Sensor-based Data Fusion for Multimodal Affect Detection in Game-based Learning Environments

Nathan L. Henderson, Jonathan P. Rowe, Bradford W. Mott, and James C. Lester
North Carolina State University
Raleigh, North Carolina, 27695, USA
{nlhender, jprowe, bwmott, lester}@ncsu.edu

ABSTRACT

Affect detection is central to educational data mining because of its potential contribution to predicting learning processes and outcomes. Using multiple modalities has been shown to increase the performance of affect detection. With the rise of sensor-based modalities due to their relatively low cost and high level of flexibility, there has been a marked increase in research efforts pertaining to sensor-based, multimodal systems for affective computing problems. In this paper, we demonstrate the impact that multimodal systems can have when using Microsoft Kinect-based posture data and electrodermal activity data for the analysis of affective states displayed by students engaged with a game-based learning environment. We compare the effectiveness of both support vector machines and deep neural networks as affect classifiers. Additionally, we evaluate different types of data fusion to determine which method for combining the separate modalities yields the highest classification rate. Results indicate that multimodal approaches outperform unimodal baseline classifiers, and feature-level concatenation offers the highest performance among the data fusion techniques.

Keywords

Affect detection, data fusion, deep learning, posture, electrodermal activity, sensor-based learning

1. INTRODUCTION

Affect detection plays a role of growing importance in educational data mining. Accurately detecting affect is vital to understanding learning. While states such as confusion or engagement have been previously correlated with positive learning outcomes [20], other emotions such as boredom have been associated with negative learning outcomes [5]. Similarly, it has been found that affect detection can potentially be used to avoid negative learning outcomes [10].

To more closely model the human cognitive perception and recognition of certain states, affective modeling techniques have expanded to include multiple parallel data streams that are processed simultaneously to form a single affect prediction or approximation; such systems are referred to as "multimodal" [2]. Each data stream, or "modality," can be provided by a wide array of sources ranging from user interaction logs to eye gaze tracking. The processing of multiple independent modalities has been shown to boost affect classifier performance [6] and provide additional insight into the various aspects of a student's interaction with an intelligent tutoring system [11]. Multimodal computing can be highly beneficial to affective computing and educational data mining tasks by providing multiple complementary perspectives on a single subject or event [3].

A common implementation of multimodal affect detection systems utilizes sensors as perceptors to capture physical data and activity. This enables the system to process different types of physiological and positional information that signify different affective states of students. Sensors are commonly deployed within multimodal systems due to their relatively low expense, flexibility with regards to hardware and software requirements, and generalization across a variety of domains. Consequently, sensor-based multimodal systems have been the focus of several research efforts in recent years. Examples of sensor-based modalities include facial expression [1], posture [9], electroencephalogram (EEG) data [24], and electrodermal activity (EDA) [15].

Sensor-based systems are not without inherent challenges [7]. Such systems can be plagued by issues such as calibration problems, mistracking, noise, irregular behavior, inconsistent data transfer, and synchronization issues. Cultural and social behaviors of participants engaged in a sensor-based system can also impact performance, as well. In certain instances, a sensor may malfunction for an extended period of time, resulting in large intervals of missing or invalid data for one or more modalities.

In this paper, we investigate sensor-based multimodal models for affect detection using data from students engaged with a game-based learning environment for emergency medicine. We utilize student posture information captured by a Microsoft Kinect, as well as EDA data captured by an Affectiva Q-Sensor. We compare the performance of support vector machine (SVM) and deep feedforward neural network models as affect classifiers using unimodal data, as well as multimodal data combining the posture and EDA data channels. Finally, we evaluate three different variations of data fusion for the multimodal affect classifiers. Results suggest improved performance of multimodal classifiers as compared to unimodal classifiers trained on separate Kinect and Q-Sensor modalities, and they reveal the impact that different data fusion techniques have on a classifier's accuracy with multimodal datasets.

2. RELATED WORK

Because of their domain independence, sensors have been integrated into a wide selection of multimodal affect detection systems. Pei et al. [23] utilize long short-term memory (LSTM) recurrent neural networks for a binary affect classification task with audio and visual recordings. Nazari et al. [18] implement a multimodal system to detect instances of narcissism in individuals using modalities such as facial expressions, dialogue, vocal acoustics, and behavioral cues. Facial tracking is paired with self-assessment post-tests to detect student engagement with MetaTutor, an adaptive learning system with a curricular focus on the human circulatory system [12]. Additionally, Muller et al. [17] implement a multimodal affect detection system based on human pose, motion tracking and speech to classify instances of four affective states (anger, happiness, sadness, and surprise) as well as estimate continuous level valence and arousal. Other sensor-based systems use modalities such as eye gaze to predict learning outcomes using gradient tree boosting algorithms [25].

The use of posture data within affect detection systems has experienced a significant increase in recent years. Low-cost sensors such as the Microsoft Kinect have allowed this modality to be easily integrated into multimodal systems. As shown in [22], Kinect-based posture data can be used by supervised and rule-based algorithms to detect various affective states. Likewise, Grafsgaard et al. [9] use Kinect data to estimate student engagement in computer-based tutoring systems used to teach introductory programming concepts. Shifts in posture have been linked to affective states such as frustration, and thus have been associated with negative learning outcomes [9]. When used in conjunction with other modalities such as facial expression and gesture tracking, posture can also be indicative of engagement, learning, and self-efficacy, as [10] demonstrates through the use of stepwise linear regression techniques. Finally, Kinect data has also been utilized for tasks involving anger detection [21] and biometric identification [24].

In addition to posture and pose-related data, advances in multimodal systems have also extended to biosignal modalities. Examples of such work include [24], where Kinect-based posture data is combined with EEG data through sensor fusion to construct a reliable biometric identification model. Additional low-cost sensors were used to capture EEG, EDA, and electromyography (EMG) data, where results indicate that a multimodal approach outperformed unimodal detectors for arousal and valence levels [8]. Using support vector machines, EEG data as well as eye gaze data was used to predict emotional response to videos [27]. The combination of EDA and EEG data has likewise been applied to the problem of stress detection [15] and frustration detection [7]. EDA has been paired with Kinect-based posture data and webcam-based facial expression data to predict students' instances of frustration and engagement in response to tutor questions in an educational environment [29].

3. DATASET

We investigate different multimodal affect classifiers within the context of a game-based environment for emergency medical training, the Tactical Combat Casualty Care Simulation (TC3Sim). Developed by Engineering and Computer Simulations (ECS), TC3Sim is widely used by the U.S. Army to provide realistic combat medic simulations for soldiers. Students assume the first-person perspective of a combat medic involved in different scenarios alongside a variety of non-player characters (NPCs). During a training scenario, participants are faced with different tasks in real time such as securing the area, applying appropriate medical care to combat victims, and preparing for evacuation. The Kinect-based posture data and EDA data collected by the Q-Sensor is captured during four different training scenarios: a leg injury scenario, an introductory training scenario, a story-driven narrative scenario, and a patient expiration scenario that portrays a combat victim expiring regardless of the actions of the player. A screenshot of a player's first-person perspective when engaged with TC3Sim is shown in Figure 1.

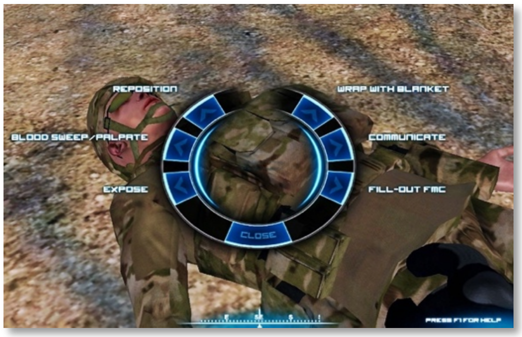

Figure 1. TC3Sim game-based learning environment.

The dataset used in this work was collected from a study with 119 cadets from the United States Military Academy (83% male. 17% female) who participated in different training sessions with TC3Sim. All participants completed the same training materials, which were administered through the Generalized Intelligent Framework for Tutoring (GIFT) framework. GIFT is a service-oriented software framework designed to aid in the development and deployment of computer-based adaptive training systems [28]. Each participant worked individually at a single workstation, and each session lasted approximately one hour. The posture activity for each participant was captured using a Microsoft Kinect for Windows 1.0 sensor. The head and torso positions and movements were captured using skeleton-tracking features contained in the GIFT framework. The data from the Kinect was sampled at a rate of 10-12 Hz. This modality contained timestamped feature vectors containing coordinates of 91 vertices. For this effort, three vertices were selected in accordance with prior research regarding affect detection with Kinect data [9]: *top_skull*, *center_shoulder,* and *head*. 73 additional features were engineered from this modality during the post-processing stage. These features were summary statistics such as the mean, variance, and standard deviation of the different vertices over time windows of 5, 10, and 20 seconds prior to each observation.

In addition to the postural modality, electrodermal activity was captured from each user using an Affectiva Q-Sensor bracelet worn by each participant. The Q-Sensor captured each user's skin temperature, electrodermal activity, and the sensor's acceleration vectors as determined by an onboard accelerometer. However, in this study, only the EDA readings were used for affect detectors. In a similar fashion to the posture modality, summary statistics were calculated for the EDA modality such as the min, max, and variance of the EDA values for each session, as well as the summary statistics across time windows of the prior 5, 10, and 20 seconds. The net changes in the EDA levels across the previous 3 and 20 seconds were also calculated. However, the Q-Sensors experienced highly inconsistent behavior with regard to the data capture, which affected approximately half of the collected data. Additionally, the interaction trace log data from each session was captured by the GIFT framework, but because this work focuses exclusively on sensor-based modalities, this data was not utilized.

To obtain ground truth labels of each student's affective states, two trained observers marked instances of different displays of affect in accordance with the BROMP protocol [19]. BROMP is a quantitative observation protocol for run-time coding of student affect and behavior during classroom-based interactions [19]. During this process, the two observers walked around the perimeter of the classroom and discreetly marked instances of affect in 20-second intervals using a handheld device. Affective states recorded include *bored, confused, engaged, frustrated,* and *surprised.*

A total of 3,066 separate BROMP observations were collected. Only observations that were collected during students' actual engagement with TC3Sim were kept, and observations where there was disagreement between the two observers were discarded. Agreeing BROMP observations were treated as a single label, and only BROMP observations recorded during the TC3Sim exercise were preserved, excluding instances during pre and post-test surveys, as well as instances occurring during the instructional PowerPoint presentation. Additional factors contributing to the significant reduction in BROMP observations were the subtlety of instances of affect in the cadets compared to classroom participants, as well as cases of multiple different affective states being observed within the same 20-second window. The resulting dataset contained 755 distinct BROMP observations; the distribution of affect instances is shown in Figure 2. Instances of engagement were by far the most common occurrence, while instances of frustration and surprise were sparse. As stated previously, the Q-Sensor experienced frequent stops in data logging. This issue resulted in 333 BROMP observations containing missing EDA information, while a subset of 422 data samples contained both the posture and the EDA modalities. The posture-based modality did not appear to suffer any data loss from the Kinect sensor.

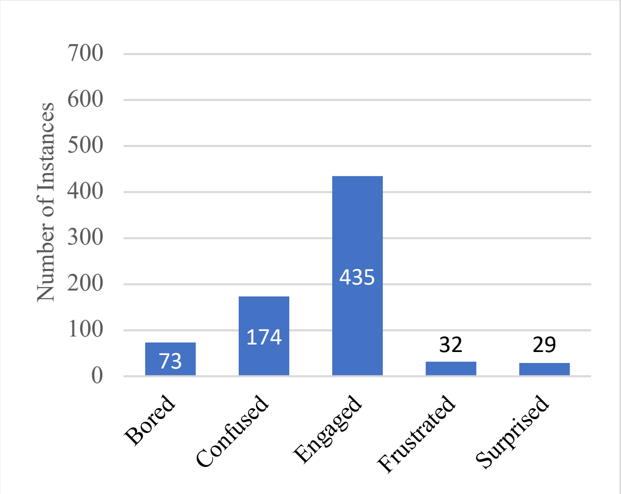

Figure 2. Distribution of affect instances from BROMP observations.

4. METHODOLOGY

The primary goal of this paper is to demonstrate the effectiveness of a multimodal classification system for affect detection using two modalities: Kinect-based posture data and electrodermal activity data. To ensure that both modalities are present in each data sample, any BROMP observation with missing or invalid EDA data was removed from the dataset. Therefore, our classifiers were trained on a dataset using 422 BROMP observations containing correlated posture and EDA data.

4.1 Data Preprocessing

After the aforementioned BROMP observations were removed from the dataset, five separate datasets were created through oversampling of each affective state. The oversampling was accomplished using a minority class cloning technique. Additionally, feature data was scaled using z-score standardization. This method ensures that each attribute of the feature vectors have the same mean and standard deviation but allows for different ranges.

4.2 Feature Selection

Prior to training the classifiers, each dataset underwent forward selection for the purpose of feature selection. This reduces the number of attributes in each dataset through a greedy algorithm that trains a model and selects the best [0, k] features based on each model's Cohen's Kappa [4]. For our work, a k value of 10 was chosen. The model used in feature selection was the sequential minimal optimization (SMO) support vector machine [7]. This polynomial-kernel model was selected due to its linear memory requirements and scalability, as a high number of models were trained to obtain the best features. An attribute was not considered unless it showed positive improvement over the currently-selected dataset, and the attribute showing the highest improvement was kept as a selected feature. The feature selection was implemented using RapidMiner 9.0 [16]. This platform was selected due to its convenience as a toolkit for implementing the data processing pipeline, as well as its use in prior work in affect detection [7].

4.3 Classifiers

Prior work has demonstrated the effectiveness of deep neural networks in affect classification tasks [14]. We utilize the same neural network approach and compare it with SVM models. The SVMs contain a radial kernel function with a convergence epsilon of 0.001 for a maximum of 100,000 iterations. The artificial neural network (ANN) architecture contained feed-forward layers of 800, 800, 500, 100, and 50 nodes, respectively, in addition to a binary classification layer. Each layer's activation function was a Rectified Linear Unit (ReLU). Each network was trained for 10 epochs with the ADADELTA adaptive learning rate [30]. A separate classifier was trained for each affective state, using the selected features of the oversampled data as described in section 4.1.

4.4 Data Fusion

To evaluate different methods of integrating the two modalities for affect classification, we implement several variations of data fusion techniques. We test two types of data fusion: feature-level fusion ("Early Fusion") and decision-level fusion ("Late Fusion"). Early Fusion involves the concatenation of features from the posture and EDA modalities prior to training the affect classifier. Late Fusion calls for the training of separate classifiers for each modality, and the predicted confidence levels of each binary class (positive or negative label of affective state) are processed by a voting schematic to produce a singular prediction of the affective state. The voting schematic can be implemented in different ways, such as majority voting, averaging, or weighting [2]. For this paper, we take the highest confidence value across the two classifiers and use the associated class as our final representative prediction. Two different variations of Early Fusion are also evaluated. The first variation, referred to in this paper as "Early Fusion 1", concatenates the features prior to the feature selection process. The other variation, referred to as “Early Fusion 2”, performs separate feature selection on the separate modalities, and only the selected features are concatenated prior to training the classifiers. A visual representation of the various data fusion pipelines is shown in Figure 3.

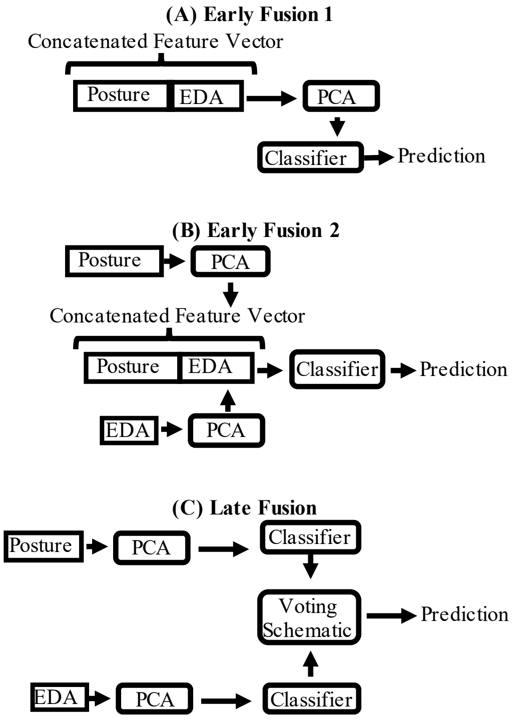

Figure 3. Data pipeline for data fusion variations.

5. RESULTS AND DISCUSSION

The classifiers were evaluated using 10-fold cross validation, with the data split on a per-session basis to ensure that all data from individual training sessions were kept in the same fold. The same batches of data were maintained across all modeling approaches to ensure fair comparisons across classifiers. The unimodal baseline classifiers and Early Fusion pipelines were implemented using RapidMiner 9.0. RapidMiner does not support decision-level fusion, so the Late Fusion pipeline was implemented using Python 3.6, while the classifiers were still implemented in RapidMiner.

Unimodal classifiers were trained on the posture and EDA modalities independently to provide a baseline for the multimodal classifiers’ performance. The results for the posture and EDA-based unimodal classifiers for each affective state are shown in Tables 1 and 2 respectively. Evaluation metrics include Cohen’s Kappa, raw accuracy, and F1 Score. Particular focus is given to Cohen’s Kappa due to its ability to account for the possibility of correct classification due to random chance.

The posture-based SVM returned the highest Kappa for four of the five affective states, and the EDA-based SVM outperformed the ANN for three of the five affective states. The ANN model performed poorly on a majority of the evaluations, returning a negative Kappa on two of the posture-based states and four of the five EDA-based states, indicating that the ANN is no better than a random classifier for a majority of states.

TABLE 1: Classifier Performance for Affective States (Posture)

Bored

Classifier	Kappa	Accuracy	F1 Score
SVM	**0.004**	0.607	0.013
ANN	-0.001	0.408	0.530

Confused

Classifier	Kappa	Accuracy	F1 Score
SVM	**0.002**	0.566	0.040
ANN	-0.003	0.566	0.040

Engaged

Classifier	Kappa	Accuracy	F1 Score
SVM	**0.065**	0.484	0.523
ANN	0.020	0.484	0.523

Frustrated

Classifier	Kappa	Accuracy	F1 Score
SVM	**0.092**	0.553	0.441
ANN	0.063	0.501	0.650

Surprised

Classifier	Kappa	Accuracy	F1 Score
SVM	-0.236	0.632	0.040
ANN	**0.02**	0.270	0.431

TABLE 2: Classifier Performance for Affective States (EDA)

Bored

Classifier	Kappa	Accuracy	F1 Score
SVM	**-0.042**	0.500	0.286
ANN	-0.047	0.360	0.478

Confused

Classifier	Kappa	Accuracy	F1 Score
SVM	**0.033**	0.533	0.319
ANN	-0.083	0.387	0.529

Engaged

Classifier	Kappa	Accuracy	F1 Score
SVM	-0.108	0.449	0.437
ANN	**-0.013**	0.541	0.682

Frustrated

Classifier	Kappa	Accuracy	F1 Score
SVM	-0.046	0.491	0.539
ANN	**0.011**	0.387	0.641

Surprised

Classifier	Kappa	Accuracy	F1 Score
SVM	**0.086**	0.607	0.357
ANN	-0.001	0.222	0.478

The posture classifiers performed relatively poorly on *boredom, confused,* and *surprised*. It is worth noting that *surprised* contains the lowest number of positive instances within the dataset, which may contribute to the poor performance. Additionally, it is possible that postural behavior may not distinguishably change between positive instances of *boredom* and *confused*, lead to common misclassifications across the two states. The EDA classifiers also performed poorly on the affective states of *bored, engaged,* and *frustrated*. However, the EDA modality contains significantly fewer features than the posture modality, and this may have caused additional misclassifications. It is also possible that the EDA modality may not contain enough variance for the classifiers to distinguish between positive and negative instances of affective states. Additionally, the EDA classifiers face the task of distinguishing between different changes in the EDA measurements, and determining whether such changes can be attributed to a particular affective state or another cause. However, this proves to be more difficult than the posture modality due to the singular dimensionality of the EDA channel. To further illustrate this issue, a graphical representation of the change in EDA throughout a session is shown in Figure 4.

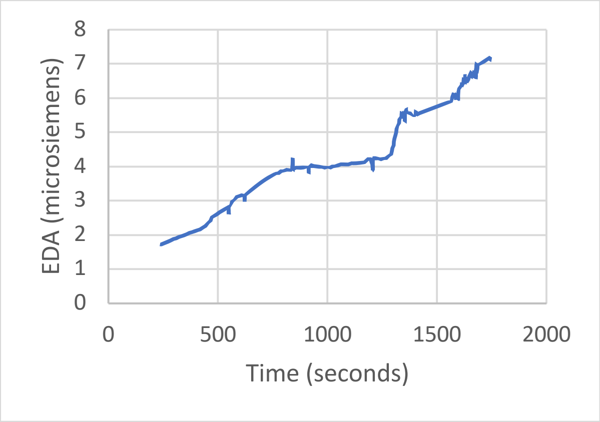

Figure 4. EDA measured over duration of a single training session.

The SVM was selected as the classifier used to implement and evaluate the data fusion methods discussed in Section 4.4. The same feature selection algorithm and classifier configuration were used as in the unimodal approach, and the same session-level groupings were also maintained. The three different data fusion approaches were evaluated for each affective state, and the results for each state are shown in Table 3.

Early Fusion 2 returned the highest Kappa for *bored, engaged,* and *frustrated*. Early Fusion 1 returned the highest value for *confused*, while the Q-Sensor baseline was the highest value for *surprised*. One possible reason that Early Fusion 2 is the highest-performing data fusion method is because feature selection is performed separately on each modality prior to each classifier. This means that if each feature selection algorithm selects up to the *k*th best features, then the combined feature vector can contain up to 2**k* features, twice as many features as allowed by Early Fusion 1. This increase in features may boost the performance of the classifier. Late Fusion can also work with 2**k* features, but the features are split between the two unimodal classifiers before decision-level fusion. Early Fusion 2 also explores the correlations between various inter-modal attributes more deeply compared to Early Fusion 1. The complex relationships between various intra-modal features are explicitly modeled in the feature selection performed on each independent modality, while the correlations between the selected inter-modal features are explored when training the primary classifier following feature selection. However, these two stages are performed simultaneously in Early Fusion 1 and certain complex relationships may not be detected as a result.

TABLE 3: Performances for Early Fusion 1, Early Fusion 2, and Late Fusion for Affective States using SVM

Bored

Classifier	Kappa	Accuracy	F1 Score
Early Fusion 1	-0.082	0.466	0.164
Early Fusion 2	**0.041**	0.5318	0.356
Late Fusion	-0.056	0.583	0.145

Confused

Classifier	Kappa	Accuracy	F1 Score
Early Fusion 1	**0.049**	0.566	0.300
Early Fusion 2	-0.004	0.515	0.321
Late Fusion	0.032	0.597	0.148

Engaged

Classifier	Kappa	Accuracy	F1 Score
Early Fusion 1	-0.064	0.446	0.393
Early Fusion 2	**0.068**	0.542	0.491
Late Fusion	-0.035	0.481	0.459

Frustrated

Classifier	Kappa	Accuracy	F1 Score
Early Fusion 1	0.191	0.657	0.656
Early Fusion 2	**0.246**	0.594	0.483
Late Fusion	0.119	0.5679	0.490

Surprised

Classifier	Kappa	Accuracy	F1 Score
Early Fusion 1	**-0.021**	0.590	0.053
Early Fusion 2	0.013	0.682	0.080
Late Fusion	-0.192	0.514	0.124

Late Fusion provides the ability to “correct” a possibly incorrect prediction across the two modalities. For example, if the postural classifier produces an incorrect prediction of TRUE with a confidence level of 0.6, but the EDA classifier produces an accurate prediction of FALSE with a confidence level of 0.8, then the EDA modality overrides the incorrect prediction because of our selected voting schematic. However, Late Fusion was not the optimal fusion method for any of the affective states, though its effectiveness as a multimodal fusion technique has been demonstrated in other affective computing tasks [14].

Of note is the performance of the multimodal classifier on the frustration dataset compared to the other affective states, as the classifier achieved substantially higher Kappa scores. One possible explanation for this behavior is that negative, high-arousal emotions such as frustration or anger have been shown to occur relatively infrequently in students engaged with computer-based learning environments [13]. This may possibly mean that the recorded instances of frustration may contain more distinguishable features compared to other common, low-arousal affective states such as boredom and engagement, encouraging higher performance from the frustration-based classifier. Additionally, frustration has been demonstrated to illicit higher EDA levels [26], indicating that the inclusion of the EDA modality with the posture modality provides additional informative features to the feature vectors, contributing to the relatively high performance of the classifier.

Although the multimodal classifiers generally outperformed unimodal classifiers, the highest-performing model returned a relatively low Kappa compared to the performance of a human BROMP labeler (~0.6). However, this threshold can vary depending on the affective state and intervention associated with each state. For example, identifying instances of *engagement* can be viewed as a lower priority than identifying instances of *frustration* or *boredom*, as these affective states often necessitate a dynamic intervention to improve learning outcomes. However, the Kappas for most of the classifiers fall below 0.05, indicating significant difficulty for several classifiers in achieving consistent performance across multiple affective states.

Previous research efforts have demonstrated that the EDA modality does not have a tightly-coupled relationship with different affective states when compared to other higher-dimensionality modalities such as facial expression and gesture [13]. The results of our work also indicate that the EDA modality resulted in at least one classifier returning a negative Kappa for all five affective states. Possible explanations for this behavior include an inadequate amount of training data, lack of variance or distinguishable trends across the observed time windows, or lack of useful features (17 EDA features vs. 75 posture features). However, our results indicate that the EDA modality does generally improve classifier performance when used in conjunction with the posture modality.

6. CONCLUSION

In this paper, we demonstrate the effectiveness of a multimodal affect detection system based on sensor data capturing a user's posture and EDA data while engaged with a game-based learning environment. We show the improvement that multimodal classifiers achieve compared with unimodal classifiers for both modalities. We also demonstrate that SVMs outperform ANNs as a unimodal classifier in this particular domain. Finally, we demonstrate that data fusion is an effective way to combine multiple modalities, either prior to or following classification.

Results suggest several promising directions for future work. To improve model performance on smaller datasets or data containing instances of missing modalities, more sophisticated feature engineering approaches can be evaluated. The evaluation of our data fusion techniques with additional modalities can further indicate the effectiveness of this approach in a variety of multimodal systems. Additional exploration of generalizable multimodal systems should be undertaken to further utilize the flexibility of sensor-based systems. Further evaluation of classification algorithms can be investigated as well, in particular, algorithms designed for the processing of temporal data such as recurrent neural networks. The impact of additional biosignal modalities such as EEG or EMG data would provide a more in-depth perspective of the effect such modalities have on multimodal affect detection systems. Finally, the integration of multimodal affect detection into a run-time learning environment would enable adaptive pedagogical functionalities that address potentially negative learning outcomes through the use of dynamic interventions and user-tailored feedback based on learners' affective states.

7. ACKNOWLEDGMENTS

We wish to thank Dr. Jeanine DeFalco, Dr. Benjamin Goldberg, and Dr. Keith Brawner of the U.S. Army Combat Capabilities Development Command, Dr. Mike Matthews and COL James Ness of the U.S. Military Academy, Dr. Robert Sottilare of SoarTech, and Dr. Ryan Baker of 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.

8. REFERENCES

[1] Arroyo, I., Cooper, D.G., Burleson, W., Woolf, B.P., Muldner, K. and Christopherson, R. 2009. Emotion sensors go to school. In *Proceedings of the 14th International Conference on Artificial Intelligence In Education* (2009), 17–24.

[2] Baltrušaitis, T., Ahuja, C. and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 41, 2 (2018), 423–443. DOI:https://doi.org/10.1109/TPAMI.2018.2798607.

[3] Chang, C.M., Su, B.H., Lin, S.C., Li, J.L. and Lee, C.C. 2017. A bootstrapped multi-view weighted kernel fusion framework for cross-corpus integration of multimodal emotion recognition. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)* (2017), 377–382.

[4] Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*. 20, 1 (1960), 37–46.

[5] Craig, S., Graesser, A., Sullins, J. and Gholson, B. 2005. Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media*. 29, 3 (2005), 241–250. DOI:https://doi.org/10.1080/1358165042000283101.

[6] D'Mello, S. and Kory, J. 2012. Consistent but modest: A meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. *Proceedings of the 14th ACM international conference on Multimodal interaction - ICMI '12*. (2012), 31–38. DOI:https://doi.org/10.1145/2388676.2388686.

[7] DeFalco, J.A., Rowe, J.P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B.W., Baker, R.S. and Lester, J.C. 2018. Detecting and addressing frustration in a serious game for military training. *International Journal of Artificial Intelligence in Education*. 28, 2 (2018), 152–193. DOI:https://doi.org/10.1007/s40593-017-0152-1.

[8] Girardi, D., Lanubile, F. and Novielli, N. 2017. Emotion detection using noninvasive low cost sensors. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction* (2017), 125–130.

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

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

[11] Grafsgaard, J.F., Wiggins, J.B., Vail, A.K., Boyer, K.E., Wiebe, E.N. and Lester, J.C. 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* (2014), 42–49.

[12] Harley, J.M., Bouchet, F. and Azevedo, R. 2013. Aligning and comparing data on emotions experienced during

learning with MetaTutor. In *International Conference on Artificial Intelligence in Education* (2013), 61–70.

[13] Harley, J.M., Bouchet, F., Hussain, M.S., Azevedo, R. and Calvo, R. 2015. A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system. *Computers in Human Behavior*. 48, May (2015), 615–625. DOI:https://doi.org/10.1016/j.chb.2015.02.013.

[14] Henderson, N.L., Rowe, J.P., Mott, B.W., Brawner, K., Baker, R.S. and Lester, J.C. 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 (in press)* (2019).

[15] Kalimeri, K. and Saitis, C. 2016. Exploring multimodal biosignal features for stress detection during indoor mobility. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (2016), 53–60.

[16] Mierswa, I., Wurst, M., Klinkenberg, R. and Scholz, M. 2006. Yale: Rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (2006), 935–940.

[17] Muller, P.M., Amin, S., Verma, P., Andriluka, M. and Bulling, A. 2015. Emotion recognition from embedded bodily expressions and speech during dyadic interactions. *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*. (2015), 663–669. DOI:https://doi.org/10.1109/ACII.2015.7344640.

[18] Nazari, Z., Lucas, G. and Gratch, J. 2015. Multimodal approach for automatic recognition of machiavellianism. *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*. (2015), 215–221. DOI:https://doi.org/10.1109/ACII.2015.7344574.

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

[20] Pardos, Z., Baker, R., Pedro, M.S., Gowda, S.M. and Gowda, S.M. 2014. Affective states and state tests: investigating how affect and engagement during the school year predict end-of-year learning outcomes. *Journal of Learning Analytics*. 1, 1 (2014), 107–128. DOI:https://doi.org/10.1145/2460296.2460320.

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

[22] Patwardhan, A. and Knapp, G. 2016. Multimodal affect recognition using Kinect. *arXiv preprint arXiv:1607.02652*. (2016).

[23] Pei, E., Yang, L., Jiang, D. and Sahli, H. 2015. 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 (ACII)* (2015), 208–214.

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

[25] Rajendran, R., Carter, K.E. and Levin, D.T. 2018. Predicting Learning by Analyzing Eye-Gaze Data of Reading Behavior. *International Educational Data Mining Society*. (2018).

[26] Ramachandran, B.R.N., Pinto, S.A.R., Born, J., Winkler, S. and Ratnam, R. 2017. Measuring neural, physiological and behavioral effects of frustration. In *Proceedings of the 16th International Conference on Biomedical Engineering* (2017), 43–46.

[27] Soleymani, M., Pantic, M. and Pun, T. 2012. Multimodal emotion recognition in response to videos. *IEEE transactions on affective computing*. 3, 2 (2012), 211–223.

[28] Sottilare, R.A., Baker, R.S., Graesser, A.C. and Lester, J.C. 2018. Special Issue on the Generalized Intelligent Framework for Tutoring (GIFT): Creating a stable and flexible platform for innovations in AIED Research. *International Journal of Artificial Intelligence in Education*. 28, 2 (2018), 139–151. DOI:https://doi.org/10.1007/s40593-017-0149-9.

[29] Vail, A.K., Wiggins, J.B., Grafsgaard, J.F., Boyer, K.E., Wiebe, E.N. and Lester, J.C. 2016. The Affective Impact of Tutor Questions: Predicting Frustration and Engagement Alexandria. *International Educational Data Mining Society*. (2016), 247–254. DOI:https://doi.org/10.1145/1235.

[30] Zeiler, M.D. 2012. ADADELTA: An adaptive learning rate method. (2012). DOI:https://doi.org/http://doi.acm.org.ezproxy.lib.ucf.edu/10.1145/1830483.1830503.