handbook of Affective Computing, Rafael Calvo, Sidney D'Mello, Jonathan Gratch and Arvid Kappas (eds.). Oxford University Press, 233–245.

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. 41, 2 (2018). 423–443.

[4] Nigel Bosch, Huili Chen, Ryan Baker, Valerie Shute, and Sidney D'Mello. 2015. Accuracy vs. Availability heuristic in multimodal affect detection in the wild. In Proceedings of the 17th ACM International Conference on Multimodal Interaction. 267–274.

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

[6] Anthony F. Botelho, Ryan S. Baker, and Neil T. Heffernan. 2017. Improving sensor-free affect detection using deep learning. In Proceedings of the International Conference on Artificial Intelligence in Education. Springer, Cham, 40–51.

[7] Christopher Bowles, Liang Chen, Ricardo Guerrero, Paul Bentley, Roger Gunn, Alexander Hammers, David Alexander Dickie, Maria Valdés Hernández, Joanna Wardlaw, and Daniel Rueckert. 2018. GAN augmentation: Augmenting training data using generative adversarial networks. arXiv preprint arXiv:1810.10863.

[8] Cheng Chang, Lei Chen, Cheng Zhang, 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.

[9] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, (2002). 321–357.

[10] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 1 (1960). 37–46.

[11] Sidney D'Mello. 2013. A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. J. Educ. Psychol. 105, 4 (2013). 1082–1099.

[12] Sidney D'Mello and Jacqueline Kory. 2014. A review and meta-analysis of multimodal affect detection systems. ACM Comput. Surv. 47, 3 (2014). 43-79.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Advances in Neural Information Processing Systems. 2670–2680.

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

[15] Joseph Grafsgaard, Joseph Wiggins, Alexandria Vail, Kristy Elizabeth Boyer, Eric Wiebe, and James Lester. 2014. The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. In Proceedings of the Sixteenth ACM International Conference on Multimodal Interaction. ACM, 42–49.

[16] Nathan Henderson, Andrew Emerson, Jonathan Rowe, and James Lester. 2019. Improving sensor-based affect detection with multimodal data imputation. In Proceedings of the 8th International Conference on Affective Computing & Intelligent Interaction. 669–675.

[17] Nathan Henderson, Jonathan Rowe, Bradford Mott, Keith Brawner, Ryan S. Baker, and James Lester. 2019. 4D affect detection : Improving frustration detection in game-based learning with posture-based temporal data fusion. In Proceedings of The 20th International Conference on Artificial Intelligence in Education. 144–156.

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

[19] Natasha Jaques, Sara Taylor, Akane Sano, and Rosalind Picard. 2017. Multimodal autoencoder : A deep learning approach to filling in missing sensor data and enabling better mood prediction. In Proceedings of the Seventh International Conference on Affective Computing and Intelligent Interaction. 202–208.

[20] Min Kyu Lee, Dong Yoon Choi, and Byung Cheol Song. 2019. Facial expression recognition via relation-based conditional generative adversarial network. In Proceedings of the 21st International Conference on Multimodal Interaction. 35–39.

[21] Zuozhu Liu, Wenyu Zhang, Shaowei Lin, and Tony Q.S. Quek. 2017. Heterogeneous sensor data fusion by deep multimodal encoding. IEEE J. Sel. Top. Signal Process. 11, 3 (2017). 479–491.

[22] Shuang Ma, Jianlong Fu, Chang Wen Chen, and Tao Mei. 2018. DA-GAN: Instance-level image translation by deep attention generative adversarial networks. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 5657–5666.

[23] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. In arXiv preprint arXiv:1411.1784.

[24] Jaclyn Ocumpaugh, Ryan S. Baker, and Mercedes T. Rodrigo. 2015. Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual.

[25] Jennifer L. Sabourin, Jonathan P. Rowe, Bradford W. Mott, and James C. Lester. 2013. Considering alternate futures to classify off-task behavior as emotion self-regulation: A supervised learning approach. J. Educ. Data Min. 5, 1 (2013). 9–38.

[26] Ruslan Salakhutdinov and Andriy Mnih. 2008. Probabilistic matrix factorization. Adv. Neural Inf. Process. Syst. (2008). 1257–1264.

[27] Chao Shang, Aaron Palmer, Jiangwen Sun, Ko Shin Chen, Jin Lu, and Jinbo Bi. 2017. VIGAN: Missing view imputation with generative adversarial networks. In Proceedings of the IEEE International Conference on Big Data. 766–775.

[28] Kihyuk Sohn, Xinchen Yan, and Honglak Lee. 2015. Learning structured output representation using deep conditional generative models. Adv. Neural Inf. Process. Syst. (2015). 3483–3491.

[29] Robert A. Sottilare, Ryan S. Baker, Arthur C. Graesser, and James C. Lester. 2018. Special issue on the Generalized Intelligent Framework for Tutoring (GIFT): Creating a stable and flexible platform for innovations in AIED Research. Int. J. Artif. Intell. Educ. 28, 2 (2018).

[30] Randall Spain, Jonathan Rowe, Benjamin Goldberg, Robert Pokorny, and James Lester. 2019. Enhancing learning outcomes through adaptive remediation with GIFT. In Proceedings of the Interservice/Industry Training, Simulation and Education Conference. 1–11.

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

[32] Kim-Han Thung, Pew-Thian Yap, and Dinggang Shen. 2017. Multi-stage diagnosis of Alzheimer's disease with incomplete multimodal data via multi-task deep learning. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. 160–168.

[33] Marcelo Worsley, Stefan Scherer, Louis Philippe Morency, and Paulo Blikstein. 2015. Exploring behavior representation for learning analytics. In Proceedings of the 17th International Conference on Multimodal Interaction. 251–258.

[34] Suowei Wu, Zhengyin Du, Weixin Li, Di Huang, and Yunhong Wang. 2019. Continuous emotion recognition in videos by fusing facial expression, head pose and eye gaze. In Proceedings of the 20th International Conference on Multimodal Interaction. 40–48.

[35] Jianfei Yang and Kai Wang. 2018. Deep recurrent multi-instance learning with spatio-temporal features for engagement intensity prediction. In Proceedings of the 2018 International Conference on Multimodal Interaction. ACM, 594–598.

[36] Xi Yang, Yeo Jin Kim, Michelle Taub, Roger Azevedo, and Min Chi. 2020. PRIME: Block-wise missingness handling for multi-modalities in intelligent tutoring systems. In Proceedings of The International Conference on Multimedia Modeling. 63–75.

[37] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. 2018. GAIN: Missing data imputation using generative adversarial nets. In Proceedings of the 35th International Conference on Machine Learning.

[38] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. 2017. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision. 5907–5915.