Enhancing Multimodal Affect Recognition with Multi-Task Affective Dynamics 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 recognizing students' affective states is critical for enabling adaptive learning environments to promote engagement and enhance learning outcomes. Multimodal approaches to student affect recognition capture multi-dimensional patterns of student behavior through the use of multiple data channels. An important factor in multimodal affect recognition is the context in which affect is experienced and exhibited. In this paper, we present a multimodal, multi-task affect recognition framework that predicts students' future affective states as auxiliary training tasks and uses prior affective states as input features to capture bi-directional affective dynamics and enhance the training of affect recognition models. Additionally, we investigate cross-stitch networks to maintain parameterized separation between shared and task-specific representations and task-specific uncertainty-weighted loss functions for contextual modeling of student affective states. We evaluate our approach using interaction and posture data captured from students engaged with a game-based learning environment for emergency medical training. Results indicate that the affective dynamics-based approach yields significant improvements in multimodal affect recognition across four different affective states.**

***Index Terms*—multitask learning, affect recognition, multimodal interaction, game-based learning environments**

I. Introduction

Affect is a critical component of learning [1]. Positively valanced emotions such as *delight* or *flow* are often associated with improved learning outcomes and engagement [2]. However, negative emotions such as *boredom* often result in decreased learning outcomes and can be indicative of disengagement or disinterest [3]. Other emotions such as *frustration* or *confusion* often have a complex relationship with student learning. For example, *frustration* has been shown to be associated with a student's attempt at overcoming a learning impasse or challenge, which is a vital component of learning [2]. The emotion of *confusion* is complex as well, as it is uncomfortable but may coincide with experiences of cognitive disequilibrium that precedes learning [4]. As students progress through a learning environment, they may experience a wide range of affective states, which are influential in shaping their learning outcomes as well as their motivational and cognitive processes [5]. For example, *frustration* has been shown to be followed by *boredom* (potentially leading to disengagement) as well as *confusion* (potentially leading to a state of increased engagement) [4]. Often, an impasse in learning (e.g., difficulties coinciding with a state of *confusion*) that is overcome easily may result in a rapid transition to a state of *engaged concentration*. However, if a student persists in a state of *confusion* for extended periods of time, a transition to a state of *frustration* may occur, increasing the risk of potential disengagement. Potential transitions to affective states that are correlated with diminished learning outcomes can be mitigated through the implementation of affect-sensitive interventions. Affect-sensitive interventions can promote engagement and emotion regulation in support of student learning [6]. Creating affect-sensitive interventions requires the development of affect recognition models that accurately detect students' academic affective states based on observed student behavior data. The patterns and sequences of emotions that occur during student learning may provide valuable insight into a student's current emotional state, and subsequent affect recognition models that take into account affective dynamics hold potential to yield improved predictive performance.

Recent years have seen an increased focus on multimodal student affect recognition models [7] due to their ability to capture multiple concurrent perspectives on a student's behavior. This process of capturing multiple data channels from varying data sources is reflective of human perception and has demonstrated improved predictive performance over unimodal systems [7], [8]. *Sensor-based* multimodal systems capture representations of a student's physical behavior such as a student's posture [9], facial expressions [10], or speech [11] through the use of physical sensors. An alternative to sensor-based systems are *sensor-free* systems. These multimodal frameworks are typically based on trace log data that contains recordings of student activity within a learning environment, such as gameplay actions within game-based learning environments [12].

A promising approach to modeling affective sequences is predicting multiple affective states with a single output vector. A static output vector can represent a single affective sequence consisting of multiple target variables, each representing an affective state at a particular time interval. Because this necessitates a single model making multiple concurrent predictions, multi-task learning (MTL) provides a natural solution. MTL has several advantages over single-task modeling, including the ability to share feature representations

978-1-6654-0019-0/21/$31.00 ©2021 IEEE

and learned weights across multiple target variables, which introduces a form of model regularization [13]. Multi-task models also require a significantly lower number of parameters compared to the total number of parameters required by separately trained models for each individual task, while also allowing the model to inherently learn the interwoven relationships between the target variables [14]. Prior work indicates that MTL outperforms single-task learning in terms of predictive performance for a variety of tasks [15], [16]. However, the use of multiple tasks poses challenges regarding the weighting of each task's predictive performance during training, as different predictive tasks often vary in nature and intended purpose. Additionally, the appropriate balance of task-specific and shared latent representations within a multi-task model can vary as well and have a noticeable impact on model performance.

In this paper, we investigate the integration of temporal contextual features from students' affective sequences as a means to improve models of student affect through multi-task learning. We hypothesize that using students' future affective states as they engage with a game-based learning environment can be utilized as an auxiliary multi-task function to improve the predictive performance of the affect recognition models. Additionally, we explore how to optimally combine interaction-based and posture-based modalities by exploring potential deep learning architectures: multi-task fully connected feedforward neural networks and cross-stitch neural networks [17]. Finally, we examine the benefit of including a student's prior affective states as a means of providing additional affect sequence information to improve the performance of affect recognition models. Our results indicate that the use of MTL to model affective sequence patterns from each student leads to improved prediction of multiple affective states, and the use of cross-stitch neural networks further strengthens predictive accuracy.

II. Related Work

A. Multimodal Affect Recognition

Due to their multifaceted perception of student behavior and demonstrated improvements in predictive performance, multimodal approaches to affect recognition tasks have seen a growing interest in recent years. Song et al. use captured audio and facial expression data to train a recurrent neural network model to detect the presence of frustration in students [18]. Henderson et al. showed the improved performance of multimodal affect models through the use of interaction- and posture-based modalities while also investigating the impact of multimodal data fusion on predictive accuracy [8]. Wu et al. used head pose and eye gaze data to enhance the performance of a facial expression-based continuous affect recognition model through the use of a guided temporal attention mechanism [19], while Ghaleb et al. integrated temporal contextual embeddings into multimodal long short-term memory (LSTM) models trained on audio and facial expression modalities [20].

B. Affective Sequences

Affective dynamics has been the subject of growing interest. While a significant body of prior work focuses on predicting individual occurrences of various affective states [8], [21], these approaches often ignore the predictive information offered by a student's overall affective trajectory, such as how a student transitions from one affective state to another throughout a single learning session. The shifts in a student's affective states have been shown to reveal particular recurring patterns, and thus can provide predictive value in affect modeling. D'Mello and Graesser investigated a model of affective dynamics that focused on a cyclic model from engaged concentration to confusion that enhanced learning outcomes [4]. Additionally, the authors explored an alternative model that resulted in decreased learning as students transitioned from engaged concentration to confusion, frustration, and boredom, respectively. Andres et al. expanded on this work by exploring the usage of shorter transitory patterns, namely two-step patterns that consisted of only two affective states [22]. The authors investigated the presence of prolonged states of affect by analyzing four-step patterns of the same affective state, focusing on the correlation between particular affective patterns and student learning outcomes. Ocumpaugh et al. focused on the frequency of an expanded set of four-step prolonged emotions in addition to three-step patterns consisting of two affective states as they related to student actions in a blended learning system [23]. Botelho et al. investigated the performance of two-step affective transitions in students engaged with an intelligent tutoring system, in addition to investigating the time a student spent in a single affective state before transitioning to another state [24]. While there has been prior work investigating the use of sequential modeling (such as LSTMs) for predicting student affect [25], these approaches use the sequences of student behavioral data as *input* to predict a single affect label instead of predicting an affect sequence as *output*. Our work addresses this issue by exploiting the temporal information in students' affective sequences to improve the predictive performance of the affect models through the use of auxiliary output multi-task predictions during the training phase.

C. Multi-Task Learning

Recent years have seen MTL applied to a variety of tasks, including computer vision [26], natural language processing [27], and transfer learning [28]. While many multi-task models consist of a series of feedforward neural network layers, alternative deep learning architectures have been explored as well, including sluice networks [29], deep relationship networks [16], and cross-stitch networks [17]. More recently, MTL has been investigated as an alternative approach to student modeling, including for predicting post-test scores based on individual questions [15], modeling student mastery of multiple concepts [30], and estimating self-reported measures of interest and engagement with a game-based learning environment [31]. Prior applications of multi-task learning within affective computing involve the prediction of multiple states of affect by a single model [25].

III. Dataset

The dataset used to investigate our multi-task, affective dynamics-based approach consists of posture and interaction data captured during student engagement with a game-based learning environment for training emergency medical skills, *TC3Sim*. Posture data was captured using a Microsoft Kinect sensor mounted on a tripod facing the front of each student, while the interaction data was extracted from gameplay trace data logs. Students' affective states during gameplay were discreetly annotated and recorded in real time by two field observers in accordance with the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) [32]. For this study, data was obtained from a population of 119 students (83% male, 17% female) at The United States Military Academy.

A. TC3Sim Game-Based Learning Environment

TC3Sim is a serious game-based learning environment that is widely used to provide training for administering medical care within a 3D virtual environment. During the game, students assume the first-person role of a medic within various simulated narratives (Fig. 1). Students progress through the game by completing a series of scenarios that are centered on different non-player characters (NPCs) that sustain injuries and require medical attention. The students' characters administer care in accordance with medical protocol that is presented to each student prior to beginning the gameplay session. Each student engaged with *TC3Sim* individually, with each gameplay session lasting approximately one hour.

Fig. 1. *TC3Sim* game-based learning environment.

B. BROMP Protocol

Ground-truth labels of student affect were collected using the BROMP protocol [32]. BROMP is a coding procedure designed to produce quantitative labels of student affect and behavior using field observers and allows for efficient and discreet real-time annotations of learner affect based on holistic observations within real-world conditions. BROMP has been widely used in research on affect-sensitive learning technologies [33]. Notably, BROMP observations do not rely on a single data channel (e.g., facial expression). BROMP enables annotations to be contextually informed by the observers and includes practices for minimizing disruptions during annotation. Since BROMP is an observational protocol, it mitigates issues with self-reports such as recall, self-awareness, and self-presentation [34].

Observers walk around the perimeter of the classroom and discreetly annotate observed students' affective states using a hand-held device. Annotations of affect occurred in 20-second intervals and were intended to be captured as discreetly as possible to minimize the influence of the observers' presence and disruption of the students' gameplay. Prior to this study, the two observers established an inter-rater agreement exceeding 0.6 in terms of Cohen's Kappa [35].

Any observations indicating disagreement between the observers were removed from the dataset, resulting in a final dataset consisting of 755 labeled affective states. A total of 435 of the BROMP observations were labeled as *engaged concentration* (M = 0.576, SD = 0.239), 174 as *confused* (M = 0.231, SD = 0.185), 73 as *bored* (M = 0.097, SD = 0.161), 32 as *frustrated* (M = 0.042, SD = 0.182), 29 as *surprised* (M = 0.038, SD = 0.045) and 12 as *anxious* (M = 0.016, SD = 0.089). Due to the low number of observations of *anxious*, this affective state is not considered in any of the following analyses.

IV. Methodology

A. MTL with Affective Sequences

To adapt the single-task affect recognition approach to an MTL formulation, the target variables were expanded to include a one-hot representation of each possible affective state. The one-hot vector was indicative of the affective state B_{i+1} that followed the current BROMP observation B_i (Fig. 2). B_i was a binary indicator of the presence of one of the five possible affective states. Therefore, the multi-task models were modeled using a label vector of size 6 (binary indicator of a single affective state + one-hot vector of size 5). Using the affect model for *bored* as an example, the multi-task output vector for a positive annotated occurrence of *bored* followed by an annotation of *confused* would be [1, 0, 1, 0, 0, 0], while a negative annotated occurrence of *bored* followed by a subsequent annotation of *frustrated* would be [0, 0, 0, 0, 1, 0].

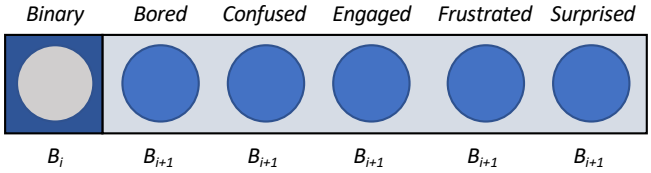

Fig. 2. Multi-task feature vector representation.

Because the multi-task models are predicting future occurrences of each affective state, it is impractical to utilize these labels as input features as this information would not be available in a run-time environment. As a result, we use these labels as auxiliary output variables for the purpose of boosting the predictive performance of the multi-task models relative to the current affective state, an approach that has been previously demonstrated to improve predictive performance [27], [36]. This process can be employed when certain features are unhelpful for predicting other output variables or are not available until after the predictions are made, allowing the features to be used to present additional information to the model during the *training process only* [36]. In this case, presenting the subsequent affective state to the multi-task model allows the model to potentially observe differential patterns in student behavior prior to transitioning to another affective state. For example, a student's postural behavior while currently in a state of *engaged concentration* may fluctuate depending on if the subsequent affective state is also *engaged concentration* or a different state such as *confusion*. By introducing additional predictive tasks, the model is trained to extract temporal features and patterns from affective sequences that can improve the model's prediction of the current affective state. The occurrences of each two-step affective sequence are shown in Fig. 3.

The most common affect sequences are persisting states of *engaged concentration* (denoted as "Concentrating" in Fig. 3), consecutive states of *confusion*, and alternating between these two states. This result aligns with the proposed model by D'Mello and Graesser [4]. Other notable sequences are students transitions between states of *bored* and *engaged concentration*, particularly as this indicates that students are often capable from returning to an engaged state while previously being in a state of relative disengagement, a behavior previously observed by Andres et al. [22].

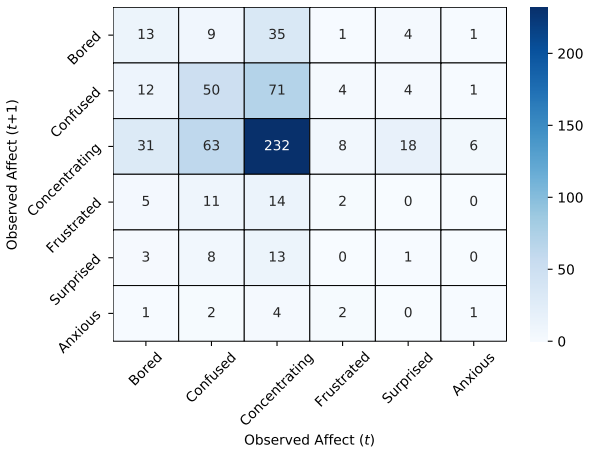

Fig. 3. Frequency of each affective state and corresponding subsequent state.

B. Cross-Stitch Networks

An active area of investigation in multi-task deep learning is determining the appropriate level of layer connectivity across each task. A multi-task model that consists of only fully connected layers contains the highest level of connectivity, as each layer propagates the same fully shared data representation across all tasks with the exception of task-specific output layers. Alternatively, to avoid any inter-task communication within the multi-task framework, separate models can be trained for each task, so that the trained weights are unique for each output. While prior work applying multi-task learning for student modeling utilizes full connectivity across tasks within the model's hidden layers [15], [30], other work within computer vision has explored the benefits of "split" neural architectures, or architectures that maintain a degree of separation between tasks within a pre-determined subset of the model's hidden layers.

Cross-stitch networks were proposed by Misra et al. as a generalizable approach to implementing "split" architectures by implementing parameterized linear combinations between a network's hidden layers that can learn optimal weightings between shared and task-specific latent representations [17]. This approach allows feature representations to be combined within certain hidden layers and shared across tasks while also maintaining separation between task-specific representations. For example, in the case of modeling two tasks (A and B), a learned weight matrix α is used to parameterize the linear combinations of multiple tasks (α_{AB}, α_{BA}) as well as activations from a single task (α_{AA}, α_{BB}) (Fig. 4). A value of 0.5 for α indicates that the representations are equally shared, with a value of 0 or 1 indicating that the representations are completely separate. Specifically, Equation 1 shows how the shared representation $\hat{x}$ is calculated at row i and column j by a cross-stitch unit that takes an input activation map, x:

$$\begin{bmatrix} \hat{x}_A^{i,j} \\ \hat{x}_B^{i,j} \end{bmatrix} = \begin{bmatrix} \alpha_{AA} & \alpha_{AB} \\ \alpha_{BA} & \alpha_{BB} \end{bmatrix} \begin{bmatrix} x_A^{i,j} \\ x_B^{i,j} \end{bmatrix} \quad (1)$$

The values of the weight matrix α are adjusted during backpropagation, with the partial derivatives easily calculatable as the cross-stitch units are modeled with linear combinations. We evaluate cross-stitch networks alongside a multi-task variant of fully connected feedforward neural networks in our modeling of student affect to investigate whether the level of connectivity within each architecture has an observable impact on the predictive performance of the affect models.

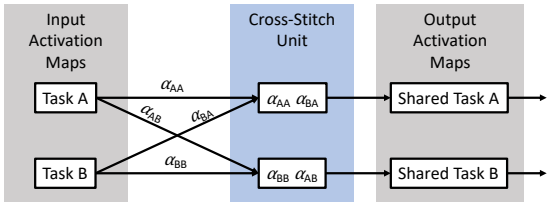

Fig. 4. Visualization of a cross-stitch network for weighting shared representations between task *A* and task *B*.

C. Posture-Based Feature Engineering

The features representing the posture data captured from the Kinect sensor are generated from three tracked vertices: *top_skull*, *center_shoulder*, and *head*. These vertices are selected based on prior literature that has investigated the effectiveness of the posture modality within affect recognition tasks [37]. Each posture-based feature is calculated based on the postural position and movement of each student that occurs within the 20-second observational window prior to each BROMP observation. Eighteen distinct features are generated for each vertex, including the most recent observed distance, minimum and maximum observed distance, median observed distance, variance in the observed distances, and most recent Z-coordinate value. In this work, "distance" refers to the Euclidean distance between each vertex and the Kinect sensor. In addition to these features, the minimum, maximum, median, and variance in the distance is calculated across the time windows of 5, 10, and 20 seconds that precede the corresponding BROMP observation. Several additional features were distilled that calculated the total change in the position (relative to the prior vertex's location in 3D coordinate space) and distance (relative to the prior vertex's distance from the Kinect sensor) across the preceding 3- and 20-second time windows. Using the median distance of the head vertex across the entire dataset, the final features were calculated to represent whether the student was leaning forward, backwards, and upright using the current position of the head vertex. These postural features were averaged across time windows of 5, 10, and 20 seconds, in addition to the entire gameplay session that had transpired prior to the current BROMP observation.

In addition to the spatial posture features, temporally based features were generated using the calculated distance between the (x, y, z) coordinates of two consecutive sensor readings from the *head* vertex. These delta values were used to generate velocity-based features averaged across time windows of the preceding 3, 5, 10, and 20 seconds prior to each BROMP observation. The mean, median, max, and variance of the calculated velocity values were used as features. Forty-eight new features were produced from this process. As a result of the high number of features generated from this process, the *center_shoulder* and *top_skull* vertices were not utilized for generating temporally based features.

D. Interaction-Based Feature Engineering

The interaction-based features are distilled from the generated log files that record each student's in-game actions and movements, and the condition of particular NPCs throughout the game [8]. Features that represent the condition of the NPCs that receive medical attention include the changes in systolic blood pressure, exposed wound type,

heart rate, and lung volume. Additional features were distilled that represented students' in-game actions such as performing a check of an NPC's vital signs and requesting an emergency medical evacuation. The interaction-based features were calculated across 20-second time intervals prior to each BROMP observation. Features were represented by using a summative count or averaging across the preceding time interval or were represented using statistical calculations such as standard deviation or median values for information such as a virtual patient's blood pressure. In total, thirty-nine interaction-based features were generated during this process.

E. Feature Selection

Due to the high number of features from the different modalities, feature selection was performed on each modality using forward feature selection. Forward feature selection iterates through a set of features in a greedy fashion beginning with a single feature and increasing the number of features according to their predictive performance on the target variable. This process iterates until a pre-determined threshold has been reached or until all features have been evaluated. However, due to the greedy search heuristic, the feature selection is weighted more heavily towards features that are evaluated earlier (e.g., the first feature evaluated is *always* selected). To mitigate any bias based on the arbitrary ordering of the feature candidates, we run 100 independent iterations of the forward feature selection. Each iteration uses a randomized feature ordering, and the features that are most frequently selected across all iterations are selected for training the affect recognition models. This approach provides a compromise between the speed of a greedy search heuristic and the computational cost of an exhaustive feature selection process [38]. Forward feature selection was performed on each modality separately, with the ten most predictive features per modality being combined at a feature level for training the affect recognition models.

F. Affect Model Evaluation

To evaluate the performance of the multi-task models (fully connected and cross-stitch networks), we train a series of single-task neural and non-neural baseline models in addition to several non-neural multi-task baseline models. The baseline models were *k*-nearest neighbor, elastic net, random forest, and feed-forward neural network. These were selected as baselines due to their capabilities of both single-task and multi-task learning. The single-task baseline models demonstrate the performance of models without any affective dynamics context, while the multi-task non-neural baseline models verify that the deep learning-based approaches (fully connected and cross-stitch networks) achieve higher performance with the affective dynamics context than non-neural multi-task models.

Each model was evaluated with nested ten-fold cross-validation, with each fold split at a student-level to prevent data leakage across the training, validation, and test sets. Within each outer cross-validation fold, the data were standardized to ensure a mean of zero and standard deviation of one prior to performing feature selection. Hyperparameter tuning was performed using three-fold cross-validation within the training data of the nested outer cross-validation. The hyperparameters evaluated were the number of nearest neighbors (*k*-nearest neighbors), ratio of *L1* and *L2* regularization (elastic net), number of estimators (random forest), and the number and size of the hidden layers (neural network). Each deep learning model's hidden layer used a hyperbolic tangent activation function due to the standardization of the data, as well as a dropout probability of 0.5 in the last hidden layer to mitigate potential overfitting. The loss function for the feedforward single-task network was binary cross entropy. Additionally, minority cloning was employed as an oversampling technique to resolve the class imbalance present within each affective state's dataset. This process clones each instance of the minority class until the class distribution is brought to a more uniform level.

An active line of investigation in MTL is determining the optimal distribution of the loss term across the different tasks. A common naïve approach to MTL loss is to assign a uniform weight to the loss term for each individual task t when calculating the summative loss term:

$$L_{total} = \sum_t W_t L_t \tag{2}$$

However, as the auxiliary tasks of predicting future affective states is distinguishable from the task of predicting the current affective state, we explore the use of a loss function that uses uncertainty weighting for each individual task [26]. The weight W_t for each task is determined by maximizing the log likelihood of an assumed multivariate Gaussian distribution. By optimizing for the model parameters θ and observation noise σ, the following loss function is derived:

$$L_{total} = \sum_t \frac{1}{2\sigma_t^2} L_t(\theta) + \log(\sigma_t) \tag{3}$$

In this way, optimizing for σ_t for each task t allows the relative weight of each task-specific loss function (i.e., the first term in Equation 3) to be learned from the data during the training process, while the second term in Equation 3 acts as a regularization term to prevent σ from increasing exponentially, which prohibits the model from learning. This allows the model to assign different weighted losses between the primary task (predicting the current affective state) and the secondary auxiliary tasks (predicting the subsequent affective state).

In addition to the single-task baseline model, four multi-task deep learning models were evaluated, uniformly weighted and uncertainty-weighted fully connected networks and cross-stitch networks. Each deep learning model was trained for 100 epochs, with early stopping implemented using the validation set and a patience of 10 epochs. Each network contained either two or three hidden layers with each layer containing either 8, 16, 32, or 64 nodes. For each cross-stitch model, each pair of hidden layers contained a cross-stitch unit. Data standardization, feature selection, and minority cloning occurred within each outer cross-validation iteration using the training folds to protect against data leakage across the validation and test folds. Each nested cross-validation fold was kept consistent across all evaluations to ensure fair comparisons between models.

V. Results and Discussion

While the five auxiliary predictive tasks (i.e., predicting the subsequent affective state) were utilized within the training process, the model evaluation focused exclusively on the predictive performance of the current affective state only. As a result, the predicted values of the auxiliary affective states are not considered in the results presented in this

TABLE I. MODEL EVALUATION RESULTS FOR SINGLE-TASK AND MULTI-TASK MODELS

Bored

Model Type	AUC	Accuracy	F1 Score
Single-Task	0.779	0.833	0.414
Fully Connected	**0.844**	**0.798**	**0.429**
Fully Connected (W)	0.839	0.848	0.414
Cross-Stitch	0.838	0.832	0.465
Cross-Stitch (W)	0.818	0.844	0.452

Confused

Model Type	AUC	Accuracy	F1 Score
Single-Task	0.546	0.570	0.301
Fully Connected	0.528	0.752	0.009
Fully Connected (W)	0.508	0.722	0.110
Cross-Stitch	**0.563**	**0.620**	**0.313**
Cross-Stitch (W)	0.561	0.614	0.314

Engaged Concentration

Model Type	AUC	Accuracy	F1 Score
Single-Task	0.586	0.554	0.589
Fully Connected	0.584	0.584	0.680
Fully Connected (W)	0.570	0.580	0.692
Cross-Stitch	**0.592**	**0.576**	**0.652**
Cross-Stitch (W)	0.561	0.586	0.666

Frustrated

Model Type	AUC	Accuracy	F1 Score
Single-Task	0.594	0.742	0.115
Fully Connected	0.537	0.478	0.046
Fully Connected (W)	0.582	0.360	0.083
Cross-Stitch	**0.602**	**0.800**	**0.095**
Cross-Stitch (W)	0.592	0.789	0.117

Surprised

Model Type	AUC	Accuracy	F1 Score
Single-Task	0.576	0.846	0.062
Fully Connected	0.523	0.702	0.055
Fully Connected (W)	0.507	0.624	0.078
Cross-Stitch	**0.646**	**0.622**	**0.099**
Cross-Stitch (W)	0.578	0.874	0.051

section. The primary evaluation metric is Area Under Curve (AUC), due to its ability to account for data imbalances. Additional evaluation metrics are the raw accuracy and the F1 score for each model. For each affective state (*bored, confused, engaged concentration, frustrated, surprised*), the highest performing single-task baseline models in addition to all multi-task model variants are shown in Table 1, with the best performing models in terms of AUC highlighted in bold.

The feedforward neural networks (FFNNs) outperformed all other non-neural baselines (single-task and multi-task) across all affective states. Of note is the fact that the uncertainty-weighted models (annotated with "(W)" in Table 1) and the cross-stitch models are all variations of FFNNs, therefore FFNNs will always be the "optimal" model. The inclusion of subsequent affective states as auxiliary outputs for multi-task modeling appeared to offer improved performance over the single-task baselines for all affective states. The uniformly weighted fully connected model was the highest performing affect model for *bored*, while the uniformly weighted cross-stitch model was the highest performing model for the remaining four affective states. Of note is the fact that, with the exception of *bored*, both variations of the fully connected model failed to outperform the single-task baselines, while the uniformly weighted cross-stitch network outperformed each baseline, and the uncertainty-weighted cross-stitch network outperformed the baseline in two affective states (*confused* and *surprised*). It did not appear that the uncertainty-weighted loss function improved performance significantly for either neural network model, achieving lower performance than almost every uniformly weighted counterpart. Although a multi-task model outperformed the single-task baseline for each affective state, the improvement was marginal at best for three affective states (*confused*, *engaged concentration*, and *frustrated*), with an incremental improvement of less than 0.02 in terms of AUC. The inclusion of the auxiliary affect sequence data showed improved performance for *surprised* and *bored*, which is surprising as the affect sequences for these two affective states typically transitioned to a state of *engaged concentration* (Fig. 3). This indicates that a student's behavior while *bored* or *surprised* may differ depending on if the student is about to transition to a state of *engaged concentration* vs. a state of *bored*, *confused*, or another uncommon affective transition.

To further investigate the impact of integrating temporal information into the multi-task affect recognition models, we include additional input features that represent the prior affective states exhibited by each student. This information is incorporated through five summative features representing the total number of observations of each of the five affective states prior to the current BROMP observation. The current BROMP observation is not included in these features as this would be a form of data leakage. This process is the natural next step in incorporating affective dynamics within each student model, so that each model can be induced using both the antecedent and subsequent affective states. This allows the model to be trained using bidirectional affective sequences. The same model architectures using the future affective states as auxiliary tasks for the multi-task models shown in Table 1 are re-evaluated while also incorporating the prior affective states as input features (Table 2). The same single-task baseline results are included in Tables 1 and 2 because the purpose of these models is to demonstrate the affect models' predictive performance without providing *any* affective dynamics context.

The addition of the preceding affect information into the input features increased the performance of the affect models for four affective states, with the exception of *surprised*. Additionally, the uncertainty-weighted loss function induced the highest performance for two affective states, *bored* and *confused*. Among the four affective states, the fully connected multi-task model was the highest performing model for one affective state, with the rest being modeled most effectively by cross-stitch networks. This provides further evidence for the enhancements offered by dynamically weighting the balance between the shared and task-specific representations within the multi-task models.

To investigate whether the improvements of the multi-task affective dynamic-based models are attributed to random chance, the results of the single-task baseline models are compared with the results of the optimal multi-task model for each affective state. The cross-validation results of the models were compared using a Wilcoxon signed-rank test, which is a non-parametric statistical test. This measure is used as the cross-validation results cannot be assumed to be

TABLE II. MODEL EVALUATION RESULTS FOR SINGLE-TASK AND MULTI-TASK MODELS

Bored

Model Type	AUC	Accuracy	F1 Score
Single-Task	0.779	0.833	0.414
Fully Connected	0.843	0.812	0.430
Fully Connected (W)	0.841	0.820	0.434
Cross-Stitch	0.854	0.811	0.435
Cross-Stitch (W)	**0.861**	**0.808**	**0.427**

Confused

Model Type	AUC	Accuracy	F1 Score
Single-Task	0.546	0.570	0.301
Fully Connected	0.559	0.618	0.323
Fully Connected (W)	0.551	0.613	0.304
Cross-Stitch	0.560	0.622	0.327
Cross-Stitch (W)	**0.614**	**0.635**	**0.368**

Engaged Concentration

Model Type	AUC	Accuracy	F1 Score
Single-Task	0.586	0.554	0.589
Fully Connected	0.597	0.619	0.685
Fully Connected (W)	0.606	0.601	0.658
Cross-Stitch	**0.627**	**0.604**	**0.676**
Cross-Stitch (W)	0.614	0.608	0.679

Frustrated

Model Type	AUC	Accuracy	F1 Score
Single-Task	0.594	0.742	0.115
Fully Connected	**0.665**	**0.600**	**0.073**
Fully Connected (W)	0.633	0.360	0.064
Cross-Stitch	0.572	0.784	0.057
Cross-Stitch (W)	0.648	0.780	0.123

Surprised

Model Type	AUC	Accuracy	F1 Score
Single-Task	**0.576**	**0.846**	**0.062**
Fully Connected	0.537	0.823	0.037
Fully Connected (W)	0.507	0.762	0.031
Cross-Stitch	0.548	0.685	0.070
Cross-Stitch (W)	0.526	0.831	0.047

normally distributed. Using a significance level of 0.05, the affective dynamics-based models were shown to demonstrate significant increases in performance for four affective states: *bored* (p=0.023), *confused* (p=0.018), *engaged concentration* (p=0.038), and *surprised* (p=0.038). Although the improvement of the affective models for *frustrated* was noticeable (0.071 AUC), it was not observed to be statistically significant (p=0.121).

There are limitations of this work that should be noted. The modeling of affective dynamics is reliant on having multiple labeled observations of affect for each student and is therefore incompatible with affective training sets that contain only a single label per student. Additionally, annotated occurrences of affect are sometimes removed in cases of inter-rater disagreement, which results in gaps in students' affective sequences throughout their learning sessions. While using summative or averaged feature representations can help mitigate this issue, this approach removes any semblance of temporal order between multiple affect labels. Therefore, an area of future research is to investigate the tradeoff between temporal context and model performance. For the purposes of this study, we utilize the annotated observations of affect to represent the prior affective states for the affective dynamics-based models. However, the models may not have access to annotated observations of a student's prior affective states in a run-time setting. In these instances, the models would require an alternative representation of prior affect, such as the model's prior predictions or confidence intervals for each affective state. Finally, the predictive performance of this work is likely dependent on the number of observations of each affective transition present in the dataset. For example, roughly 30% of the possible affective transitions in our dataset contained less than two observations. While these transitions are less likely to occur in a real-world setting, this likely impacts the overall generalizability of the models.

VI. CONCLUSION

Affect-sensitive learning environments that are designed to enhance student learning and improve student engagement could significantly benefit from accurate affect recognition models. In this work, we demonstrate the effectiveness of utilizing affect sequences to improve the predictive performance of multimodal affect recognition models through multi-task learning. Our approach allows models of students affect to capture patterns of student behavior based on unidirectional and bi-directional sequences of affect using auxiliary classification tasks and input feature engineering combined with multi-task models. Further, we investigate the use of different multi-task models in addition to uniformly weighted and uncertainty-weighted loss. Results indicate that the use of cross-stitch networks as a multi-task modeling technique leads to increased predictive accuracy for the majority of the affective states evaluated, and the use of affective dynamics context within the models increases the accuracy of all affective states. Additionally, further improvements to the models' performance can be induced through the use of bi-directional affective sequence data, in addition to uncertainty-weighted loss functions within the multi-task models.

The results suggest several promising directions for future research. First, the tradeoff between temporal information and model generalizability can be further explored by using different representations of the subsequent affective states instead of the one-hot encoding of a single state. Additionally, the generalizability of the affective dynamics-based multi-task modeling approach should be evaluated using different learning environments, student populations, and modalities. Our approach is dependent on the coding scheme used to annotate students' affective states in our dataset, so evaluating the performance of our models using different observational protocols or another method such as self-report would provide further insight into the impact that these factors have on our modeling approach. Finally, it is important to investigate the practicality of affective dynamics-based multi-task modeling through the integration of the models into a run-time setting, enabling user-adaptive features such as affect-sensitive feedback and guidance to improve learning outcomes and increase student engagement.

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

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

[3] R. Baker, S. D'Mello, Ma. M. Rodrigo, and A. Graesser, "Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments," *Int. J. Hum. Comput. Stud.*, vol. 68, no. 4, pp. 223–241, 2010.

[4] S. D'Mello and A. Graesser, "Dynamics of affective states during complex learning," *Learning and Instruction*, vol. 22, no. 2, pp. 145–157, 2012.

[5] K. Loderer, R. Pekrun, and J. Lester, "Beyond cold technology: A systematic review and meta-analysis on emotions in technology-based learning environments," *Learning and Instruction*, vol. 70, pp. 1–15, 2020.

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

[7] S. D'Mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surv.*, vol. 47, no. 3, p. 43:1-43:36, 2015.

[8] N. Henderson, J. Rowe, L. Paquette, R. Baker, and J. Lester, "Improving affect detection in game-based learning with multimodal data fusion," in *Proceedings of the 21st International Conference on Artificial Intelligence in Education*, 2020, pp. 228–239.

[9] J. DeFalco, J. Rowe, L. Paquette, V. Georgoulas-Sherry, K. Brawner, B. Mott, R. Baker, and J. 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.

[10] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Trans. Affect. Comput.*, vol. 7, no. 1, pp. 17–28, 2016.

[11] T. Krokotsch and R. Böck, "Generative adversarial networks and simulated+unsupervised learning in affect recognition from speech," in *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction*, 2019, pp. 28–34.

[12] W. Min, B. Mott, J. Rowe, R. Taylor, E. Wiebe, K. Boyer, and J. Lester, "Multimodal goal recognition in open-world digital games," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2017, pp. 80-86.

[13] Y. Zhang and Q. Yang, "Learning sparse task relations in multi-task learning," in *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017, pp. 2914–2920.

[14] C. Shui, M. Abbasi, L.-É. Robitaille, B. Wang, and C. Gagné, "A principled approach for learning task similarity in multitask learning," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 3446–3452.

[15] M. Geden, A. Emerson, J. Rowe, R. Azevedo, and J. Lester, "Predictive student modeling in educational games with multi-task learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 654–661.

[16] M. Long, Z. Cao, J. Wang, and P. Yu, "Learning multiple tasks with multilinear relationship networks," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 1593–1602.

[17] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, "Cross-stitch networks for multi-task learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3994–4003.

[18] M. Song, Z. Yang, A. Baird, E. Parada-Cabaleiro, Z. Zhang, Z. Zhao, and B. Schuller, "Audiovisual analysis for recognising frustration during game-play: introducing the multimodal game frustration database," in *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction*, 2019, pp. 517–523.

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

[20] E. Ghaleb, M. Popa, and S. Asteriadis, "Multimodal and temporal perception of audio-visual cues for emotion recognition," in *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction*, 2019, pp. 552–558.

[21] Y. Jiang, N. Bosch, R. Baker, L. Paquette, J. Ocumpaugh, J. M. A. L. Andres, A. Moore, and G. Biswas, "Expert feature-engineering vs. deep neural networks: Which is better for sensor-free affect detection?," in *Proceedings of the International Conference on Artificial Intelligence in Education*, 2018, pp. 198–211.

[22] J. M. A. L. Andres, J. Ocumpaugh, R. Baker, S. Slater, L. Paquette, Y. Jiang, S. Karumbaiah, N. Bosch, A. Munshi, A. Moore, and G. Biswas, "Affect sequences and learning in Betty's Brain," in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 2019, pp. 383–390.

[23] J. Ocumpaugh, R. Baker, S. Karumbaiah, S. A. Crossley, and M. Labrum, "Affective sequences and student actions within reasoning mind," in *Proceedings of the 21st International Conference on Artificial Intelligence in Education*, 2020, pp. 437–447.

[24] A. Botelho, R. Baker, J. Ocumpaugh, and N. Heffernan, "Studying affect dynamics and chronometry using sensor-free detectors," in *Proceedings of the 11th International Conference on Educational Data Mining,* 2018, pp. 157-166.

[25] A. Botelho, R. Baker, and N. Heffernan, "Improving sensor-free affect detection using deep learning," in *Proceedings of the 18th International Conference on Artificial Intelligence in Education*, 2017, pp. 40–51.

[26] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7482–7491.

[27] X. Liu, J. Gao, X. He, L. Deng, K. Duh, and Y.-Y. Wang, "Representation learning using multi-task deep neural networks for semantic classification and information retrieval," *NAACL*, 2015.

[28] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2014, pp. 3320–3328.

[29] S. Ruder, J. Bingel, I. Augenstein, and A. Søgaard, "Latent multi-task architecture learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 4822–4829.

[30] N. Henderson, V. Kumaran, W. Min, B. Mott, Z. Wu, D. Boulden, T. Lord, F. Reichsman, C. Dorsey, E. Wiebe, and J. Lester, "Enhancing student competency models for game-based learning with a hybrid stealth assessment framework," in *Proceedings of the 13th International Conference on Educational Data Mining*, 2020, pp. 92–103.

[31] R. Sawyer, J. Rowe, R. Azevedo, and J. Lester, "Modeling player engagement with bayesian hierarchical models," in *Proceedings of the 14th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2018, pp. 215–221.

[32] J. Ocumpaugh, R. Baker, and M. M. Rodrigo, "Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual," pp. 1–72, 2015.

[33] R. Baker, J. Ocumpaugh, and J. M. A. Andres, "BROMP Quantitative Field Observations: A Review," in *Learning Science: Theory, Research, and Practice*, New York, NY: McGraw-Hill, pp. 127–156, 2020.

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

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

[36] R. Caruana and V. De Sa, "Promoting poor features to supervisors: Some inputs work better as outputs," in *Proceedings of the 9th International Conference on Neural Information Processing Systems*, 1996, pp. 389–395.

[37] J. Grafsgaard, K. E. Boyer, E. Wiebe, and J. Lester, "Analyzing posture and affect in task-oriented tutoring," in *FLAIRS Conference*, 2012, pp. 438–443.

[38] N. Henderson, W. Min, A. Emerson, J. Rowe, S. Lee, J. Minogue, and J. Lester, "Early prediction of museum visitor engagement with multimodal adversarial domain adaptation," in *Proceedings of the Fourteenth International Conference on Educational Data Mining*, 2021, pp. 93–104.