Improving Sensor-Based Affect Detection with Multimodal Data Imputation

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

Andrew Emerson
Department of Computer Science
North Carolina State University
Raleigh, NC, USA
ajemerso@ncsu.edu

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

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

***Abstract*— Utilizing sensors for affect detection in adaptive learning technologies has been the subject of growing interest in recent years. This extends to the collection of multiple concurrent sensor-based input channels to enable multimodal affective modeling. However, sensors pose significant challenges to affect detection, including sensor connectivity issues, background noise, inconsistent data logging, and loss of data due to hardware failure. In this paper, we introduce a framework for multimodal data imputation to improve automated detection of student affect in adaptive learning technologies. Through the use of an autoencoder neural network trained on Microsoft Kinect-based posture data and electrodermal activity data with synthetic noise injection, we approximate missing values within the original dataset while still preserving the inter-related context between features when reconstructing the dataset. The reconstructed dataset can be used in conjunction with multimodal data fusion techniques to further boost affect detector accuracy. Results indicate that this framework improves the effectiveness of multimodal affect detectors when compared to unimodal baseline models, as well as models using baseline data imputation techniques such as mean imputation. Further, it maintains cross-modality information that influences the multimodal affect detectors' performance, as the approach also outperforms previous work using the latent representation of the imputed dataset as training data instead of a complete reconstruction of the original dataset's dimensionality.**

Keywords—data imputation, deep learning, affect detection, multimodal analytics, data fusion

I. INTRODUCTION

Affect plays a pivotal role in learning [1]. Affective states such as boredom have been associated with poorer learning outcomes [2], while states such as confusion and engagement have been associated with positive learning outcomes [3]. Other affective states, such as frustration, have a complex relationship with learning, as they have been shown to have a correlation with both negative and positive learning outcomes in different contexts [3], [4]. Affective models are critical to adaptive learning technologies [5] due to their role in detecting and intervening in student emotional states to enhance learning and engagement [6], [7].

Recent years have seen increased interest in multimodal affect detection within adaptive learning technologies. The inclusion of multiple independent data streams has been shown to provide a significant boost in affect detector performance [8], as well as additional insight into a student's tendencies and behavior when engaged with a learning environment [9]. Multimodal machine learning models are of particular interest because of their ability to emulate human perception of emotion based on multiple concurrent modalities (e.g., visual, auditory, tactile) [10]. As a result of increased interest in multimodal affect detection, multimodal machine learning techniques have been applied to a wide range of tasks including detection of stress [11], anger [12], engagement [13], and biometric features [14].

Multimodal affect detection frequently involves the deployment of multiple physical hardware sensors, each designed to capture a distinct modality. For example, recent work on multimodal sensor-based affect detection has captured modalities such as facial expression, eye tracking, posture, gesture, speech, electrodermal activity (EDA), and electroencephalography (EEG) [7], [13], [15]. Sensor-based methods have shown significant promise for affect detection in adaptive learning technologies because of their potential for generalizability across domains and learning environments and because they need not rely upon domain-specific feature representations. Additionally, sensors can provide a relatively inexpensive alternative to more costly input sources because many sensors do not require specialized hardware support, as they use built-in webcams, eye trackers, microphones, and motion-tracking cameras.

However, there are inherent challenges with sensor-based systems [7]. Sensors can frequently experience issues such as poor calibration, mistracking, background noise, inconsistent behavior, and loss of data due to storage or transfer constraints. These problems are exacerbated when the number of concurrent sensors is increased, since the proportion of incomplete data samples increases with intermittent sensor failures [16]. Often, a sensor can malfunction for an extended period of time, producing large volumes of data that contain significant noise or that are missing altogether.

To address these challenges, we propose a novel framework for handling missing data in multimodal affect detection. We investigate this framework in the context of student affect detection in a game-based learning environment for emergency medical training, TC3Sim. We investigate multimodal affect detection using posture-tracking data and EDA data, where the EDA data is missing from approximately half of the raw dataset. We train an

971-1-7281-3888-6/19/$31.00 ©2019 IEEE

autoencoder neural network with the subset of data containing all modalities. The autoencoder is trained to reconstruct the original dataset using artificial noise injection, thus simulating the missing modalities, and then it is used to impute the missing EDA values. We also investigate the use of the encoder portion of the autoencoder as an alternative to dimensionality reduction during feature extraction. Afterward, we identify the highest-performing classifier for each of five affective states, comparing several classifiers trained on the reconstructed multimodal dataset using alternate data fusion techniques. Results indicate that autoencoder-based data reconstruction outperforms other data imputation methods based on classifier performance, and multimodal affect detection yields improved classifier performance compared to unimodal affect detection.

II. Related Work

Factors such as the decreasing cost of sensors, hardware flexibility, and support for multimodal systems have led to increased interest in sensor-based affect detection. Pei et al. utilize LSTM recurrent neural networks to perform multimodal affect detection on a dataset containing both audio and video-based modalities [15]. Patwardhan et al. calculate spatiotemporal features from Kinect posture data to perform similar affect classification tasks using a hybrid model of supervised and rule-based learning [17]. In a similar fashion, Grafsgaard et al. explore the use of Kinect data to determine user engagement in an adaptive tutoring system for teaching introductory programming concepts [18]. The association between affective states, such as frustration and engagement, and learning outcomes was explored in prior work by Grafsgaard et al., using facial expression recognition algorithms applied to Kinect posture and gesture data [19]. DeFalco et al. utilize posture-based Kinect data for the development of affect detection systems for boredom, confusion, engagement, frustration, and surprise [7].

Additional multimodal systems have been constructed to take advantage of biosignal modalities such as electroencephalogram data [20], electrodermal activity [21], and blood pulse volume (BVP) [11]. Harley et al. explore the relationship between facial expressions, EDA, and 19 different self-reported affective states for each user engaged with MetaTutor, an adaptive hypermedia-based learning environment [21]. The use of EEG, EDA, electromyographic (EMG) signals, and various other biosignals have been used with multimodal machine learning for the detection of low and high valence and arousal in subjects watching a series of videos [22].

There has been limited work on multimodal analytics that has addressed the issue of missing sensor data through the use of autoencoders. One exception is recent work that used a version of a denoising autoencoder [23] to learn a latent representation of artificially noisy data [16]. This can be used to represent missing data in a compact representation for the purpose of improving classification accuracy. In our work, we seek to actually reconstruct the missing data, as opposed to learning an encoded representation. Previous approaches have also focused on multimodal data generation using text and images, such as in [24] and [25]. However, that work did not use sensor-based multimodal data streams. There has been prior work investigating imputation of missing sensor data using various neural network-based fusion techniques [26]. One area where we improve upon prior approaches is by handling blocks of missing values, such as when a sensor is unavailable for a period of time. This is in contrast to imputing missing values that are sparsely distributed. Other communities, such as the medical community, often need to handle missing Electronic Health Record data, but typically use latent representations for classification rather than using deep neural architectures for imputation [27].

Fig. 1. Screenshot of injured soldier in the *TC3Sim* game-based learning environment.

III. Dataset

Our investigation into multimodal affect detection utilizes a game-based learning environment focused on training military medical personnel, TC3Sim. This simulation environment was developed by *Engineering and Computer Simulations (ECS)* and is frequently deployed by the U.S. Army to provide realistic training simulations of combat medic scenarios. Within the game, users assume the first-person role of a combat medic alongside various computer-generated non-player characters (NPCs). The story-driven scenarios feature a series of combat-based simulations with the end result being injuries received by one or more NPCs. Participants are tasked with executing a number of tactical combat and medical tasks including securing the working perimeter, applying the correct treatment to the appropriate victims, and preparing for eventual evacuation. This work derives the primary dataset from sensor data corresponding to student interactions with four separate training scenarios from TC3Sim: a tutorial scenario, leg injury scenario, a patrol scenario involving an IED attack, and a scenario where the patient expires regardless of treatment received. A screenshot of the user's perspective when engaging with an injured NPC is shown in Fig. 1.

The dataset used to develop affect detectors was collected through a previous study consisting of observations of 119 students (83% male, 17% female) as they engaged with TC3Sim. TC3Sim was deployed using the Generalized Intelligent Framework for Tutoring (GIFT), which is an open-source software framework designed for building and deploying adaptive training systems. Each participant in the study worked at a single workstation, with the session lasting approximately one hour per user. The Microsoft Kinect for Windows 1.0 sensor was used to capture the posture of each individual throughout the duration of the training session. The Kinect was positioned to face directly at the individual, while capturing all head and body movements at a sampling rate of 10-12 Hz. The data underwent a filtering process within GIFT before being exported for external processing. Additionally,

each user wore an Affectiva Q-Sensor bracelet, which captured timestamped electrodermal activity, as well as the acceleration vectors for the sensor. Acceleration data was not utilized in this effort.

Ground truth labels of student affect were collected by two trained observers using a quantitative observation protocol called BROMP [28]. Each observer walked around the classroom, routinely observed each participant, and marked instances of affective behavior in 20-second intervals. The inter-annotator agreement between the two BROMP observers had a Cohen's Kappa that was higher than 0.6.

Seven affective states were observed during this time: *engaged, confused, bored, surprised, frustrated, contempt,* and *other*. The resulting dataset consisted of 3,066 BROMP observations between the two observers. Any observation where there was disagreement between the two observers was removed from the dataset, leaving 755 BROMP observations during the subset of time that students used TC3Sim during the study. A total of 435 observations of *engagement* were recorded, with 174 instances of *confused*, 73 instances of *boredom*, 32 instances of *frustration*, 29 instances of *surprise*, and 12 as *contempt*. Due to the small number of observed instances of *contempt* and *other*, these affective states are not considered in this study.

The Kinect sensor tracked and recorded data for 91 vertices, of which we selected three based on prior work investigating Kinect-based affect detection [18]: *top_skull*, *center_shoulder*, and *head*. A total of 73 posture features were distilled from the Kinect vertex data providing a summary of the posture of the student, with the mean, variance, and standard deviation calculated over time windows of 5, 10, and 20 seconds preceding the BROMP observation. The Q-Sensor returned data consisting of a timestamp and an EDA reading. In a similar fashion to the Kinect posture data, summary statistics were calculated for the EDA modality including attributes such as *min_eda, max_eda, variance_eda,* and *median_eda*. These statistics were also calculated across time windows of 5, 10, and 20 seconds prior to the BROMP observation. Additionally, the net change in the EDA readings across time windows of 3 and 20 seconds was calculated, resulting in 18 EDA features.

During the data collection process, the Q-Sensor experienced a significant amount of inconsistent behavior. This resulted in the loss of data for varying durations, and occasionally the EDA modality was lost for entire sessions. Out of the 755 data samples in our dataset, 333 instances were shown to have missing EDA data. The Kinect modality did not appear to suffer from significant data loss. To train our autoencoder, we derived a complete dataset from the original "raw" dataset by removing the 333 incomplete data samples, leaving a subset of 422 data samples containing both Kinect-based posture data as well as the associated EDA readings.

IV. METHODOLOGY

In this work, we handled missing data by employing a specialized variation of a denoising autoencoder that is based on Multimodal Autoencoders (MMAE) [16] [23]. The model used in this process involves feature-level fusion of the multiple modalities, which are then used to train an autoencoder neural network. The model is trained to reconstruct the original dataset by converting the dataset to a latent representation following artificial noise injection on select modalities. In the MMAE approach, an autoencoder is trained on artificially noisy data where all modalities are present. We begin by taking the complete set of data where all modalities are present, and we then normalize all features to be in the range [0,1]. Before passing the complete dataset through the autoencoder, each observation is injected with two types of noise: a simple masking noise and a complete removal of one or more modalities. For the simple masking noise, 5% of the features for the observation are randomly selected and these values are set to 0. For the removal of the modalities, this is performed by randomly selecting one or more modalities—for this work, we select one of either the Kinect or EDA modality—and setting each feature within this modality to -1. The MMAE is then trained to reconstruct the original data by using this compromised dataset. This process enables the autoencoder to more accurately reproduce a full multimodal dataset when faced with missing or invalid data within certain modalities.

Once the MMAE has been trained on the complete data where all modalities are present, we can feed the full dataset, including missing data, to the autoencoder. The output of the autoencoder network includes imputed values for all observations. To be consistent with the training data, all features of the full dataset are normalized to be in the range [0,1]. The values that are missing or invalid, which can include blank cells, unique identifiers, and other representations, are all set to -1. We feed the full dataset through the MMAE network. Instead of using the latent layer of the autoencoder as the input to classification algorithms, as in [16], we propagate the full set of observations through the trained network, including the decoder portion of the MMAE, resulting in output of the same dimensionality as the input. The output then contains imputed values for all original values. We take advantage of having the original values where data was not missing, and we replace the corresponding imputed values with their original values. To emphasize this, we end the imputation process with our original full dataset where missing values are replaced with imputed values produced from the trained autoencoder. A visual representation of this process is found in Fig. 2, for a dataset containing *m* attributes and *n* data samples.

This approach confers a significant advantage over using the latent representation of the data: it yields an interpretable set of features. Performing feature selection on this imputed data can then generate a more human-understandable set of features, as opposed to the result of the latent autoencoder representation. In addition, it also affords the ability to use other dimensionality reduction techniques such as PCA or even another neural network architecture to find more compact representations of the data.

The autoencoder used to perform this imputation has a single hidden layer with 30 nodes, which performed well in terms of input reconstruction. The layers within the autoencoder utilized a sigmoidal activation function due to the normalization of the training data. We trained the model using the ADADELTA optimization method [29] for 5,000 epochs, using mean squared error as the cost function. The training and forward propagation of the missing data was performed with the Keras deep learning toolkit with a TensorFlow backend.

Once the missing data has been imputed, we still have the problem of class imbalance in our dataset. Thus, for each affective state, we take the input data with imputed values and then oversample the minority class. We oversample each minority class observation by cloning at a rate equivalent to the ratio of the majority to minority class labels for that

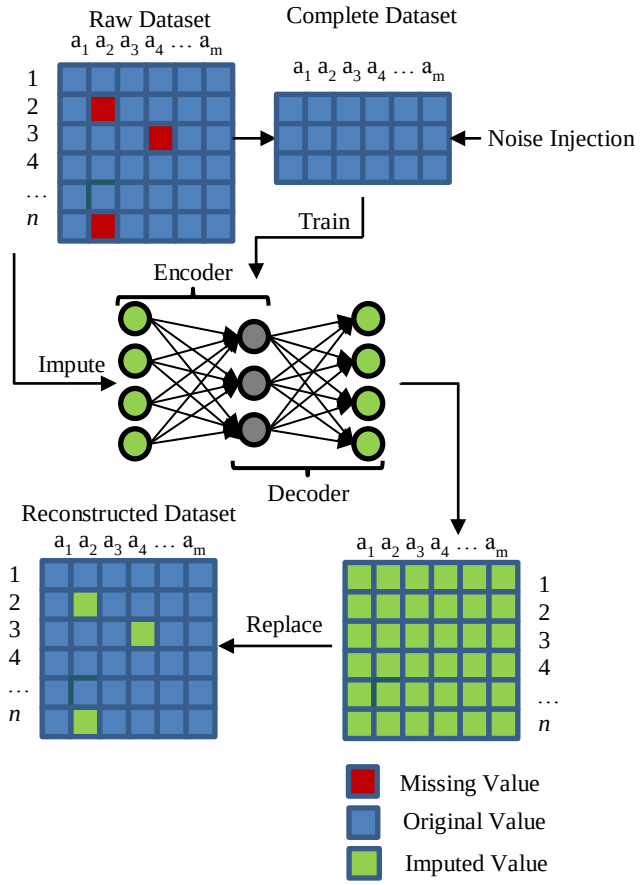

Fig. 2. Multimodal data reconstruction process.

affective state, resulting in five different oversampled datasets corresponding to the five affective states.

Next, we establish a data processing pipeline with the objective of investigating a set of classifiers to determine the optimal affect detection model for each individual affective state. Because of the high number of available features within our multimodal dataset, we employ principal component analysis (PCA) to reduce the number of dimensions in the classification task. PCA helps remove noise from a high dimensionality relative to the size of the data, and it accounts for potential multicollinearity among features. For fair comparison in our experiments, we transform the data with PCA using the same number of components that the latent layer of the autoencoder has in terms of dimensionality. These new orthogonal features will be comparable to the latent features produced by the encoder portion of the autoencoder.

To determine the optimal classifier for each affective state, we investigate five different model types: support vector machine (SVM), J48 decision tree, JRip propositional rule learner, logistic regression, and deep neural network. Each classifier was trained using student-level 10-fold cross-validation, meaning that data from a single student session is never split across both the training and test sets, which could lead to positively biased results. Oversampled positive instances of emotions were also removed from the test set, to avoid inflated results as well.

After the optimal classifier for each affective state was determined, we evaluated different types of multimodal data fusion to determine if feature-level fusion or decision-level fusion boosted the performance of the classifier. We evaluated three data fusion variations. Fig. 3 shows a visual representation of the three data fusion techniques. Early Fusion 1 involves concatenating the posture and EDA features prior to the PCA dimensionality reduction and

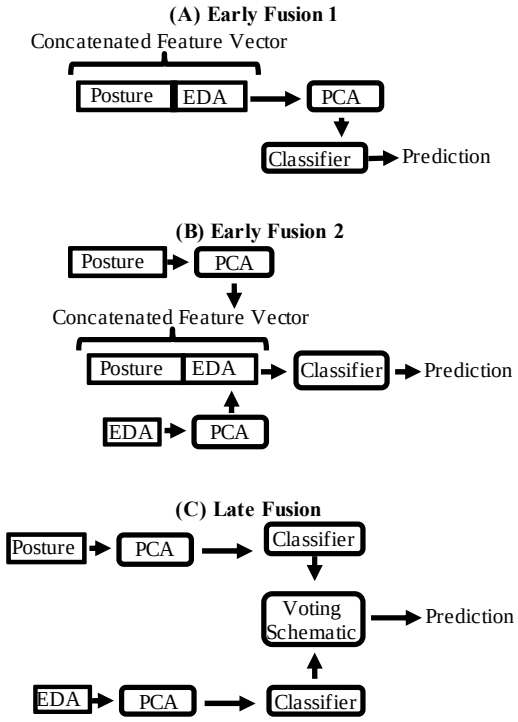

Fig. 3. Data pipeline for data fusion variations.

training a single classifier on a dataset containing *n* attributes. Early Fusion 2 uses PCA dimensionality reduction on each individual modality, producing two individual datasets, each consisting of *n/2* attributes. These attributes are then concatenated in a similar fashion to Early Fusion 1, and subsequently used to train a single classifier.

Late Fusion involves performing PCA dimensionality reduction on two separate modalities with each resulting data channel containing *n* attributes each. Each data channel is then used to train two separate unimodal classifiers. The output of each classifier is a two-element confidence vector representing whether a certain data sample contains a positive or negative instance of the target affective state. A voting schematic is then used to determine the overall representative prediction of the data fusion system. We experiment with two different voting schematics: highest confidence level, and highest average confidence level. For the former, the class with the highest confidence level is selected as the prediction. For the latter, the confidence levels for each class are averaged across each classifier, and the class with the highest average is selected as the prediction. This data pipeline was implemented and evaluated using RapidMiner 9.0 [32], while the data filtering and distillation, noise injection, and data imputation were performed using Python 3.

Finally, we evaluate classifiers trained on the encoded data produced by the encoder portion of the autoencoder [16] against the classifiers trained on the decoded, reconstructed data. Our final results from the multimodal classifiers are then compared with unimodal classifiers trained solely on the posture data to determine whether the addition of the EDA modality through data imputation improved affect detector performance.

V. Results and Discussion

We compare our method of data imputation to mean imputation, a commonly used approach that imputes missing data points using the mean of the available data for a given feature [30]. We train two separate support vector machines,

one on a reconstructed dataset using our autoencoder, and the other trained on a dataset completed using mean imputation. Cohen's Kappa [31] is used as our primary evaluation metric for classifier performance due to its ability to determine a classifier's ability to perform at a higher success rate than chance. The multimodal dataset is comprised of vectors that contain posture and EDA data concatenated at a feature level. Table I shows a comparison of Kappa values for each trial with each affective state, with the best imputation method for each classifier shown in bold. Results indicate that autoencoder-based data imputation yields higher-performing classifiers than mean imputation across all five affective states. The results of the best classifier selected for each affective state are shown in Table II, with Cohen's Kappa, Area Under Curve (AUC), accuracy, and F1 Scores shown. The approach to handling the multimodal data in this experiment was Early Fusion 1. The SVM achieved the highest classification performance for three affective states: *bored*, *confused*, and *surprised*. JRip and J48 achieved the highest performance for *frustrated* and *engaged*, respectively.

Logistic regression performed relatively well for two affective states (*confused* and *bored)* but did not achieve the highest performance for any affective state. Notably, deep neural networks performed less effectively than the best classifier for each category and yielded poor results for a few affective states as well. This can possibly be attributed to an insufficient amount of training data, as well as overfitting of the autoencoder or the classifier itself.

Following this procedure, we used each affective state's top-performing classifier to evaluate Early Fusion 2 and Late Fusion. Additionally, we evaluated Late Fusion based on two voting schematics: highest confidence (HC) and highest average confidence (HAC). Table III displays the results of Early Fusion 2 and both variations of Late Fusion.

The results in Table III indicate that variations of data fusion do not improve the results of the classifier for any of the affective states, and in several cases, the results were significantly worse. One explanation for the relatively poor performance of Early Fusion 2 is that this method forces an even balance of attributes across modalities used to train the classifier. While PCA in Early Fusion 1 is able to select its own ratio of 30 principle components from the posture and EDA modalities to comprise the 30 attributes for the classifier, Early Fusion 2 forces each PCA algorithm to select exactly 15 attributes per modality. Thus, if a modality such as the EDA data is inherently less informative than other modalities, Early Fusion 2 is replacing potentially useful attributes with less helpful attributes, resulting in lower performances across the classifiers.

Previous work has found that EDA data does not have a tightly-coupled relationship with various affective states, as compared to other modalities such as facial expression [21]. It is also a possibility that the EDA modality does not contain enough variance across multiple instances of each affective state for each classifier to distinguish between them effectively. This problem is amplified during examples of mild or suppressed expressions of affective states. Additionally, modalities such as the Kinect posture data inherently contain higher dimensionality than the EDA data and therefore potentially contain more distinguishing factors between affective states. Data fusion methods such as Early Fusion 2 and Late Fusion embrace an equal emphasis on all modalities present, which likely led to a tradeoff between informative Kinect features and less informative EDA features that adversely impacted classifier performance for those two data fusion techniques. However, the EDA modality did appear to contain useful contextual information that mostly improved classifier performance when used in conjunction with the Kinect posture modality.

To determine whether the addition of the EDA modality was indeed beneficial to the performance of each classifier, we trained a unimodal classifier on the complete posture data

TABLE I. Comparison between mean imputation and autoencoder imputation for classifying student affective states.

Affective State	Mean Imputation	Autoencoder
Bored	0.087	**0.184**
Confused	0.068	**0.107**
Engaged	0.029	**0.037**
Frustrated	0.023	**0.049**
Surprised	0.019	**0.020**

TABLE II. Results for best-performing classifier for each affective state using Early Fusion 1.

Bored				
Classifier	**Kappa**	**AUC**	**Accuracy**	**F1 Score**
SVM	**0.1100**	0.6160	0.6897	0.2350
Confused				
Classifier	**Kappa**	**AUC**	**Accuracy**	**F1 Score**
SVM	**0.1340**	0.6210	0.6398	0.3685
Engaged				
Classifier	**Kappa**	**AUC**	**Accuracy**	**F1 Score**
J48	**0.1460**	0.5650	0.5799	0.6014
Frustrated				
Classifier	**Kappa**	**AUC**	**Accuracy**	**F1 Score**
JRip	**0.078**	0.5550	0.9174	0.1389
Surprised				
Classifier	**Kappa**	**AUC**	**Accuracy**	**F1 Score**
SVM	**0.154**	0.5000	0.7007	0.2736

TABLE III. Comparison of multimodal data fusion techniques with best performing classifiers for each affective state.

Bored	
Fusion Method (SVM)	**Kappa**
Early Fusion 1	**0.1100**
Early Fusion 2	0.0650
Late Fusion (HC)	0.0960
Late Fusion (HAC)	0.1059
Confused	
Fusion Method (SVM)	**Kappa**
Early Fusion 1	**0.1340**
Early Fusion 2	0.0730
Late Fusion (HC)	0.02932
Late Fusion (HAC)	0.02932
Engaged	
Fusion Method (J48)	**Kappa**
Early Fusion 1	**0.1460**
Early Fusion 2	-0.020
Late Fusion (HC)	0.0651
Late Fusion (HAC)	0.0651
Frustrated	
Fusion Method (JRip)	**Kappa**
Early Fusion 1	**0.0780**
Early Fusion 2	0.0070
Late Fusion (HC)	0.0186
Late Fusion (HAC)	0.0259
Surprised	
Fusion Method (SVM)	**Kappa**
Early Fusion 1	**0.1540**
Early Fusion 2	0.0170
Late Fusion (HC)	-0.0172
Late Fusion (HAC)	-0.0136

only and used the classifiers' performance as a baseline for each affective state. The baselines and best results from the multimodal approach for each affective state (Early Fusion 1) are shown in Table IV. The addition of the partially imputed EDA modality improved classifier performance on all affective states with the lone exception of *boredom*. However, a significant majority of results indicate that multimodal data imputation for affect detection is beneficial relative to unimodal classification techniques.

Prior research demonstrated the effectiveness of using the encoded latent feature vectors produced by an autoencoder to train a classifier [16]. We compare this approach to our approach of reconstructing the original dataset using decoding of latent representations. After producing a reconstructed dataset, we replace any values that have associated existing values in the original dataset. This process ensures that only the values determined to be missing or invalid are imputed, and values that existed in the original dataset are not overwritten with imputed values. Upon completion of this process, we train the same selected classifier model for each affective state on two variations of data: the encoded latent representations, and the decoded, reconstructed data. The comparison of each classifier's performance on the encoded and decoded data is shown in Table V.

The performance of the classifiers trained on the reconstructed dataset lead the classifier to achieve higher performance for every affective state. A possible explanation includes the preservation of original values after the data reconstruction. This ensures that the dataset contains the original underlying, complex relationships between multiple attributes, which often is an important aspect of multimodal machine learning [10]. This problem extends to the encoded dataset, as reducing the dimensionality through the latent representation contains the inherent risk of losing contextual information that may affect the performance of a classifier.

VI. CONCLUSION

Missing data is a persistent problem in sensor-based computational systems, particularly in affect detection for adaptive learning technologies. Given recent interest in multimodal affect detection, it is critical to devise effective methods for coping with situations where one or more modalities suffer from noisy or incomplete data. Removing incomplete data samples risks loss of important contextual information contained within inter-related modalities. Standard imputation methods, such as mean imputation, allow all data samples to be retained, but they only retain contextual information across a single feature.

We have introduced a multimodal data imputation framework that uses an autoencoder to capture contextual relationships across attributes spanning multiple modalities. We investigated the framework using Kinect-based posture tracking and Q-Sensor-based electrodermal activity data collected during student interactions with a game-based learning environment for emergency medical training. An empirical evaluation shows that the multimodal data imputation framework significantly improves the performance of multimodal sensor-based affect detection.

There are several promising directions for future work. Additional feature reduction and feature selection techniques should be explored to investigate their impact on classifier performance. The multimodal data imputation framework should be investigated across a broader range of modalities, including student facial expression and gesture, as well as additional datasets to evaluate the generalizability of our overall multimodal data pipeline. Finally, there is significant promise in investigating more sophisticated denoising techniques related to the framework's noise injection approach, which holds significant potential for further improving the performance of multimodal affect detectors.

TABLE IV. Comparison of Kinect-only unimodal vs. multimodal classifiers.

Bored (SVM)				
Classifier	**Kappa**	**AUC**	**Accuracy**	**F1 Score**
Unimodal	**0.1280**	0.6310	0.7817	0.2235
Multimodal	0.1100	0.6160	0.6897	0.2350
Confused (SVM)				
Classifier	**Kappa**	**AUC**	**Accuracy**	**F1 Score**
Unimodal	0.0280	0.5490	0.6151	0.1913
Multimodal	**0.1340**	0.6210	0.6398	0.3685
Engaged (J48)				
Classifier	**Kappa**	**AUC**	**Accuracy**	**F1 Score**
Unimodal	0.0480	0.5480	0.5496	0.6353
Multimodal	**0.0710**	0.5960	0.5774	0.6904
Frustrated (JRip)				
Classifier	**Kappa**	**AUC**	**Accuracy**	**F1 Score**
Unimodal	-0.001	0.4870	0.8783	0.0606
Multimodal	**0.0780**	0.5550	0.9174	0.0926
Surprised (SVM)				
Classifier	**Kappa**	**AUC**	**Accuracy**	**F1 Score**
Unimodal	-0.0210	0.4110	0.5913	0.0435
Multimodal	**0.1540**	0.5000	0.7007	0.2736

TABLE V. Comparison of decoded dataset vs encoded dataset on classifier performance.

Bored (SVM)				
Classifier	**Kappa**	**AUC**	**Accuracy**	**F1 Score**
Decoded	**0.1100**	0.6160	0.6897	0.2350
Encoded	0.093	0.649	0.6247	0.2270
Confused (SVM)				
Classifier	**Kappa**	**AUC**	**Accuracy**	**F1 Score**
Decoded	**0.1340**	0.6210	0.6398	0.3685
Encoded	0.0530	0.5540	0.5633	0.3072
Engaged (J48)				
Classifier	**Kappa**	**AUC**	**Accuracy**	**F1 Score**
Decoded	**0.1460**	0.5650	0.5799	0.6014
Encoded	-0.0200	0.492	0.5339	0.6587
Frustrated (JRip)				
Classifier	**Kappa**	**AUC**	**Accuracy**	**F1 Score**
Decoded	**0.0780**	0.5550	0.9174	0.1389
Encoded	-0.0250	0.4780	0.8750	0.0204
Surprised (SVM)				
Classifier	**Kappa**	**AUC**	**Accuracy**	**F1 Score**
Decoded	**0.1540**	0.5000	0.7007	0.2736
Encoded	0.0070	0.3940	0.6671	0.0543

ACKNOWLEDGMENTS

We wish to thank Dr. Jeanine DeFalco, Dr. Benjamin Goldberg, and Dr. Keith Brawner at the U.S. Army Combat Capabilities Development Command, Dr. Mike Matthews and COL James Ness at the U.S. Military Academy, Dr. Robert Sottilare at SoarTech, and Dr. Ryan Baker at the University of Pennsylvania for their assistance in facilitating this research. 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] S. D'Mello, "A selective meta-analysis on the relative incidence of discrete affective states during learning with technology," *J. Educ. Psychol.*, vol. 105, no. 4, pp. 1082–1099, 2013.

[2] S. Craig, A. Graesser, J. Sullins, and B. Gholson, "Affect and learning: An exploratory look into the role of affect in learning with AutoTutor," *J. Educ. Media*, vol. 29, no. 3, pp. 241–250, 2005.

[3] Z. Pardos, R. Baker, M. S. Pedro, S. M. Gowda, and S. M. Gowda, "Affective states and state tests: investigating how affect and engagement during the school year predict end-of-year learning outcomes," *J. Learn. Anal.*, vol. 1, no. 1, pp. 107–128, 2014.

[4] S. D'Mello and A. Graesser, "The half-life of cognitive-affective states during complex learning," *Cogn. Emot.*, vol. 25, no. 7, pp. 1299–1308, 2011.

[5] D. G. Cooper, I. Arroyo, and B. P. Woolf, "Actionable affective processing for automatic tutor interventions," in *New Perspectives on Affect and Learning Technologies*, New York, NY: Springer, 2011, pp. 127–140.

[6] J. F. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "Predicting learning and affect from multimodal data streams in task-oriented tutorial dialogue," in *Proceedings of the Seventh International Conference on Educational Data Mining*, 2014, pp. 122–129.

[7] J. A. DeFalco, J. P. Rowe, L. Paquette, V. Georgoulas-Sherry, K. Brawner, B. W. Mott, R. S. Baker, and J. C. Lester, "Detecting and addressing frustration in a serious game for military training," *Int. J. Artif. Intell. Educ.*, vol. 28, no. 2, pp. 152–193, 2018.

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

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

[10] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2018.

[11] K. Kalimeri and C. Saitis, "Exploring multimodal biosignal features for stress detection during indoor mobility," in *Proceedings of the 18th International Conference on Multimodal Interaction*, 2016, pp. 53–60.

[12] A. Patwardhan and G. Knapp, "Aggressive actions and anger detection from multiple modalities using Kinect," *CoRR*, 2017.

[13] N. Bosch, S. K. D'mello, J. Ocumpaugh, R. S. Baker, and V. Shute, "Using video to automatically detect learner affect in computer-enabled classrooms," *ACM Trans. Interact. Intell. Syst.*, vol. 6, no. 2, pp. 1–26, 2016.

[14] W. Rahman and M. L. Gavrilova, "Emerging EEG and Kinect face fusion for biometric identification," in *Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017, pp. 1–8.

[15] E. Pei, L. Yang, D. Jiang, and H. Sahli, "Multimodal dimensional affect recognition using deep bidirectional long short-term memory recurrent neural networks," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 2015, pp. 208–214.

[16] N. Jaques, S. Taylor, A. Sano, and R. Picard, "Multimodal autoencoder : A deep learning approach to filling in missing sensor data and enabling better mood prediction," in *Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction*, pp. 202–208, 2017.

[17] A. Patwardhan and G. Knapp, "Multimodal affect recognition using Kinect," *arXiv Prepr. arXiv1607.02652*, 2016.

[18] J. Grafsgaard, K. Boyer, E. Wiebe, and J. Lester, "Analyzing posture and affect in task-oriented tutoring," in *Proceedings of the International Conference of the Florida Artificial Intelligence Research Society*, 2012, pp. 438–443.

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

[20] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 211–223, 2012.

[21] J. M. Harley, F. Bouchet, M. S. Hussain, R. Azevedo, and R. Calvo, "A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system," *Comput. Human Behav.*, vol. 48, pp. 615–625, 2015.

[22] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, and A. Paiva, "Automatic analysis of affective postures and body motion to detect engagement with a game companion," in *Proceedings of the 6th International Conference on Human-robot Interaction*, 2011, pp. 305–312.

[23] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. 3, pp. 3371–3408, 2010.

[24] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in Neural Processing Systems*, 2012, pp. 2222–2230.

[25] M. Wu and N. Goodman, "Multimodal generative models for scalable weakly-supervised learning," in *Advances in Neural Processing Systems*, 2018, pp. 5580–5590.

[26] Z. Liu, W. Zhang, S. Lin, and T. Q. S. Quek, "Heterogeneous sensor data fusion by deep multimodal encoding," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 3, pp. 479–491, 2017.

[27] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep Patient: An unsupervised representation to predict the future of patients from the electronic health records," *Sci. Rep.*, vol. 6, pp. 1–10, 2016.

[28] J. Ocumpaugh, R. S. Baker, and M. T. Rodrigo, "Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual," 2015.

[29] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," 2012.

[30] T. G. Weiss, T. Carayannis, R. Jolly, T. G. Weiss, T. Ca, and R. Jolly, "Missing data: A systematic review of how they are reported and handled," *Epidemiology*, vol. 12, no. 5, pp. 729–732, 2016.

[31] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

[32] I. Mierswa, M. Wurst, R. Klinkenberg, and M. Scholz, "Yale: Rapid prototyping for complex data mining tasks," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006, pp. 935–940.