A Multimodal Approach for Real-Time Engagement Monitoring in E-Learning Using Machine Learning

Rohan Shankar
Incoming Freshman / Outgoing Senior
Cornell University / Mountain View High School
Ithaca, NY, USA / Mountain View, CA, USA
ORCID: 0009-0005-2328-4549

***Abstract*—This full research paper describes a multimodal approach to detect student engagement in an e-learning scenario. Remote learning has transformed access to education and helped reshape learning by offering students expanded access to resources, while also enabling educators to reach several new demographics. It has rightly been identified as one of the Grand Challenges for improving life on our planet. Unfortunately, gauging learner engagement has remained a significant challenge, limiting improvements and optimization of teaching methods. Machine Learning (ML), despite being an underutilized tool in the domain of e-learning, holds substantial potential for the real-time recognition of students' behavioral patterns, thereby helping provide personalized feedback to educators.**

This research paper aims to explore the potential of ML to compare multiple modalities such as facial action units (AUs), gaze vectors, and emotion indicators. AUs and gaze vectors are derived from the Facial Action Coding System (FACS), a comprehensive system developed to represent human facial movements. Detection of emotions is accomplished by employing a Convolutional Neural Network (CNN) that has been pre-trained on the AffectNet image database. A deep neural network model that combines these multiple modalities is finally used to estimate student engagement. Its performance is evaluated on the University of Massachusetts Amherst / Boston University student engagement dataset. The performance of the neural network model is characterized over a number of epochs to improve its bias-variance performance, thereby optimizing the model's overall accuracy. The performance of each modality is first explored. We then demonstrate the effectiveness of the multimodal approach, showcasing its potential.

This paper first describes the current state of the art research into student engagement detection in e-learning settings, and concludes by outlining recommendations for future work in this up and coming area of study. This paper helps further our understanding of how suitable tools can be developed to detect student behaviors, and thereby ultimately improving the effectiveness of digital learning experiences in the context of remote learning.

Keywords—machine learning, educational technology, learning management systems, gaze tracking, emotion recognition

INTRODUCTION

The advent of e-learning platforms has transformed the way knowledge is shared and acquired. E-learning has enabled accessibility, flexibility, and scalability. It enables learners to engage with educational content at their own pace, and in conditions that suit them. However, one big drawback is the lack of a two-way feedback loop between the teacher and the student. In cases where it is important to ensure critical content is grasped correctly, say for safety and compliance training, it is not possible to monitor the student's attention and engagement on a real-time basis. The only mechanism is to grade a test with questions on a subset of topics. It is far more proactive and beneficial to have a real-time ability to detect, measure, and enhance engagement of students consuming the e-learning content [1, 2]. Engagement detection will also allow educators to measure the effectiveness and quality of their teaching style and revise / experiment with them more freely [3].

E-learning environments have unique challenges and obstacles. Learners often operate in isolation, without the physical presence of instructors, fellow peers, or a classroom [4]. As a result, traditional methods of assessing engagement, which involve in-person observation of body-language and two-way interaction with students are not possible [2, 3].

In 2008, the National Academy of Engineering (NAE) identified 14 Grand Challenges that are considered to be game-changing goals for improving life on our planet. These challenges fall into four cross-cutting themes: Sustainability, Health, Security, and Joy of Living. Quite rightly, "Advance Personalized Learning" is one of these key Grand Challenges [5].

ML can be a powerful tool to address challenges in assessing e-learning engagement - specifically, real-time monitoring, pattern recognition, and personalization. ML algorithms can monitor each student in real-time [6, 7]: while a teacher may only be able to focus on one student or small groups, ML can facilitate engagement with and ensure oversight of all students. ML algorithms can potentially pick up subtle nuances in facial expressions that humans could miss,

and so there is an opportunity for a more meaningful engagement than previously possible. Finally, a targeted algorithm can be personalized to each student, and patterns between students can be compared to each other [7], almost like a personal tutor.

In the context of e-learning, engagement is often defined as the extent to which learners are actively involved, motivated, and focused on the learning process and content [1]. Engagement goes beyond mere looking at a screen. It involves participation, and encompasses emotional, cognitive, and behavioral aspects of the learner's interaction with the educational material. Engaged learners are more likely to absorb and retain information, leading to enhanced learning outcomes.

One way to utilize the measurement of engagement would be to improve classroom lessons. For example, when faced with a particularly challenging concept, a teacher could choose to explain it using either a video, or a live demonstration. Based on the engagement results for each approach, the teacher could emphasize one method or the other for future lectures.

How to measure engagement itself is a very challenging topic. Initial research in this area has relied on using gaze as a key indicator - i.e. whether the student is looking at the screen or not. However, the topic of attention and interaction is much more complex, and simple situations where the student is in deep thought and looking out of the window, or if a student is looking at educational content on a second monitor that is not positioned next to the primary camera, or if the student looks down at a notebook to take notes, can easily break gaze-based simplistic assumptions. Researchers have therefore begun to embrace a multimodal approach that integrates various indicators - not just gaze [8]. Each indicator can offer unique insights into the complex interactions that together make up engagement[9]. This form of improved understanding of human interaction also holds a strong potential for the upcoming field of humanoid robotics as well.

In this paper, an approach that uses multiple modalities to detect engagement is presented. Additionally, the effectiveness of these modalities is tested on a new, prominent engagement dataset within the e-learning context. The research goal is to discover if AUs, gaze vectors, and emotions are accurate methods to detect student engagement.

I. Methodology

In this study, three modalities (AUs, Gaze, and Emotion) are tested on a prominent engagement dataset that was developed by observing behaviors during recorded e-learning sessions.

Four different modalities were analyzed to understand their effectiveness: AUs-only, Gaze-only, Emotion-only and all Combined. The individual modalities are first introduced. Then the philosophy of training models effectively by optimizing the number of epochs is introduced.

A. Dataset

For this study, the following models were tested on the core Student Engagement Dataset [9, 19]. This dataset of 3400 images has been collected and labeled by students at the University of Massachusetts and is being hosted by Boston University. It captures images of consenting college students solving math problems in an online learning environment and is annotated to indicate whether students' attention is engaged or not engaged. The authors also emphasize the importance of student engagement in online learning and highlight the correlation between engagement, emotion, and learning gains.

The dataset itself contains data in video and image formats. For this study, only the image formats are used, since the focus is not on any temporal analysis (the use of video feeds for temporal analysis is left as future work).

B. Action Units (AUs)

Facial Action Units, otherwise known as AUs, are a popular modality used in emotion and engagement detection. A Facial Action Coding System (FACS) was described in a seminal manual by Eckman and Friesen in 1978 [10] and more recently revised in 2002. It is based on facial muscle groups, and so is anatomically accurate in its description of facial expressions. These AUs offer a window into a learner's emotional states by detecting facial expressions [11]. ML algorithms can then accurately interpret these expressions, unveiling subtle cues that may go unnoticed by human observers [12]. AUs can enable real-time assessment of emotional responses contributing to the understanding of learner reactions, and are displayed in Fig. 2.

An AU based classifier is implemented for its range of features. The list of AUs include the shape of eyes, eyelids, brow, nose, cheek, lips and head. Typical reactions such as winking, squinting, furrowing, pursuing, pressing, raising, tilting etc. are measured with two metrics - the presence or absence of them (1 or 0), and a continuum of intensity of their expression (ranging from 1 - lightly to 5 - highly intense) [13-18]. The wide variety of features detected by AUs has the potential to detect hidden trends in the data that other classifiers (with a more specific focus) may not have tracked.

C. Gaze Vectors

Gaze is another, more direct, tool that can be used for understanding student engagement by observing the student's pose and eye-focus area. Gaze vectors estimate periods of high engagement by noting that attention is being focused on the screen, while disengagement is measured by noting wandering or glazed eyes.

It is worth noting that Gaze vectors are limited to their direction vectors only, without the "intensity" nuance that AUs offer. However, they serve an important purpose due to the nature of the dataset used (i.e. the BU dataset [9, 19]). The BU dataset classified images for engagement level by the direction in which the subject was looking (i.e. either towards a screen or away from it).

D. Emotion

Engagement is also closely related to emotional attachment to a topic [1]. Emotions can be measured in two separate methods. The first of these methods is the 8 basic emotions- Anger, Contempt, Disgust, Fear, Happy, Neutral, Sad, and

Surprise. A second method represents human emotion by using the arousal-valence model - arousal or intensity of the response (ranging from calm or low intensity, to excited or high intensity), and valence or pleasantness of response (ranging from negative to positive) [24]. Both techniques can be used to measure the expressiveness of the student to determine the underlying sentiment, which is then used to detect engagement, in keeping with the definition of emotional engagement attachment.

E. Multimodal Analysis

Finally, a pipeline with all three combined modalities: emotion, AUs, and gaze, is implemented to discover any specific patterns that exist amongst them. Rather than each modality working individually, it is theorized that specific patterns may emerge when they are combined. Each modality's underlying characteristics (described in sections B-D) are preserved, creating a model that analyzes a multimodal fusion of features.

F. Model Training

In addition to modalities utilized, the number of epochs each model runs on is varied. Epochs refer to the number of training passes the model runs through the training dataset. If a model passes through the training dataset once, it has run for 1 epoch. It logically follows that while a higher number of epochs will take more training time, it will also allow the ML model to establish stronger trends amongst the data. However, if the number of epochs is too high, it will result in the model overfitting the data. This study implemented a bias-variance optimization to minimize both the bias and the variance, over the corresponding number of epochs.

Bias refers to the systematic error that arises when a model makes simplified assumptions or overlooks important patterns in the data. This oversimplification often occurs at low epoch counts, and the model is likely to underfit the data. Variance, on the other hand, refers to a model's sensitivity to fluctuations or noise in the training data. Models with high variance typically perform exceptionally well on data they have seen before, but struggle with testing data that is somewhat new. High variance is usually correlated to a model being overfit as well.

Generally, both parameters have an inverse relationship: a model with high variance usually implies low bias, and vice versa. The aim is to find an optimized number of epochs that minimize the total error from both parameters. This optimization, depicted in Fig. 1, is done by measuring the bias and variance of the model at several different epoch numbers.

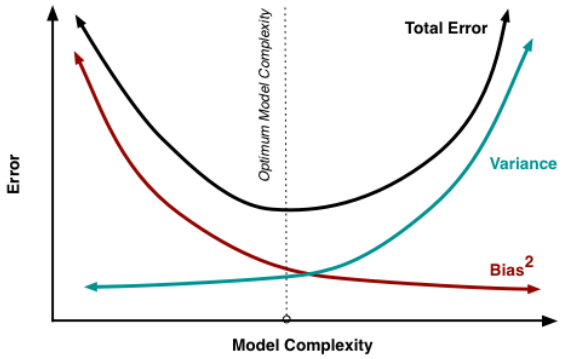

Fig. 1. Bias-Variance curve, and subsequent optimization, [20]

II. Experimental Procedure

Four different modalities are analyzed to understand their effectiveness to predict engagement. AUs-only, Gaze-only, Emotion-only, and all Combined. In this section, the proposed multimodal model for engagement detection that utilizes AUs, Gaze Data, and Emotion recognition is explained. First, the details of how each modality was extracted from the dataset is explained. Then, their processing for the model is explained.

A. Additional Datasets Used

The AffectNet dataset [24, 25], displayed in Fig.3 is used to train the emotion detection model. AffectNet is an extensive repository of 1M images of facial expressions found in the wild. Of these, a subset of 440,000 images have been manually labeled for the presence of eight discrete facial expressions (neutral, happy, angry, sad, fear, surprise, disgust, contempt) along with their intensity of valence and arousal. It is one of the largest such databases available today thansks to the efforts of the University of Denver.

B. Software Tools Utilized

Several tools and libraries form the backbone of modern ML research today, facilitating the development and implementation of robust models. Python, and the TensorFlow library were both used to code the models used in this study.

Openface was used for the extraction of Gaze data, and AUs. OpenFace [14] is an open-source toolkit specifically designed for comprehensive facial behavior analysis. Leveraging advanced algorithms, it excels in detecting facial landmarks, recognizing AUs, estimating head poses, and tracking gaze vectors from analyzing images and video frames Fig. 3.

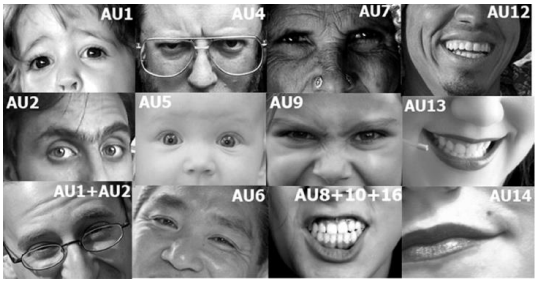

Fig. 2. Facial Action Coding System (FACS) is a way to classify and encode movements of individual facial muscles making it possible to represent any anatomically possible facial expression, and deconstructing it into the specific AUs [13].

C. Machine Learning Models Utilized

Two ML models were utilized for the research, namely SVMs, and CNNs. This section will detail how each function, as well as their specific usage.

A Support Vector Machine (SVM) is a powerful ML algorithm primarily used for classification tasks. It is a particularly popular technique for affective computing applications. In this study, SVMs are used to process data extracted through OpenFace, such as AUs and Gaze Data, to predict engagement, for the combined model.

A Convolutional Neural Network (CNN) is a useful neural network architecture built for processing images. It is a popular and robust model for image classification. CNNs were used to extract emotions from the BU Student Engagement Dataset.

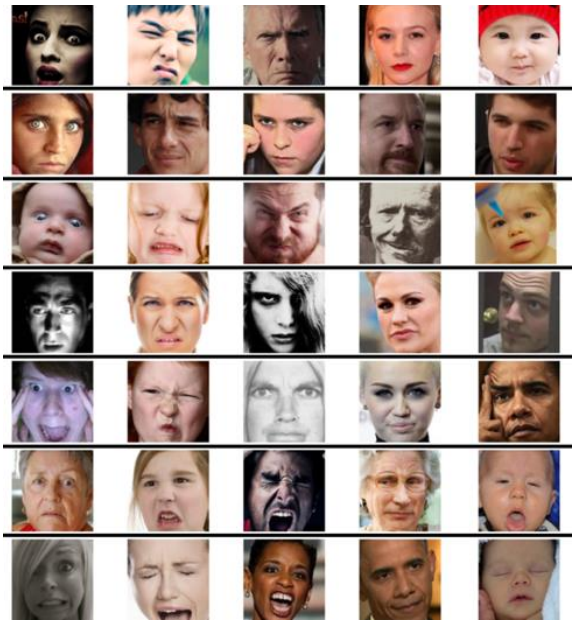

Fig. 3. AffectNet contains about 1M facial images collected from the Internet by querying three major search engines using 1250 emotion related keywords in six different languages. About half of the retrieved images (~440K) have been manually annotated for presence and intensity of emotions [24].

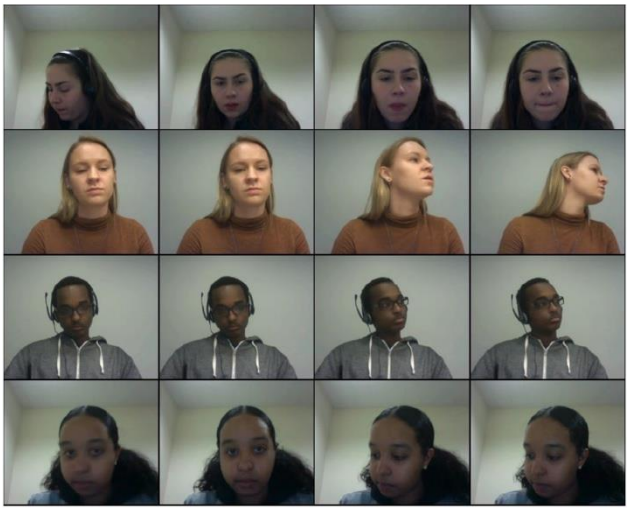

Fig. 4. The Student Engagement dataset, collected by the University of Massachusetts Amherst, captures student participants solving math problems on MathSpring, an online mathematics learning software tool.

D. Experimental Procedure

In this study, a CNN model is trained first on the AffectNet image dataset [24, 25], one of the most prominent emotion-labeled image datasets available. Then this model is used to extract emotions from the Student Engagement Dataset. In parallel, OpenFace is used to extract AUs and gaze vectors from the Student Engagement Dataset as well. Now, with emotions, gaze vectors, and AUs extracted from the Student Engagement Dataset, several different ML models are trained on the extracted data: one with emotions only, another with gaze vectors only, another with AUs only, and a final model with all 3 proxies combined. Now, the accuracy of each of the models can be compared in order to determine which are the most effective in predicting the student engagement labels.

III. Experimental Results

In this section, the accuracy of each of the models (Gaze, AUs, Emotion, and all Combined) are compared, in order to determine which is the most effective in predicting the student engagement labels.

A. Emotion Classifications

Emotion prediction is a very challenging area of research and the most recent state of the art (SOTA) model, POSTER++, can still only achieve 63.77% accuracy [26, 27]. In fact, even trained human labelers agreed on their manual emotion classifications only 60.7% of the time [25].

To minimize training and test time for development, a random subset of the larger labeled AffectNet image database i.e. 6% images (26,138 images) of the 440,000 original dataset was extracted. Of these 80% of images (20,914 images) were used for training, and 20% of images (5,224 images) were used for testing. A simple 5-layer CNN model to identify emotions was architected and then subsequently trained using the subset

of Affectnet images. The CNN model was composed of 3 hidden Convolution 2D layers each incorporating a ReLu activation function and a MaxPooling 2D layer, and 2 output Dense layers with a ReLu and SoftMax activation functions. The output layers also included a 50% dropout to prevent the model from overfitting the data.

A confusion matrix showing results of the model is shown below in Fig. 5. This simple model tested with 58.6% accuracy (training accuracy was 61.0%), which compares respectfully to the SOTA POSTER ++ [26,27] model with 63.8% accuracy. The labels in the confusion matrix are Anger, Contempt, Disgust, Fear, Happiness, Neutrality, Sadness, and Surprise.

As can be seen, the model struggles to correctly classify Disgust and Fear emotions (only 14% and 16% accurate respectively). Disgust is more often mis-classified as Anger, while Fear is more often mis-classified as Surprise. Regardless of these limitations, the overall performance is quite reasonable.

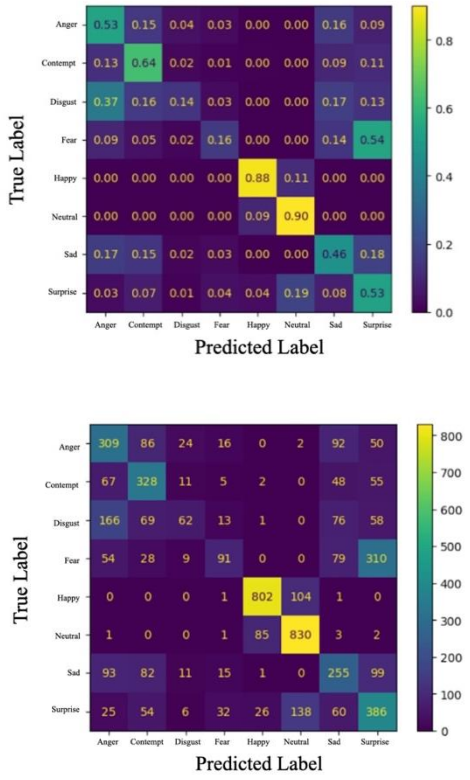

Fig. 5. Confusion Matrix showing degree of match between True (labeled) emotion and Predicted (by author's model). Figure on bottom shows the raw count of images classified, while the figure on the top shows the same information normalized to the total number of True (labeled) classifications.

TABLE I. EMOTION CORRELATION TO ENGAGEMENT LEVEL

Dataset Label	Detected Emotion				
	Fear	***Neutral***	***Sad***	***Surprise***	***Total***
Engaged	69	59	661	37	**826**
Not Engaged	30	13	462	4	**509**
Total	**99**	**72**	**1123**	**41**	**1335**

Dataset Label	Detected Emotion				
	Fear	***Neutral***	***Sad***	***Surprise***	***Total***
Engaged	8.4%	7.1%	80.0%	4.5%	**100.0%**
Not Engaged	5.9%	2.6%	90.8%	0.8%	**100.0%**
Total	**7.4%**	**5.4%**	**84.1%**	**3.1%**	**100.0%**

Results from applying Emotion Classifier model to labeled images in Student Engagement Dataset. The top table shows the raw image counts, while the bottom table shows the % split for each label category

When the pre-trained emotion classifier model was applied to the Student Engagement Dataset, a surprisingly large number of images seemed to be predicted to be Sad. In addition, there did not seem to be a clear distinction amongst the mix of emotions detected in the two Dataset images categories (Engaged or looking at the screen, vs Not Engaged or wandering and looking away from screen):

When utilizing the fully connected CNN utilizing these detected Emotions to predict engagement levels, the results are quite poor. Model accuracy and errors do not change with the number of Epochs. In fact, at a 1000 epochs, the Test Accuracy is higher than Train Accuracy indicating that the model performance has peaked, and further training will not be able to improve its performance any further.

TABLE II. PERFORMANCE VS # OF EPOCHS FOR MODEL USING EMOTIONS TO PREDICT ENGAGEMENT

Measured Metric	Epoch Level			
	10 Epochs	***100 Epochs***	***300 Epochs***	***1000 Epochs***
Test Accuracy	64.0%	61.4%	59.6%	62.9%
Train Accuracy	61.3%	62.0%	62.5%	61.6%
Bias Error	38.7%	38.0%	37.6%	38.4%
Variance Error	-2.7%	0.6%	2.9%	-1.3%
Bias + Variance	36.0%	38.6%	40.5%	37.1%

B. Gaze and AU Based Classification

The Gaze vectors and AUs (extracted by OpenFace) are used to train a Deep Neural Network to predict engagement levels of the Student Engagement Dataset. The number of epochs in the model are varied from 10 to 1000 to examine its

impact of training and test accuracy and resulting bias and variance errors.

TABLE III. PERFORMANCE VS # OF EPOCHS FOR MODEL USING GAZE VECTORS TO PREDICT ENGAGEMENT

Measured Metric	Epoch Level			
	10 Epochs	***100 Epochs***	***300 Epochs***	***1000 Epochs***
Test Accuracy	81.8%	82.5%	85.9%	82.9%
Train Accuracy	81.9%	84.5%	89.1%	100.0%
Bias Error	18.1%	15.5%	10.9%	0.0%
Variance Error	0.1%	2.0%	3.3%	17.1%
Bias + Variance	18.2%	17.5%	14.1%	17.1%

As can be observed from Table III, as the number of Epochs increases, Training accuracy steadily increases (and thereby Bias errors steadily decrease). However, Testing accuracy first increases (indicating improved model performance) and then decreases (indicating the model has become overfit to its data) and thereby the Variance Errors increase significantly. An optimal # of epochs then is likely somewhere between 300 and 1000 epochs.

In contrast, AUs show a different behavior. As the number of epochs increase, Bias errors decrease and stabilize, while Variance errors increase slightly before also stabilizing. An excellent Test accuracy of 94.8% is achieved after 1000 epochs, and it is possible that a slightly higher number of epochs can improve the model performance further.

TABLE IV. PERFORMANCE VS # OF EPOCHS FOR MODEL USING AUS TO PREDICT ENGAGEMENT

Measured Metric	Epoch Level			
	10 Epochs	***100 Epochs***	***300 Epochs***	***1000 Epochs***
Test Accuracy	92.6%	93.3%	94.1%	94.8%
Train Accuracy	94.8%	99.4%	99.5%	99.4%
Bias Error	5.2%	0.6%	0.5%	0.6%
Variance Error	2.2%	6.1%	5.5%	4.6%
Bias + Variance	7.4%	6.7%	6.0%	5.2%

C. Multimodal Classifications

Finally, a Deep Neural Network model that combines emotion, gaze and AUs is trained on the data, and its performance examined vs that of individual models.

Both AUs and the "Combination" model perform very well in predicting engagement of the student. A Gaze-only model is also quite effective, though not as well-performing as AUs only or the Combined model performance. Emotion-only performance falls quite behind.

TABLE V. PERFORMANCE WITH 1000 EPOCHS FOR MODEL PREDICTING ENGAGEMENT USING INDIVIDUAL (EMOTION, GAZE, AUS) AND COMBINED VARIABLES

Measured Metric	Model Used			
	Emotion	***Gaze***	***AUs***	***Combined***
Test Accuracy	62.9%	82.9%	94.8%	94.8%
Train Accuracy	61.6%	100.0%	99.4%	99.7%
Bias Error	38.4%	0.0%	0.6%	0.3%
Variance Error	-1.3%	17.1%	4.6%	4.9%
Bias + Variance	37.1%	17.1%	5.2%	5.2%

IV. DISCUSSION

Ultimately, one of the major drawbacks in the available Student Engagement Dataset is that its labeling is primarily focused on use of student gaze as the primary measure for determining engagement. A more realistic learning scenario will need to consider multiple other parameters to determine true student engagement.

This study shows even with this initial labeling, both AUs only and a model that incorporates AUs in addition to other parameters can be effective in predicting student engagement. AUs enable algorithms to incorporate facial expressions into understanding student engagement, making them quite powerful. This finding underscores the importance of including AUs in future studies of student engagement. One can hypothesize that a Combined model incorporating multiple modalities including AUs will be more effective in realistic learning environments. This multimodal approach can align better with the complex nature of true human engagement, allowing for a more comprehensive analysis.

V. CONCLUSION

This paper reviewed the importance of assessing student engagement in e-learning settings. Various ML techniques were surveyed and software tools used to study this topic were presented. A multimodal approach to engagement detection was introduced and its results on the BU

Student Engagement Dataset, showed the effectiveness of a multimodal approach. In spite of some limitations in its labeling, the usefulness of AUs in predicting engagement was successfully demonstrated. Emotion detection was challenged due to the difficulty of state-of-the-art algorithms. A multi-modality approach (that could still utilize emotions) can still be a powerful technique in realistic e-learning settings, but a more representative and better labeled dataset is needed to be able to assess its efficacy.

Predicting student engagement ultimately holds the power to improve education and learning techniques.

VI. Limitations and Future Work

The current landscape in engagement assessment has a fundamental challenge: limited availability of properly labeled and un-biased datasets. The Student Engagement Dataset, while being an excellent start, is limited in focusing too heavily on gaze-based indicators, and thereby overlooking a broader spectrum of engagement modalities. The limitation restricts proper study in effectiveness of models that can rely on nuanced indicators like emotional states or intricate facial AUs to better predict level of student engagement.

Future work will necessitate the development of comprehensive datasets that encompass diverse modalities beyond gaze. These datasets should encapsulate behavioral, cognitive, and affective elements, fostering a more accurate and holistic understanding of engagement levels. A more comprehensive and well-labeled e-Learning dataset can aid further research in this important area.

An extension of this research can involve K12 students. The current dataset only contains college students, limiting its applicability to younger populations. Additional enhancements can incorporate student performance metrics such as test scores and course grades to observed engagement levels. This correlation can provide a more quantitative measurement of each student's success, rather than the qualitative observations taken in this study.

An additional area of research can be to incorporate temporal analysis/indicators -- such as from the usage of videos. This can ultimately be more effective than only relying on images for engagement prediction, as they are mere snapshots in time.

Acknowledgment

I'd like to thank Dr. Marwa Mahmoud, Lecturer (Assistant Professor) at the University of Glasgow, for her time and assistance as my research mentor, and for our invaluable research discussions involving the role of machine learning in emotion recognition.

I'd also like to thank Cambridge Center for International Research (CCIR) for their incredible support in facilitating this 1-1 research mentorship experience.

I'd finally like to thank Dr. Kobad Bugwadia at Mathnasium for his insights, and for providing me with a real-life opportunity to teach and observe students as their Mathematics Instructor.

References

[1] Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., Movellan, J.R.: The faces of engagement: Automatic recognition of student engagement from facial expressions. IEEE Transactions on Affective Computing. 5, 1, 86–98 (2014).

[2] Zhou, W., Cheng, J., Lei, X., Benes, B., Adamo, N. (2020). Deep Learning-Based Emotion Recognition from Real-Time Videos. In: Kurosu, M. (eds) Human-Computer Interaction. Multimodal and Natural Interaction. HCII 2020. Lecture Notes in Computer Science(), vol 12182. Springer, Cham. https://doi.org/10.1007/978-3-030-49062-1_22

[3] Bahreini, K., Nadolski, R., & Westera, W. (2016). Towards multimodal emotion recognition in E-learning environments. Interactive LearnIng Environments, 24(3), 590-605. https://doi.org/10.1080/10494820.2014.908927

[4] Rößler, J.; Sun, J.; Gloor, P. Reducing Videoconferencing Fatigue through Facial Emotion Recognition. *Future Internet* **2021**, *13*, 126. https://doi.org/10.3390/fi13050126

[5] "Advance Personalized Learning." *Grand Challenges - Advance Personalized Learning*, http://engineeringchallenges.org/challenges/learning.aspx. Accessed 22 Dec. 2023.

[6] Chen, X., Zou, D., Xie, H., Cheng, G., & Liu, C. (2022). Two Decades of Artificial Intelligence in Education: Contributors, Collaborations, Research Topics, Challenges, and Future Directions. Educational Technology & Society, 25 (1), 28-47.

[7] Nezami, O.M., Dras, M., Hamey, L., Richards, D., Wan, S., & Paris, C. (2018). Automatic Recognition of Student Engagement Using Deep Learning and Facial Expression. *ECML/PKDD*.

[8] S. L. Happy, A. Dasgupta, P. Patnaik and A. Routray, "Automated Alertness and Emotion Detection for Empathic Feedback during e-Learning," 2013 IEEE Fifth International Conference on Technology for Education (t4e 2013), Kharagpur, India, 2013, pp. 47-50, doi: 10.1109/T4E.2013.19.

[9] K. Deglado, J. M. Origgi, T. Hansapoor, H. Yu, D. Allessio, I. Arroyo, W. Lee, M. Betke, B. Woolf, S. A. Bargal. "Student Engagement Dataset". ICCVW, 2021.

[10] P. Ekman and W. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto, CA, 1978.

[11] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In Proceedings of theIEEE/CVF Conference on Computer Vision and PatternRecognition, pages 7680–7689, 2021.

[12] T. Baltrušaitis, M. Mahmoud and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic Action Unit detection," 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 2015, pp. 1-6, doi: 10.1109/FG.2015.7284869.

[13] TadasBaltrusaitis Action units. In: GitHub. https://github.com/TadasBaltrusaitis/OpenFace/wiki/Action-Units. Accessed 22 Dec 2023

[14] OpenFace 2.0: Facial Behavior Analysis Toolkit Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency, IEEE International Conference on Automatic Face and Gesture Recognition, 2018

[15] Convolutional experts constrained local model for facial landmark detection A. Zadeh, T. Baltrušaitis, and Louis-Philippe Morency. Computer Vision and Pattern Recognition Workshops, 2017

[16] Constrained Local Neural Fields for robust facial landmark detection in the wild Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. in IEEE Int. Conference on Computer Vision Workshops, 300 Faces in-the-Wild Challenge, 2013.

[17] Rendering of Eyes for Eye-Shape Registration and Gaze Estimation Erroll Wood, Tadas Baltrušaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling in IEEE International Conference on Computer Vision (ICCV), 2015

[18] Cross-dataset learning and person-specific normalisation for automatic Action Unit detection Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson in Facial Expression Recognition and Analysis Challenge, IEEE International Conference on Automatic Face and Gesture Recognition, 2015

[19] N. Ruiz, H. Yu, D. Allessio, M. Jalal, A. Joshi, T. Murray, J. Magee, J. Whitehill, V. Ablavsky, I. Arroyo, B. Woolf, S. Sclaroff, M. Betke. "Leveraging Affect Transfer Learning for Behavior Prediction in an Intelligent Tutoring System" FG, 2021.

[20] Fortmann-Roe, S. (2012). Understanding the bias-variance tradeoff. https://api.semanticscholar.org/CorpusID:61975862

[21] Sharmin, Nusrat & Saif, Sarwar. (2019). Undergraduate Thesis: Automatic Human Brain Segmentation using Machine Learning Techniques. 10.13140/RG.2.2.35711.74404.
[22] Bre, Facundo & Gimenez, Juan & Fachinotti, Víctor. (2017). Prediction of wind pressure coefficients on building surfaces using Artificial Neural Networks. Energy and Buildings. 158. 10.1016/j.enbuild.2017.11.045.
[23] Kaplanoglou, Pantelis. (2017). Content-Based Image Retrieval using Deep Learning. 10.13140/RG.2.2.29510.16967.
[24] AffectNet. In: Mohammad H. Mahoor, PhD. http://mohammadmahoor.com/affectnet/. Accessed 22 Dec 2023
[25] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor, "AffectNet: A New Database for Facial Expression, Valence, and Arousal Computation in the Wild", IEEE Transactions on Affective Computing, 2017.
[26] Mao, J., Xu, R., Yin, X., Chang, Y., Nie, B., & Huang, A. (2023). POSTER V2: A simpler and stronger facial expression recognition network. *ArXiv, abs/2301.12149*.
[27] Papers with code - Affectnet benchmark (facial expression recognition (FER)). In: The latest in Machine Learning. https://paperswithcode.com/sota/facial-expression-recognition-on-affectnet. Accessed 22 Dec 2023a