Multimodal Machine Learning in Adaptive Instructional Systems: A Survey

Nathan Henderson, Jonathan Rowe, and James Lester

North Carolina State University

INTRODUCTION

The development and evaluation of multimodal machine learning approaches is an ongoing research area that has been the subject of growing interest in recent years. By emulating human sensory perception using multiple concurrent data channels, or “modalities,” multimodal machine learning has shown promise for a range of domains related to education and training, particularly in adaptive instructional systems (AISs). Multimodal machine learning has been shown to yield improved models compared to unimodal methods, particularly in the area of affect detection and classification (Baltrušaitis, Ahuja, & Morency, 2018; Grafsgaard et al., 2014).

Recent advances in sensor technologies have enabled a growing list of applications of multimodal machine learning to different sensor-based modalities, including eye gaze data, facial expression, speech, posture, gesture, electrodermal activity (EDA), and electroencephalography (EEG). These data streams are complementary to sensor-free modalities, such as keystroke date, mouse movements, and interaction trace log data. Multimodal machine learning has been used for tasks such as automated classification and identification of affective states, including frustration, boredom, and engagement. Multimodal systems have also been devised to induce computational models for assessment (Grafsgaard et al., 2014) and metacognition (Azevedo & Aleven, 2013). Data for training multimodal machine learning models in AISs has been collected in a number of different environments, including laboratory (Taub et al., 2017), classroom (Bosch et al., 2016), and military training settings (DeFalco et al., 2018).

Multimodal machine learning shows significant promise for enabling personalized support functionalities to enhance learning outcomes and engagement in AISs. However, multimodal AISs raise challenges as well. Sensors with high sample rates generate large volumes of data to be filtered and processed, raising issues of data storage, computational resources, scalability, and modality interdependence. Sensors that rely on external hardware can also break or fail, raising issues of noise, data loss, calibration, mistracking, and interference. The inclusion of multiple parallel data streams requires each modality to be aligned and represented in a way that is compatible with a chosen multimodal machine learning algorithm (Baltrušaitis, Ahuja, & Morency, 2018). Additionally, multimodal machine learning is ideally configured to take advantage of multi-dimensional information available across the various modalities; otherwise, a multimodal approach is unlikely to be any better than an ensemble of unimodal models.

In recent years, the Generalized Intelligent Framework for Tutoring (GIFT) has emerged as an important testbed for the development and deployment of AISs. GIFT is a service-oriented framework of software tools and methods designed to streamline the process of designing, developing, and deploying AISs. Notably, GIFT provides built-in support for collecting multimodal data during student interactions with an AIS. This is enabled by the GIFT Sensor Module, which provides a configurable interface to several hardware sensors, including webcams, motion-tracking cameras, and EDA bracelets. However, much of this support is focused on collection of multimodal data. Significant gaps exist in available tools and support for the development and implementation of multimodal machine learning systems for AISs, including tools for multimodal data preprocessing, modeling, and analysis. GIFT does include some integration with existing data mining toolkits, such as RapidMiner (Mierswa, Wurst, Klinkenberg, & Scholz, 2006).

However, significant programming effort is required to utilize these features, and there is limited support for many prominent machine learning toolkits.

In this paper, we provide an overview of recent research on applications of multimodal machine learning in AISs. We describe machine learning techniques that have been used in different multimodal AISs, as well as non-instructional systems that raise parallel challenges. We discuss issues such as data fusion, data imputation, and data alignment, along with relevant algorithms. Practical considerations for multimodal data collection and analysis are noted as well. Finally, we offer several suggested directions for future enhancements to GIFT to facilitate development and utilization of multimodal machine learning in AISs, especially those developed with GIFT authoring tools.

OVERVIEW OF MULTIMODAL MACHINE LEARNING

Multimodal machine learning has its roots in audio-visual speech recognition (Baltrušaitis, Ahuja, & Morency, 2018), but with recent advances in sensor technologies and computational resources, multimodal machine learning has expanded to a wide variety of roles. In AISs, a common application of multimodal machine learning is affect detection. Affect serves a key role in shaping learning outcomes (Grafsgaard, Wiggins, Boyer, Wiebe, & Lester, 2014). By devising computational models that dynamically measure learner affect at run-time, it is possible to detect and intervene in negative affective states, such as frustration and boredom, to enhance student learning outcomes (DeFalco et al., 2018; Harley, Bouchet, Hussain, Azevedo, & Calvo, 2015). Multimodal machine learning has also been utilized to predict positive affective states that correlate positively with student learning, such as engagement (Grafsgaard et al., 2014).

Sensor-Based Multimodal Affect Detection

Sensor-based multimodal systems have seen a significant increase in usage in recent years, primarily due to their inherent generalizability across a multitude of domains. This is attributable to rapid decreases in the cost and size of many sensors. Many types of sensors no longer require purchase of specialized hardware, and instead are widely available through universal platforms such as built-in webcams, microphones, eye trackers, and motion-tracking cameras like the Microsoft Kinect. Because these sensors are free of specialized hardware restrictions, they can offer a more cost-effective alternative to more expensive input channels. In a survey of multimodal affect detection systems, D’Mello & Kory (2014) observed a large number of affect detection models and detailed contemporary trends. They observed that facial expression and voice were the most commonly used modalities, occurring in over 75% of observed studies. They also noted that other sensor-based inputs such as posture, body movement, and other physiological modalities were individually present in at least 10% of studies (D’Mello & Kory, 2014).

We review a number of recent works involving sensor-based multimodal affect detection systems and observe their respective methodologies for classification, as well as their utilized modalities. Harley et al. (2015) captured multimodal data from 67 undergraduate students engaged in MetaTutor, an adaptive science-based learning environment. They captured facial expressions using a webcam in conjunction with automatic facial expression recognition software (FaceReader 5.0). They also measured physiological arousal using electrodermal activity (EDA) data using an Affectiva Q-Sensor bracelet and analyzed trends in the two modalities in conjunction with learners’ self-reported affective states. Their findings indicated high level of agreement between facial expression and affective states, but low correlation between the EDA modality and affective state (Harley et al., 2015). In a similar fashion, Cooper, Arroyo, and Woolf (2011) used EDA data alongside posture and facial tracking to investigate learner engagement with Wayang Outpost, a mathematics intelligent tutoring system. They also utilized a mouse sensor that captured grip pressure. They utilized stepwise linear regression to detect a series of affective states (confident, excited, interested, or frustrated). Grafsgaard et al. (2012) utilized data from a Microsoft Kinect sensor to identify frustration, focused attention, decreased involvement, and disengagement in students interacting with a computer-mediated tutoring system for introductory Java programming. Multiple postural features were correlated with different affective states and student learning outcomes. In general, it appeared that the more a user shifted their overall posture, the less engaged and more frustrated they were.

Facial expression and posture have been investigated alongside EDA data and mouse pressure in work by Arroyo et al. (2009) aimed at classifying learner confidence, frustration, excitement, and interest among students engaged with an AIS designed for teaching geometry (Arroyo et al., 2009). Using stepwise linear regression, their work indicated that increased mouse pressure was correlated with rising frustration levels, and facial expression was indicative of approximately 60% of students’ instances of affect. In addition to facial expression and posture data, Grafsgaard et al. (2014) analyzed textual dialogues between a tutoring system and student to analyze engagement, frustration, and normalized learning gains in students. Bosch et al. (2016) used clustering and Bayes Nets to construct binary classifiers for boredom, confusion, delight, engagement, and frustration. Their experiment took place in a classroom environment, where students were engaged with Physics Playground, an educational game about qualitative physics, using facial expression, head and torso positioning, and gross body movement as input modalities. Eye gaze tracking has also effectively been utilized as an indicator of learning outcomes, such as work by Rajendran, Carter, and Levin (2018) that used this modality to train a gradient tree boosting algorithm to model students’ reading performance.

Additional modalities have also been investigated for classifying learners’ affective states. However, the sensors required to capture these types of data are often more intrusive than facial expression or posture analysis sensors, such as webcams and motion-tracking cameras. Many multimodal systems that leverage biometric-based modalities have been devised for environments outsides of educational settings. EEG data has been used alongside Kinect data for biometric identification tasks using K-nearest neighbor clustering with histogram-oriented gradient features (Rahman & Gavrilova, 2017). EEG, EDA, and EMG modalities were modeled using Naïve Bayes classifiers, support vector machines (SVMs), and J48 decision trees for the purpose of identifying individuals’ levels of arousal and valence while watching online videos (Girardi, Lanubile, & Novielli, 2017). Results show that EEG and EDA yielded the highest classification rates for arousal when used with an SVM, while all three modalities produced the highest classification rate for valence. Similarly, EDA and EEG have been simultaneously utilized to detect stress levels and cognitive load among visually-impaired people navigating an unfamiliar environment (Kalimeri & Saitis, 2016). The classifier investigated in this experiment was a random forest model. Soleymani et al. explored the use of EEG data with facial expression data for the approximation of valence and arousal levels of students watching a series of emotion-invoking video clips. They used long short-term memory recurrent neural networks (LSTM-RNN) and continuous conditional random fields for their classification models. Their results indicated that facial expression data was inherently more informative than EEG data for their task, and a majority of the EEG features were a result of facial expression contamination. However, the EEG modality was beneficial when used in a complimentary role alongside the facial expression modality (Soleymani, Asghari-Esfeden, Fu, & Pantic, 2016).

Multimodal Deep Learning

Deep learning techniques, such as LSTM-RNNs, have seen a huge increase in interest in recent years, particularly due to significant improvements in computational hardware such as graphical processing units. In a survey paper focused on deep multimodal learning, Ramachandram and Taylor (2017) attribute increased interest in deep learning to its ability to form a hierarchical representation of each modality simultaneously, offering a distinct advantage over unimodal classifiers. Neural network architectures such as convolutional neural networks, autoencoders, LSTM-RNNs, and feedforward neural networks have also been shown to serve as effective multimodal machine learning models for tasks such as affect detection, sentiment analysis, image annotation, and speech classification (Ramachandram & Taylor, 2017). Common modalities for these solutions include audio-visual information, text, speech/dialogue data, and optical flow. Deep feedforward neural networks have been shown to yield improved performance in tasks such as frustration detection over non-neural models, such as SVMs (Henderson et al., 2019), which may be attributable to their innate ability to learn complex relationships across high-dimensional data as well as their capacity to process data while keeping spatial and temporal context intact (Pei, Yang, Jiang, & Sahli, 2015).

CHALLENGES IN MULTIMODAL MACHINE LEARNING

Although multimodal machine learning shows significant promise within AISs, there are still a multitude of risks and issues to be addressed, particularly when dealing with sensor-based models. Some concerns are raised when dealing with the dimensionality of the data itself. For example, multimodal systems often require high-dimensionality data, which can improve performance but also increase computational workloads and hardware constraints. Another issue is the spatial nature of many modalities. Although temporal information in different modalities has been demonstrated to be beneficial to multimodal classifiers (Henderson et al., 2019), many sensors only capture positional or spatial data at discrete time points; temporal context is not explicitly recorded. Other hardware-related issues can also arise, yielding significant noise or the loss of an entire data channel altogether. Because multimodal machine learning often involves the creation of a singular model for a multitude of data streams, it becomes imperative that each modality is equally accessible and interpretable. This raises issues of data alignment and representation, as well as calibration of the sensors themselves. In this section, we discuss several common issues and obstacles that are raised by multimodal machine learning-based systems, as well as different efforts to remedy several of these issues.

Temporal Context

As stated in the previous paragraph, sensors often capture limited temporal context about subjects. While this does not directly inhibit the creation or deployment of multimodal systems, such information can provide insight into the state of the subject captured by the sensor. For example, this was recently shown to be beneficial to the creation of a multimodal machine learning-based model for run-time affect detection in a game-based learning environment for emergency medical training (Henderson et al., 2019). A single modality containing spatial posture data (i.e., torso position) was captured using a Microsoft Kinect sensor. A second, synthetic modality that captured temporal data (i.e., torso velocity) was generated by taking the derivative of each captured instance of participants’ postural positions. This temporal information improved the performance of the affect detector over a previously published baseline that utilized only the spatial Kinect features. Another example of this approach took (1) body lean angle, (2) slouch factor, (3) quantity of motion, and (4) contraction index from a single postural modality (Sanghvi et al., 2011). These input vectors were utilized for the classification of elementary school students’ engagement with an automated companion in a game-based learning environment (iCat). These features served as artificial temporal modalities, an attempt to solve the issue of missing temporal context from sensor-based data. An alternative approach is to employ machine learning to derive temporal context from a multimodal dataset using techniques such as recurrent neural networks (Chen & Jin, 2015). Continuous generation models have also been used as a generative approach to preserve temporal information across modalities (Baltrušaitis, Ahuja, & Morency, 2018).

Data Preprocessing

There are several issues that arise during data preparation prior to the application of multimodal machine learning techniques. For example, a common issue in affect detection is the problem of imbalanced class labels. In many educational settings, a student is more likely to exhibit displays of concentration than frustration or surprise, which may adversely impact a classifier’s ability to accurately detect certain affective states. This calls for application of oversampling techniques to training data. One method to address this problem is minority cloning. This oversamples positive instances of a sparse affective state, bringing the data to a balance that is closer to an even ratio of positive and negative instances. A more sophisticated approach that is commonly used in affect detection is Synthetic Minority Over-sampling Technique (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). SMOTE generates synthetic samples from the minority class based on existing samples.

An important step in many practical applications of multimodal machine learning is feature selection and feature reduction. Because each modality can have high dimensionality, it is important to eliminate redundant or irrelevant features to save computational time and resources. There are range of feature selection algorithms that have been utilized to distill feature representations in AISs. One example is forward selection: this is a greedy selection algorithm that trains a model on each feature and selects the feature whose model returns the highest performance. This process continues until a preset number of features is reached. Other feature selection methods include univariate selection, tree-based feature selection, and removing features with low variance. A common approach to feature reduction is principal component analysis (PCA). PCA involves reducing a multivariate dataset to lower-dimensionality linearly correlated values called “principal components” (Chandrashekar & Sahin, 2014). Autoencoders have also seen use as a feature reduction method, and they are particularly useful due to their ability to decode the reduced data to its original representation (Jaques, Taylor, Sano, & Picard, 2017).

Data Imputation

Multimodal AISs, particularly sensor-based systems, face inherent risks associated with hardware failure. Physical hardware can be unreliable or inconsistent, leading to issues such as data noise, data loss, and outliers. A common consequence is missing data for one or more modalities. Another common issue is data noise within a modality. This can occur due to background activity captured by a sensor (i.e., someone walking in the background), or inconsistent behavior from the sensor itself. While a common solution is to discard data samples that contain missing or invalid data, an alternative approach is to impute missing values using the available samples. One simple approach to data imputation is *mean imputation*, which involves replacing each missing value with the mean of existing values for that particular feature. A more sophisticated method is the use of autoencoders to impute data (Jaques, Taylor, Sano, & Picard, 2017). This involves training an autoencoder with a subset of data that does not contain any missing values. This trained model is then used to approximate the missing values. This method can be effectively applied to sparse missing data, as well as entire missing modalities. Autoencoders have also been commonly used as a denoising technique for noisy or inconsistent data.

Data Fusion

Data fusion deals with the integration of data from multiple modalities for the purpose of classification or regression. Thus, it is a critical step in the creation of multimodal machine learning models (Baltrušaitis, Ahuja, & Morency, 2018). The majority of data fusion techniques are model-agnostic; that is, the data fusion is not reliant on a particular machine learning algorithm, and it occurs prior to, or after, classification or regression has taken place (Baltrušaitis, Ahuja, & Morency, 2018). These approaches are commonly divided into three categories: early fusion (feature-level), late fusion (decision-level fusion), and hybrid fusion.

Early fusion involves the concatenation of feature vectors from multiple modalities, and it occurs immediately after feature extraction. This is arguably the simplest data fusion method, since concatenation is a relatively straightforward operation and the method requires only a single machine learning model. Late fusion involves training a unimodal classifier for each modality and then fusing the resulting predictions from each classifier. Fusion can be accomplished in several possible ways, including averaging, voting, weighting, or applying another machine learning-based model (Baltrušaitis, Ahuja, & Morency, 2018). This approach does allow for different machine learning models to be used on each modality, which may increase overall model performance. This method is also more robust, allowing for models to be trained even in the presence of missing modalities. Hybrid fusion combines predictions from early fusion with additional unimodal predictors. Recent efforts to evaluate data fusion methods include work by Rahman and Gavrilova (2017), which used a form of late fusion on EEG and Kinect posture data for biometric identification. Similarly, Kalimeri and Saitis (2016) used early fusion on EEG and EDA modalities for the detection of stress levels, and Patwardhan and Knapp (2016) used a variety of body tracking modalities in combination with late fusion for the purpose of affect detection. Finally, Henderson et al. evaluated both early and late fusion techniques with a combination of spatial and temporal posture modalities for frustration detection in a game-based learning environment (Henderson et al., 2019).

There are also data fusion techniques that implicitly handle multimodal data; we refer to these as *model- based approaches* (Baltrušaitis, Ahuja, & Morency, 2018). One example of a model-based approach is multiple kernel learning models. These are an extension of SVMs, but apply the kernel-based learning approach to multiple modalities. Another alternative is probabilistic graphical models. Originally devised using hidden Markov models and Bayesian networks, probabilistic graphical models have expanded to include conditional random fields. Graphical models are useful due to their ability to process spatial and temporal features from multimodal data, and they often lead to interpretable models (Baltrušaitis, Ahuja, & Morency, 2018). As stated before, deep learning has become a prominent method for multimodal machine learning. This is partly due to its innate ability to fuse encoded features from multiple modalities. Deep neural networks offer the ability to learn complex relationships across high-dimensionality datasets, as well as the ability to process spatial and temporal information through the use of CNNs and RNNs, respectively.

Data Alignment

Multimodal data alignment is a process that accounts for the relationships between sub-components of two or more modalities (Baltrušaitis, Ahuja, & Morency, 2018). Data alignment is particularly relevant when multiple modalities operate or sample at differing frequencies. This task can be undertaken using either explicit or implicit alignment. Explicit alignment is an alignment process that is the primary objective of a modeling analysis. Implicit alignment is an intermediate step within a larger, overarching task. Implicit alignment often involves a latent representation of the separate modalities, and it often occurs during model training. As a result of this approach, the models do not explicitly align the data, but latently align the data during the training phase.

Explicit alignment aims to increase the correlation between two modalities’ components. This allocates increased emphasis on a similarity metric that evaluates the quality of the alignment between two or more modalities. One common approach is called dynamic time warping (DTW). DTW is a dynamic programming approach that can be applied to time-series datatemporal alignment is often a primary issue in multimodal datasetsparticularly when multiple sensors or input channels are involved. DTW computes the similarity between two modalities and inserts additional frames within the modalities to find an optimal match (Baltrušaitis, Ahuja, & Morency, 2018). This requires timesteps between the two modalities to be comparable and compatible with the given similarity metric. More recently, canonical correlation analysis has been used as a linear transformation serving as the similarity metric for DTW. This method allows DTW to discover linear relationships across multiple modalities in the temporal dimension, but it does not work well with non-linear relationships. Deep learning techniques have also been used for data alignment, particularly to measure similarity between modalities. However, because deep neural networks are typically utilized in supervised fashion, there is an implicit requirement for pre-aligned data to be used to train deep learning models. Often, datasets lack a subset of explicitly annotated data, restricting the utility of supervised data alignment techniques (Baltrušaitis, Ahuja, & Morency, 2018).

Implicit alignment is used for tasks where explicit alignment is either not useful or feasible, such as speech recognition or machine translation. Early work in implicit alignment involved the use of graphical models. However, the usage of this method has waned over time due to the need for manual construction of the graphical mapping between modalities and the need for previously-aligned training data. More recently, deep neural networks have become a primary method of implicit alignment. Often, this takes place through the use of autoencoder models as well as cross-modality retrieval models.

MULTIMODAL MACHINE LEARNING IN GIFT

GIFT has been used to develop and deploy AISs in a range of research studies (Aleven et al., 2018; DeFalco et al., 2018; Goldberg & Cannon-Bowers, 2015). Several studies have utilized GIFT’s Sensor Module to collect multimodal data. However, GIFT provides limited support for downstream analysis and modeling of multi-channel data using multimodal machine learning techniques. We offer several recommendations for potential enhancements to GIFT which would facilitate the development and deployment of AISs that leverage multimodal data streams. Currently, GIFT supports sensors for posture, gesture, facial expression analysis, EDA, and EEG. Additional data channels can be added to GIFT-based AISs by integrating additional hardware sensor types, such as electromyography and eye tracking. Further, sensor integration need not be restricted to a single sensor for each modality; it is conceivable that there would be benefits to supporting multiple sensors concurrently that focus on a single modality (i.e. multiple Microsoft Kinect sensors positioned at different locations around a learner).

Additional data preprocessing techniques could also be integrated into the GIFT Sensor Module to improve the preparation of data prior to it being sent to the Learner Module. One example is feature scaling, typically performed through data normalization or standardization, which is a step that is often necessary in machine learning analysis, particularly in deep learning. Providing solutions to address class label imbalances in recorded data could also prove useful, including support for algorithmic oversampling techniques such as SMOTE. Notably, this would only be applicable to recorded data that is labeled prior to oversampling. Another prospective enhancement is integrating feature selection and feature reduction techniques. This would enable GIFT to present the Learner Module with data that omits redundant or uninformative information that can be captured in raw sensor data. Examples include forward feature selection or elimination of features that fall below a pre-set variance threshold. Alternately, dimensionality reduction techniques, such as PCA, can be integrated as well. These techniques will decrease the amount of data processed by the Learner Module, thus reducing run-time computational requirements.

Another prospective improvement to GIFT would be the integration of data imputation methods. This would reduce the negative impact of missing data in GIFT’s analysis pipeline, and it also ensures the preservation of each data point in the raw sensor data. Simple imputation methods, such as mean imputation, ensure that missing or invalid data do not adversely impact the processing pipeline, and they do not require previously-labeled data or model training. More complex methods, such as the autoencoder-based methods, impute missing data more accurately, but they require pre-existing trained models. This renders the approach ineffective in instances where a modality is missing a majority of its data. However, the trained autoencoder can also be utilized to denoise the data, which can boost classifier performance.

Data fusion techniques could also be implemented in GIFT to aid the Learner Module in dealing with multiple data channels. Feature-based data fusion (Early Fusion) is programmatically simple to integrate as it requires modalities to be concatenated prior to being passed to other modules. Alternative feature-level fusion methods have been shown to offer improvement over simple feature concatenation. For example, performing feature selection on individual modalities prior to feature concatenation has been demonstrated to improve affect classification results (Henderson et al., 2019). Decision-level fusion is more complex to implement due to the need for a decision selection schematic, as well as the need for a machine learning model for each modality.

Expanded support of machine learning models and tools would also introduce to GIFT the capability to implement an entire multimodal data processing pipeline, including initial data capture, preprocessing, imputation, and modeling. Recent years have seen growing interest in deep learning-based models in AISs for a variety of learner modeling tasks, including run-time assessment and affect detection. Enhanced support for deep neural networks, including LSTM-RNNs, as well as other ML algorithms within GIFT would provide an expanded range of modeling options. It should be noted that the addition of these ML techniques stipulates a requirement for labeled training data, as well as computational resources for data- intensive deep learning algorithms, such as RNNs.

The most significant challenge to the integration of multimodal machine learning in AISs is handling disparate data streams. This is a common problem for multimodal systems deploying sensors operating at different sampling rates and within different time intervals. For modeling techniques such as data fusion to be possible, this issue must be addressed. We recommend two types of data alignment techniques to address this problem: explicit and implicit alignment. Temporal misalignment can be explicitly handled through DTW, although this method requires that the time axis between modalities be compatible with the similarity metric in the DTW algorithm. Additionally, non-linear relationships between modalities increases the difficulty of modality alignment. Although deep learning techniques have emerged as an effective approach to implicit alignment, it requires pre-labeled training data, which is usually not readily available in multimodal AISs involving disparate data streams. Data alignment continues to be widely researched and is an area of significant interest in the development of multimodal machine learning systems.

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

Multimodal machine learning will have a critical role in the design, development, and evaluation of AISs. Multimodal data streams enable the creation of data-rich models of student learning and engagement, which can be utilized to inform adaptive interventions to improve student outcomes. We have provided a survey of recent work on multimodal machine learning in AISs. We detail common machine learning techniques used to implement multimodal AISs, including key components of the multimodal data processing pipeline. Further, we detail specific challenges faced by developers of multimodal AISs, including common issues in data collection, alignment, and modeling.

GIFT has significant promise for facilitating future development of multimodal AISs. We recommend that future research and development efforts focus on integrating an expanded range of machine learning algorithms, addressing common issues raised by sensor-based AISs, and implementing solutions to data misalignment issues, particularly along the temporal dimension. By extending GIFT to include enriched multimodal machine learning capabilities, significant strides can be made to increase access to computational solutions for enhancing learner models and enabling adaptive pedagogical functionalities that improve learning outcomes and instructional effectiveness.

REFERENCES

Aleven, V., Sewall, J., Andres, J. M., Sottilare, R., Long, R., & Baker, R. (2018). Towards adapting to learners at scale: integrating MOOC and intelligent tutoring frameworks. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale* (Article no. 14). New York, NY: ACM.

Arroyo, I., Cooper, D. G., Burleson, W., Woolf, B. P., Muldner, K., & Christopherson, R. (2009). Emotion sensors go to school. In *Proceedings of the 14th International Conference on Artificial Intelligence In Education* (pp. 17– 24). <http://doi.org/10.3233/978-1-60750-028-5-17>

Azevedo, R., & Aleven, V. (2013). Metacognition and learning technologies: an overview of current interdisciplinary research. In *International handbook of metacognition and learning technologies.* (pp. 1–16). Springer, New York, NY. <http://doi.org/10.1007/978-1-4419-5546-3_47>

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

Bosch, N., Mello, S. K. D., Dame, N., Dame, N., Baker, R. S., Shute, V., … Zhao, W. (2016). Detecting student emotions in computer-enabled classrooms. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence* (pp. 4125–4129).

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16-28.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over- sampling technique. *Journal of Artificial Intelligence Research*, *16*(1), 321–357. <http://doi.org/10.1613/jair.953>

Chen, S., & Jin, Q. (2015). Multi-modal dimensional emotion recognition using recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge* (pp. 49–56). ACM. <http://doi.org/10.1145/2808196.2811638>

Cooper, D. G., Arroyo, I., & Woolf, B. P. (2011). Actionable affective processing for automatic tutor interventions.

In *New perspectives on affect and learning technologies* (pp. 127–140). New York, NY: Springer. <http://doi.org/10.1007/978-1-4419-9625-1>

D’Mello, S. K., & Kory, J. (2014). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, *47*(3), 43. <http://doi.org/10.1145/2682899>

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

Girardi, D., Lanubile, F., & Novielli, N. (2017). Emotion detection using noninvasive low cost sensors. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 125-130). IEEE.

Goldberg, B., & Cannon-Bowers, J. (2015). Feedback source modality effects on training outcomes in a serious game: Pedagogical agents make a difference. *Computers in Human Behavior*, *52*, 1-11.

Grafsgaard, J., Boyer, K., Wiebe, E., & Lester, J. (2012). Analyzing posture and affect in task-oriented tutoring. In *FLAIRS Conference* (pp. 438–443). Retrieved from <http://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS12/paper/download/4447/4843>

Grafsgaard, J. F., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., & 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* (pp. 122–129). London, UK: International Educational Data Mining Society. <http://doi.org/10.1182/blood-2013-10-529982.The>

Grafsgaard, J. F., Wiggins, J. B., Vail, A. K., Boyer, K. E., Wiebe, E. N., & 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* (pp. 42–49). ACM. <http://doi.org/10.1145/2663204.2663264>

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

Henderson, N. L., Rowe, J. P., Mott, B. W., Brawner, K., Baker, R. S., & 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)*.

Jaques, N., Taylor, S., Sano, A., & Picard, R. (2017). Multimodal Autoencoder : A Deep Learning Approach to Filling In Missing Sensor Data and Enabling Better Mood Prediction, 0–6.

Kalimeri, K., & Saitis, C. (2016). Exploring multimodal biosignal features for stress detection during indoor mobility. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (pp. 53–60). ACM. <http://doi.org/10.1145/2993148.2993159>

Mierswa, I., Wurst, M., Klinkenberg, R., & 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* (pp. 935–940). <http://doi.org/10.1080/14672715.1968.10405148>

Patwardhan, A., & Knapp, G. (2016). Multimodal affect recognition using Kinect. *ArXiv Preprint ArXiv:1607.02652*. Retrieved from <http://arxiv.org/abs/1607.02652>

Pei, E., Yang, L., Jiang, D., & 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)* (pp. 208–214). IEEE. <http://doi.org/10.1109/ACII.2015.7344573>

Rahman, W., & Gavrilova, M. L. (2017). Emerging EEG and Kinect face fusion for biometric identification. In

*Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1–8). IEEE. Rajendran, R., Carter, K. E., & Levin, D. T. (2018). Predicting Learning by Analyzing Eye-Gaze Data of Reading

Behavior. *International Educational Data Mining Society*.

Ramachandram, D., & Taylor, G. W. (2017). Deep Multimodal Learning: A Survey on Recent Advances and Trends. *IEEE Signal Processing Magazine*, *34*(6), 96–108. <http://doi.org/10.1109/MSP.2017.2738401>

Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P. W., & Paiva, A. (2011). 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* (pp. 305–312). ACM. <http://doi.org/10.1145/1957656.1957781>

Soleymani, M., Asghari-Esfeden, S., Fu, Y., & Pantic, M. (2016). Analysis of EEG Signals and Facial Expressions for Continuous Emotion Detection. *IEEE Transactions on Affective Computing*, *7*(1), 17–28. <http://doi.org/10.1109/TAFFC.2015.2436926>

Taub, M., Mudrick, N. V., Azevedo, R., Millar, G. C., Rowe, J., & Lester, J. (2017). Using multi-channel data with multi-level modeling to assess in-game performance during gameplay with CRYSTAL ISLAND. *Computers in Human Behavior*, *76*, 641–655. <http://doi.org/10.1016/j.chb.2017.01.038>

ABOUT THE AUTHORS

**Nathan Henderson** is a PhD student in Computer Science at North Carolina State University. He received his BS degree in Electrical and Computer Engineering from Auburn University and his MS degree in Computer Science from The University of Alabama in Huntsville. His research focuses on multimodal machine learning techniques for advanced learning technologies. Prior to graduate school, he was employed as a software developer for multiple U.S. Department of Defense contractors, where he supported several machine learning-related contracts.

**Dr. Jonathan Rowe** is a Research Scientist in the Center for Educational Informatics at North Carolina State University, as well as an Adjunct Assistant Professor in the Department of Computer Science. He received the PhD and MS degrees in Computer Science from North Carolina State University, and the BS degree in Computer Science from Lafayette College. His research focuses on the intersection of artificial intelligence and human-computer interaction for advanced learning technologies, with an emphasis on game-based learning environments, intelligent tutoring systems, multimodal analytics, learner modeling, and computational models of interactive narrative generation.

**Dr. James Lester** is Distinguished University Professor of Computer Science at North Carolina State University, where he is Director of the Center for Educational Informatics. His research centers on transforming education with technology-rich learning environments. With a focus on adaptive learning technologies, his research spans intelligent tutoring systems, game-based learning environments, affective computing, and tutorial dialogue. The adaptive learning environments he and his colleagues develop have been used by thousands of students in K-12 classrooms. He received his PhD in Computer Science from the University of Texas at Austin in 1994. He is a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI).