Engagement Intensity Detection in an E-learning Setting: Factors Affecting Effectiveness and Efficiency

by

Name: Chia-Yu Chen
Student-Nr.: S158800

Name: Tomasz Tadeusz Rogoz
Student-Nr.: S136267

Exam: Data Mining, Machine Learning, and Deep Learning (CDSCO1004E) - Oral exam based on written product (IC)

Study: MSc Business Administration and Data Science
Date: 19. May 2023
Pages: 15
Characters incl. spaces: 25,389
Dataset: https://drive.google.com/file/d/1yrk_wyhZ-c7q0Mcyyi888ylFkl_JDELi/view

Engagement Intensity Detection in an E-learning Setting: Factors Affecting Effectiveness and Efficiency

Authors

Chia-Yu Chen (158800)

Tomasz Tadeusz Rogoz (136267)

{chch22ag, taro19ad}@student.cbs.dk

Students - Copenhagen Business School

MSc. Business Administration and Data Science

Abstract

Due to the lack of interaction between teachers and students in the e-Learning environment, engagement detection is a critical part of measuring the efficiency of online education. Most recent research focuses on complex multi-model methods to increase model accuracy based on spatial-temporal features extracted from facial expression data. However, body language also plays a role in emotional expression. In this paper, we propose the 3D CNN model, which utilizes spatial-temporal data including body movement. We choose a low resolution to emphasize the importance of body language. The results show a 56% accuracy rate on the DAiSEE dataset, a publicly accessible video dataset measuring student engagement, with less than 1% of all available data being fed into the learning algorithm. This result justifies the hypothesis that by considering body language, the model can achieve a competitive accuracy rate with lower computational cost.

Keywords: Engagement Intensity Classification, 3D CNN, Spatial-Temporal Analysis, Computation Cost, Body Language Analysis, Multiclass Classification

1. Introduction

Due to advancements in technology, e-Learning has become widely adopted in university education, skill development, and other domains. In comparison to in-person teaching methods, e-Learning lectures lack direct interaction with students, making it difficult to assess the suitability of course content and the actual quality of online courses.

Traditional online survey methods suffer from low participation rates, and students may hide their true emotions or have a low willingness to answer surveys (D'Mello et al., 2014; Dewan et al., 2019). However, machine learning technologies can help institutions analyze students' reactions without interrupting their learning, providing a more accurate and cost-effective way to interpret education quality (Dewan et al., 2019).

According to D'Mello et al. (2017), human emotions are dynamic. Analyzing students' real feelings based on still images taken in class without considering the temporal aspect of emotions may not be sufficient. Moreover, body language, such as head position, gestures, and body movements, also contributes to emotional expression. Therefore, this research focuses on classifying subjects' engagement based on spatial-temporal data. Our goal is to determine if including body language features in videos can achieve competitive results with a lower computational cost under the limitations of computational resources.

2. Literature Review

In this literature review section, we focus on research that predicts engagement intensity based on the DAiSEE dataset to gain methodological inspiration.

Gupta et al. (2016) proposed DAiSEE, a dataset used to benchmark engagement intensity prediction. They employed a set of CNN-based architectures, including ImpectNet V3, which relies solely on spatial features, and the 3-dimensional CNN (C3D) model, which performs

spatial-temporal analysis. The accuracy rates achieved were 46.6% and 48.6%, respectively.

Liao et al. (2021) proposed the Deep Facial Spatiotemporal Network (DFSTN), a CNN-based architecture that incorporates spatial and temporal feature extraction, noise removal, and dataset imbalance adjustment. They used a 1/15 sampling rate and doubled the number of video samples to increase the training instances. Their model, which utilized Multi-Task Convolutional Neural Network (MTCNN) and pre-trained SENet for spatial feature extraction, achieved an accuracy rate of 58.84%. However, the GALN and doubled sampling led to overfitting issues.

Abedi and Khan (2021) proposed a hybrid neural network architecture that combines the Residual Network (ResNet) and Temporal Convolutional Network (TCN). The ResNet extracts spatial data, while the TCN extracts temporal features. To address the issue of data distribution imbalance, Abedi and Khan introduced custom sampling strategies to ensure that each class is represented in the training, test, and validation sets. They also applied the CE (Cross Entropy) loss optimizer to handle the imbalance problem. The model achieved an accuracy rate of 63.9%. However, despite the inclusion of custom sampling strategies and weighted loss, the recall rate for the 'Low' intensity classification was just over 10%.

Mehta et al. (2022) proposed the DenseNet self-attention Neural Network, which is a 3D CNN model-based architecture that combines a self-attention layer to enhance the features that represent students' affective states. The paper employed a 1/10 sampling rate and a face region resolution of 244x244, cropped by the Dlib face detector. In terms of noise removal, the first ten frames of the video were excluded from the sampling range, as students might be distracted at the beginning. Additionally, frames with facial occlusions were removed as noise. The pre-processed data of size 30 frames x 244 height x 244 width x 3 channels were fed into the 3D CNN to extract spatial-temporal relationships. The self-attention layer was then used to enhance

deep learning by focusing on more relevant features (spatial, temporal, or hybrid) within the frames. Finally, the CE, Class-Balanced Cross-Entropy (CB-CE), and Class-Balanced Focal loss (CB-FL) were employed to address the data imbalance problem in the dataset. The results indicated that the DenseNet combined with the Hybrid self-attention layer and CB-FL slightly improved the accuracy rate to 63.59%. However, the probability of correctly predicting 'Very low' and 'Low' engagement levels was still less than 20% in the DAiSEE dataset.

In the literature review, we observed that weighted loss approaches and more complex models that enhance relevant features do not effectively address the dataset imbalance issue but increase the computational cost during model training. Additionally, not all the data we need to analyze may suffer from data imbalance problems in practical scenarios. Therefore, instead of proposing a more complex model or weighted adjustment methods, we propose incorporating body position and language, which are often overlooked aspects of emotional expression. Moreover, to emphasize the importance of body language, we can reduce the frame resolution, thereby increasing computational efficiency and reducing the training cost of the engagement classifier.

3. Conceptual Framework

3.1 Data preparation

The data preparation in this research consists of four main stages: video transformation with frame sampling, image transformation, sequence aggregation, and data separation (see Figure 1). Since our data is presented in the 30-fps video format, the video needs to be transformed into a sequence for further analysis. The preparation starts with video transformation, which includes the extraction and sampling of frames from each video. Then, image transformation is applied to reduce the computational requirements. Firstly, the frames are transformed into grayscale, which reduces the number of channels. Secondly, to focus more on body movement,

the resolution is reduced, which also discards some detailed facial patterns. The images extracted from the same video are aggregated into a 30x20x11x1 (DxHxWxC) sequence data and a 30x40x22x1 (DxHxWxC) sequence data to fit the input of the 3D CNN. Here, D refers to the number of frames in an instance, H refers to the height of the image, W refers to the width of the image, and C refers to the number of channels. Finally, the full dataset in a DxHxWxC format is separated into a training set, validation set, and test set in a ratio of 64:16:20.

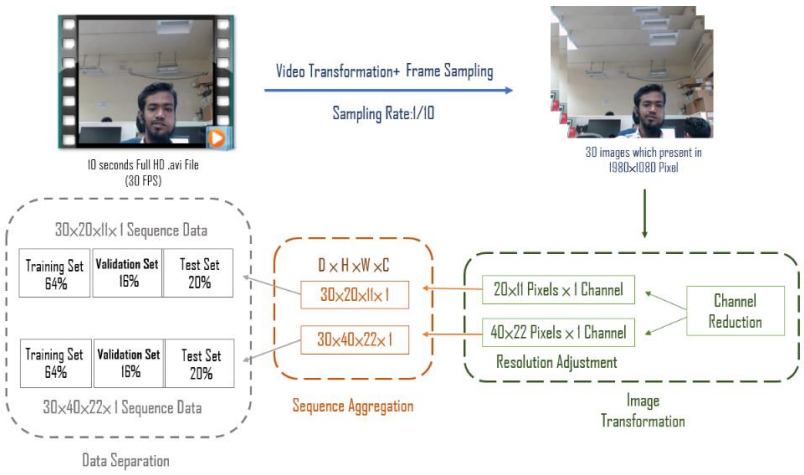


Figure 1 Data Preparation Process

3.2 Training Strategy

Since the DAiSEE dataset is the only publicly accessible dataset that simulates multiple learning illuminations, learning environments, and participants from different countries and genders, the 3D CNN model is fully trained using the DAiSEE dataset to allow the algorithm to learn these patterns. The data is first split into a training and test set using an 80:20 ratio with the model further splitting the training data fed into it into a training and validation set using the same ratio. Additionally, higher-resolution data (40×22) and lower-resolution data (20×11) are trained separately to compare the effects on computation cost and accuracy. Both resolutions are trained using batches of sizes 16, 32, and 64 with the higher resolution data additionally trained with batch sizes of 128 and 256. Different batch sizes are tested as the smaller ones introduce more randomness into the training process while larger ones estimate the gradient of

the loss function more accurately. Both have advantages and disadvantages associated with escaping local minima, reducing the impact of outliers or overfitting that are difficult to evaluate in advance and, therefore, require testing. Results are presented in terms of a classification report and a confusion matrix as well as a display of total training time.

In addition to comparing the data with different resolutions and training with different batch sizes, this paper will also compare the accuracy rate with the accuracy rate obtained by other researchers that are also based on the DAiSEE dataset.

4. Methodology

4.1 Dataset Description

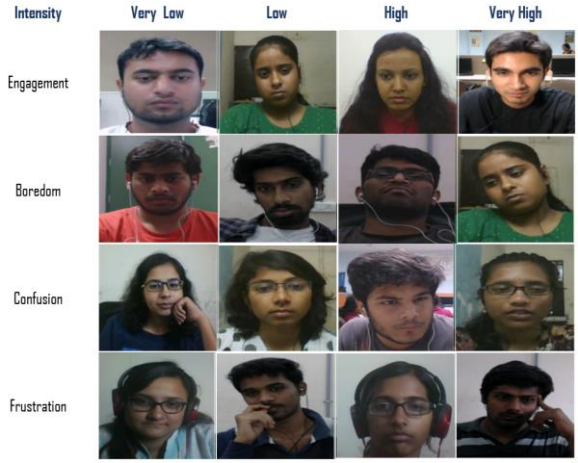


Figure 2 The example of different affective states and their intensity. In this example we can also see the different learning environments and illuminations and participants presented in the example. From Gupta et al. (2016).

The algorithm in this research is trained using the publicly accessible video dataset called DAiSEE, which aims to help analyze students' experiences with online learning. The dataset contains 9,068 different videos with 112 different subjects. To compare the emotional expressions between e-learning and normal video watching, each subject in the video is presented with two different videos: one with educational content and another with recreational

content. Since the classroom environment for students may vary, the dataset includes six different types of locations and three different levels of brightness to simulate data collected in e-learning institutions (see Figure 2). To increase the coverage of students attending online education, the dataset includes subjects of both genders, ranging in age from 18 to 30, and from different Asian countries.

The dataset is presented in 10-second Full HD (1920x1080 pixels) and 30 frames per second video format. Unlike most datasets that classify emotions into anger, joy, fear, sadness, surprise, and disgust, the DAiSEE dataset focuses on boredom, engagement, confusion, and frustration, which are less commonly used but more useful for understanding students' attitudes towards education quality. Each video has been annotated with four emotions and four different intensity levels (1) very low, (2) low, (3) high, and (4) very high (Gupta et al., 2016). The distribution of each category of data can be checked in Table 1. To ensure the accuracy of data annotation, the data has been annotated via crowdsourcing and compared with professional suggestions for verification.(Gupta et al., 2016)

Table 1 The distribution of different categories.

	Engagement (%)	Bored (%)	Confused (%)	Frustration (%)
Very High	4071 (44.89%)	334 (3.86%)	101 (1.11%)	87 (0.96%)
High	4477 (49.37%)	1934 (21.33%)	752 (8.29%)	346 (3.82%)
Low	459 (5.06%)	2931 (32.32%)	2191 (24.16%)	1649 (18.18%)
Very Low	61 (0.67%)	3869 (42.67%)	6024 (66.43%)	6986 (77.04%)

According to the table, we can see the sample distribution is imbalance in all the four categories. Especially under the confused and frustration label, most samples are labelled as the 1 (‘Very Low’). From Gupta et al. (2016).

4.2 Data Pre-processing

4.2.1 Video Transformation

We combine the sampling with frame extraction since this operation can reduce the extraction workload by 90%. Inspired by the work of Ai et al. (2022), which also applied the 3D CNN-based model, the sampling rate is set as 1/10. However, since Gupta et al. (2016) do not mention whether videos are snipped when the emotion is at its highest intensity or if they start at the point when people start to express their emotion, we choose not to remove the frames at the beginning and end of the video.

4.2.2 Image Transformation

Due to the constraint of computational resources, the 3-channel, RGB representation has been transformed into a 1-channel grayscale expression. Since we want our model to take body language into account, such as head position, gestures, and body movement, we reduce the resolution rate to 20x11 as body language detection suffers less from a reduced resolution than facial expression detection. Additionally, we also experiment with a resolution rate of 40x22 to investigate the improvement rate resulting from an increase in resolution.

4.3 3-Dimensional Convolutional Neural Network

Compared with traditional 2D CNNs, which can only process the height and width of an image, 3D CNNs can also capture temporal features (Tran et al., 2015). Since the engagement intensity is presented as a spatial-temporal variant sequence, the 3-Dimensional CNN, which efficiently learns both spatial and temporal relationships, has been introduced (Mehta et al., 2022; Tran et al., 2015).

Due to the amount of data fed in the model, we decided to use the 3D CNN with 2 COV3D layers, 2 MaxPooling3D layers and 2 Dense (full connected) layers. The kernel size selection is based on Tran et al. (2015) which suggest 3×3×3 kernel size performs the best on the video

multiclass classification task. The pool size selection is based on our data input. Additionally, inspired by Mehta et al. (2022), ReLU activation function is chosen to introduce the non-linearity.

Due to the large amount of data fed into the model, we decided to use a 3D CNN with 2 COV3D layers, 2 MaxPooling3D layers, and 2 Dense (fully connected) layers. The selection of kernel size is based on Tran et al. (2015), which suggests that a 3×3×3 kernel size performs the best for video multiclass classification tasks. The choice of pool size is based on our data input. Additionally, inspired by Mehta et al. (2022), the ReLU activation function is chosen to introduce non-linearity.

Once the data is fed into the model, the COV3D layers extract relevant spatial-temporal features, and then the MaxPooling3D layers extract the most important features. After passing through these layers, the output image cubes are flattened into a 1D array in preparation for further classification. This 1D feature vector is then fed into two dense layers, with the first layer using ReLU to learn more complex relationships, and the second dense layer using the Softmax function to produce the probability distribution of the four intensity classes. Finally, the model is compiled with the CE loss and Adam optimizer (Kingma & Ba, 2015) to optimize its performance on the multiclass classification task.

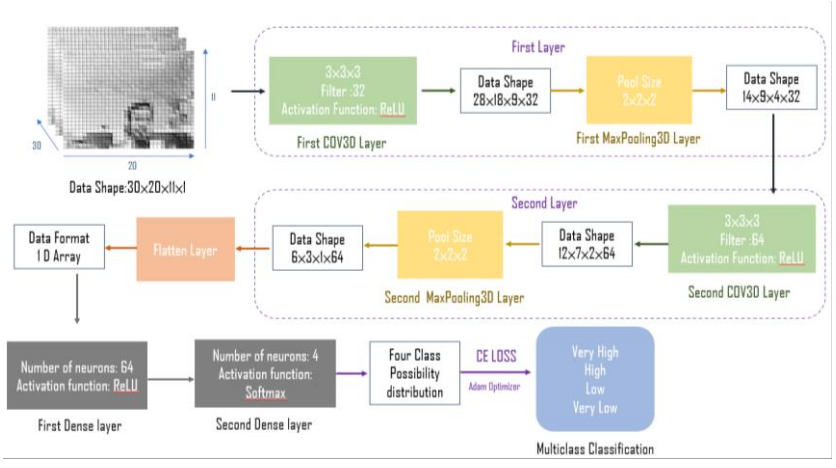


Figure 3 Framework of 3D CNN modal (take 20×11 data as an example)

4.4 Performance Evaluation

4.4.1 Evaluation Metric

Although accuracy rate is commonly used as the primary measure of model performance, it only focuses on the correct predictions without considering the reasons behind the model's decisions or the specific types of errors it makes. This comparison becomes particularly risky in imbalanced datasets where over 80% of instances are labeled as 'High engagement' or 'Very High Engagement', as the model could simply guess these classes to achieve a high accuracy rate. Therefore, precision, recall rate, and F1 score have been introduced to evaluate the model's performance, considering false positive and false negative predictions. The functions below present the calculations of precision, recall, and F1 score.

$$Precision = \frac{\#\ of\ True\ Positive}{\#\ of\ True\ Positive + \#\ of\ False\ Positive}$$

$$Recall = \frac{\#\ of\ True\ Positive}{\#\ of\ True\ Positive + \#\ of\ False\ Negative}$$

$$F1\ Score = \frac{\#\ of\ Correct\ Predicted\ Instances}{\#\ of\ Total\ Instance}$$

The F1 score combines precision and recall rates with equal weight. A lower score indicates a higher possibility of high accuracy based on uneven guessing (predictions concentrated on a specific class). Low precision and recall can reveal the preference of the model for a particular class.

4.4.2 Confusion Matrix

Since the model we applied is aimed at conducting multiclass classification, precision and recall rates can only indicate if the ratio of false positives and false negatives in the classification task is high or not. However, which classes are most frequently misclassified as another class, or whether the algorithm simply guesses the class with the highest number, should be interpreted using the confusion matrix. Therefore, in this paper, the accuracy and evaluation metrics will

be examined first, and then the confusion matrix will be used to investigate the model.

5. Results

Method		Image Size (HxWxC)	Accuracy	
ER System	(Dresvyanskiy et al., 2021)	224x224x3	39%	
InceptionNet V3	(Gupta et al., 2016)	229x229x3	46%	Benchmark
3D CNN	(Gupta et al., 2016)	229x229x3	48%	Benchmark
ResNet + TCN with sampling and weighted loss	(Abedi & Khan, 2021)	224x224x3	54%	
3D CNN with body language [40x22]	(Proposed)	**20x11x1**	55%	
3D CNN with body language [20x11]	(Proposed)	**40x22x1**	56%	
LRCN	(Gupta et al., 2016)	229x229x3	58%	
DFSTN	(Liao et al., 2021)	224x224x3	59%	
3D DenseNet + SA + FC	(Mehta et al., 2022)	224x224x3	**62.15%**	

Table 1 Model Accuracy Comparison with other works

To verify the effectiveness of our model, we compared it with other related works. Gupta et al. (2016) provided five different benchmarks for further research. We selected the most basic video classification model, the 3D CNN model, which is the same as our model. We also considered the LRCN model, which achieved the highest accuracy rate. The results listed in the table show that even though we only used 0.2% of the data they used, our model still performed 12% and 8% better in terms of accuracy compared to the benchmark model. We also included the ResNet + TCN model with sampling and weighted loss proposed by Abedi & Khan (2021). This work also used a 3D CNN-based architecture but included more advanced techniques for extracting temporal features and adjusting the imbalanced dataset. Even with a lower amount of data fed into the algorithm, our model still achieved a 2% higher accuracy rate compared to that model. We also included other models that utilized more advanced feature extraction techniques, such as DFSTN proposed by Liao et al. (2021), 3D DenseNet with a self-attention layer and focal loss for addressing data imbalance (Mehta et al., 2022), and the ER system (Dresvyanskiy et al., 2021) specifically designed for handling imbalanced datasets. Despite our model only using basic optimization techniques and fewer features, the inclusion of body

language still allows our model to achieve a competitive accuracy rate.

5.1 Time Complexity Analysis

The models were trained on Ucloud's 32 Intel Xeon Gold 6130 vCPU. The training time for each model depended mostly on the resolution of the images and was about 750 seconds for the 20 x 11 resolution and about 3500-4100 seconds for the 40 x 22 resolution. The fourfold increase in the amount of data, therefore, led to a slightly greater than fourfold increase in training time. The greater batch sizes usually slightly reduced training time except for the higher resolution data where it increased it. The lack of significant reduction in training time for larger batches was due to a lack of utilization of GPU support. There was also an additional computational cost associated with video transformation, image transformation and sequence aggregation, each of which took time on the order of hours to complete. Without any code optimization, increasing the sampling rate and resolution significantly would greatly increase the computational costs.

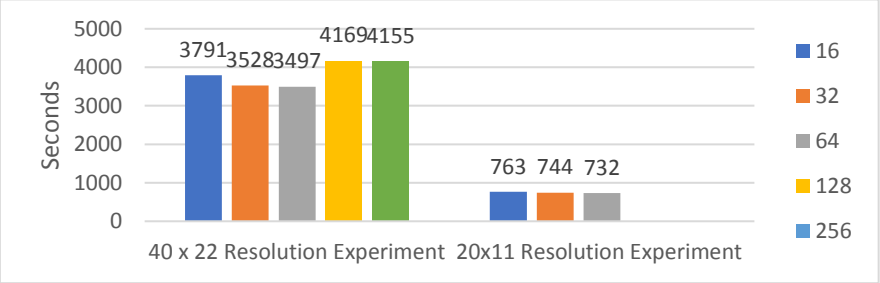


Table 2 The running time of different resolution and different training batch

6. Discussion

By including the language, the model shows a competitive result based less than 1% of data, reducing computational cost on training the data significantly. The same accuracy rate present in the 40 x 22 and 20 x 11 experiment suggests that the increasing in the resolution in our work might be too small to perform the more significant improvement on the model performance. The 56% of accuracy rate also implied the huge improvement space of the engagement analysis

on this imbalanced dataset.

6.1 Error Analysis

Intensity	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Very Low	0.00	0.00	0.00	0.00	0.00	0.00
Low	0.14	0.01	0.02	0.10	0.01	0.02
High	0.58	0.53	0.56	0.55	0.73	0.63
Very High	0.52	0.65	0.58	0.59	0.46	0.51
Resolution	40 x 22	Accuracy:	55%	20 x 11	Accuracy:	56%

Table 3 Evaluation Metric

The F1 Score present as 0.00 in the 'Very Low engagement' and 0.02 in the 'Low engagement' under both solution rate shows that the model basically ignores these 2 classes, predicting only two major classes. These results might cause by the uneven distribution of instance in this dataset since only 6% of instance in this dataset annotated as Low or Very Low Engagement, and 94% of video perform the High or Very High Engagement. The Low and Very Low Engagement classes are too small to contain enough information to let the algorithm learn the pattern.

Resolution	40x22					Resolution	20x11				
	Very Low	Low	High	Very High	Support		Very Low	Low	High	Very High	Support
Very Low	0	0	7	7	14	Very Low	0	0	11	3	14
Low	0	1	50	38	89	Low	0	1	74	14	89
High	0	5	426	373	804	High	0	9	583	212	804
Very Hight	0	1	247	460	708	Very Hight	0	0	384	324	708
	0	7	730	878	1615		0	10	1052	553	1615

Table 4 Confusion Metrix

The confusion matrix for both solutions also verifies that none of the instances were classified as very low engagement. Moreover, the algorithm misclassified them as High and Very High engagement, which are supposed to have more distinct features compared to Low engagement. This indicates that the model has a strong preference for predicting at high and very high intensity. However, compared to the 20 x 11 resolution model, which made more than two-thirds of predictions on High Engagement (the class with the most instances in the test set), the

40 x 22 model makes the predictions more evenly. This fact might indicate that passing more features into the model increases the likelihood of overcoming the data imbalance issue.

Additionally, according to Abedi and Khan's (2021) simulation on multiple well-known engagement classification models combined with different weighted adjustment and feature enhancement techniques, most of the currently available models still suffer from the data imbalance issue and forecast only the majority class (see Appendix 1). The RE system proposed by Dresvyanskiy et al. (2021) has been specifically designed to handle imbalanced data, resulting in more balanced predictions on the DAiSEE dataset (see Appendix 2). However, the trade-off is a low accuracy rate of 39%.

Besides the skewed distribution of data, the arbitrary data annotation method might also pose difficulties in learning the patterns in the minor classes (Liao et al., 2021). Since the data annotation is based on annotators' feelings (Gupta et al., 2016), the reasons for annotating a subject as High Engagement can vary, such as the subject appearing highly engaged throughout the entire video or the subject exhibiting extremely high engagement for a short period. The unclear annotation criteria can lead to an unstable data annotation pattern, reducing the algorithm's ability to learn the feature patterns and perform accurate multiclass classification tasks.

7. Conclusion and Further Work

Based on the results we can conclude with some confidence that body language plays a significant role in detecting levels of engagement. Despite the lower resolution than other researchers, combined with a less sophisticated model architecture, we were able to obtain comparable results in terms of accuracy despite others focusing exclusively on facial expression details. Increasing resolution from 20 x 11 to 40 x 22 had a negligible effect despite a significant increase in training time. The overall unremarkable performance both in our case and in the

case of other researchers was likely due to the deficiencies of the dataset itself, mostly in terms of data labeling. Detecting user engagement in e-learning settings with a high degree of precision remains an open problem with further work still needed to be done.

7.1 Further Work

In this work we present the result that including body language and movement in the engagement can perform the competitive result under the less information provided.

Since the algorithm is learning from different features, the face region cropped operation using in the other paper is the way to normalize the data, providing the better learning environment for algorithm to learning from the facial pattern. (Liao et al., 2021) However, the change in body pose and the gestures will also change the ratio of body in the frame in a non-even zoom in and out way and the square or rectangular cropped technique cannot really accommodate the change in gestures or turn the head. We hope the researchers, content creators and data scientist can propose a new human body detected way and standardize way which can extract and normalize the body language data, increasing the accuracy in engagement intensity analysis that take body language into account.

Additionally, due to the computation resources constraint, the information we used in fed the model in 1/10,000 of the original data and 1/2,500 of original data, although it still presented the competitive result, this work hasn't explored the full potential of the model. Therefore, we recommend to using the higher resolution to check the highest accuracy rate can generate by adding the body movement into consideration, and for the optimize combination, the amount of data we can save compared with solely analysis the engagement level on facial features.

References

Abedi, A., & Khan, S. S. (2021). *Improving state-of-the-art in Detecting Student Engagement with Resnet and TCN Hybrid Network*. https://doi.org/10.1109/crv52889.2021.00028

Ai, X., Sheng, V. S., & Li, C. (2022). Class-attention Video Transformer for Engagement Intensity Prediction. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2208.07216

Dewan, M. a. A., Murshed, M., & Lin, F. (2019). Engagement detection in online learning: a review. *Smart Learning Environments*, *6*(1). https://doi.org/10.1186/s40561-018-0080-z

D'Mello, S. K., Dieterle, E., & Duckworth, A. L. (2017). Advanced, Analytic, Automated (AAA) Measurement of Engagement During Learning. *Educational Psychologist*, *52*(2), 104–123. https://doi.org/10.1080/00461520.2017.1281747

D'Mello, S. K., Lehman, B., Pekrun, R., & Graesser, A. C. (2014). Confusion can be beneficial for learning. *Learning and Instruction*, *29*, 153–170. https://doi.org/10.1016/j.learninstruc.2012.05.003

Dresvyanskiy, D., Minker, W., & Karpov, A. (2021). Deep Learning Based Engagement Recognition in Highly Imbalanced Data. In *Lecture Notes in Computer Science* (pp. 166–178). Springer Science+Business Media. https://doi.org/10.1007/978-3-030-87802-3_16

Gupta, A., D'Cunha, A., Awasthi, K., & Balasubramanian, V. (2016). DAiSEE: Towards User Engagement Recognition in the Wild. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1609.01885

Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1412.6980

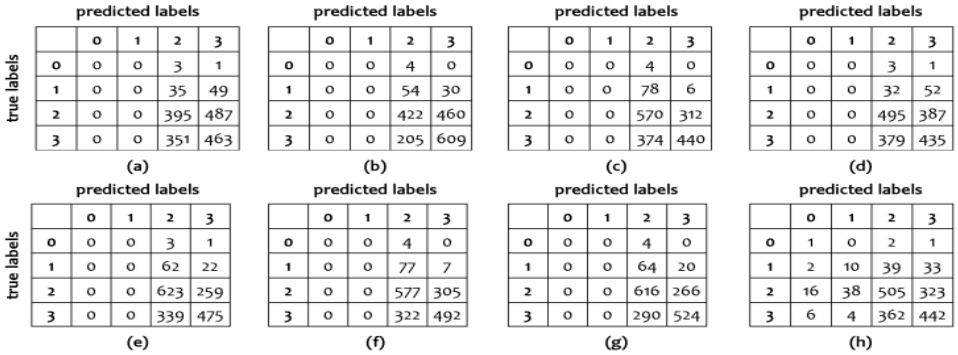
Liao, J., Liang, Y., & Pan, J. (2021). Deep facial spatiotemporal network for engagement prediction in online learning. *Applied Intelligence*, *51*(10), 6609–6621. https://doi.org/10.1007/s10489-020-02139-8

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). *The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression*. https://doi.org/10.1109/cvprw.2010.5543262

Mehta, N. K., Prasad, S. N., Saurav, S., Saini, R., & Singh, S. (2022). Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement. *Applied Intelligence*, *52*(12), 13803–13823. https://doi.org/10.1007/s10489-022-03200-4

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). *Learning Spatiotemporal Features with 3D Convolutional Networks*. https://doi.org/10.1109/iccv.2015.510

Appendix 1: Engagement-level confusion matrices of different methods on the DAiSEE dataset



predicted labels / true labels

	0	1	2	3
0	0	0	3	1
1	0	0	35	49
2	0	0	395	487
3	0	0	351	463

(a)

	0	1	2	3
0	0	0	4	0
1	0	0	54	30
2	0	0	422	460
3	0	0	205	609

(b)

	0	1	2	3
0	0	0	4	0
1	0	0	78	6
2	0	0	570	312
3	0	0	374	440

(c)

	0	1	2	3
0	0	0	3	1
1	0	0	32	52
2	0	0	495	387
3	0	0	379	435

(d)

	0	1	2	3
0	0	0	3	1
1	0	0	62	22
2	0	0	623	259
3	0	0	339	475

(e)

	0	1	2	3
0	0	0	4	0
1	0	0	77	7
2	0	0	577	305
3	0	0	322	492

(f)

	0	1	2	3
0	0	0	4	0
1	0	0	64	20
2	0	0	616	266
3	0	0	290	524

(g)

	0	1	2	3
0	1	0	2	1
1	2	10	39	33
2	16	38	505	323
3	6	4	362	442

(h)

Figure 4 Engagement-level confusion matrices of different methods on the DAiSEE dataset , (a) C3D feature extraction, (b) C3D fine tuning, (c) C3D + LSTM , (d) C3D averaging + LSTM [30], (e) ResNet + LSTM ,(f) C3D + TCN, (g) ResNet + TCN, (h) ResNet + TCN with weighted sampling and weighted loss. From Abedi, A., & Khan, S. S. (2021). Improving state-of-the-art in Detecting Student Engagement with Resnet and TCN Hybrid Network.

Appendix 2: The confusion metrix of RE system proposed by Dresvyanskiy et al. (2021)

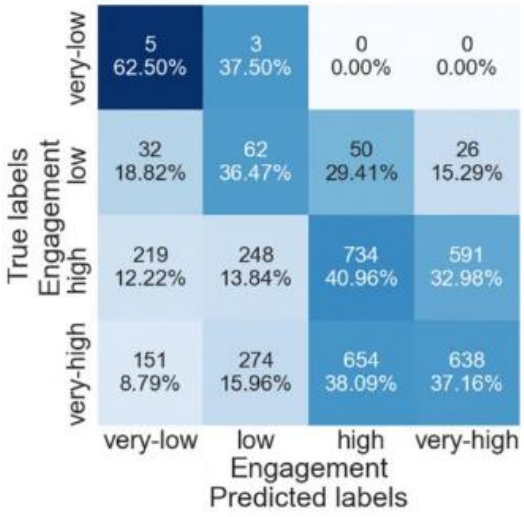


Figure 5 Confusion matrix of ER system on DAiSEE. The prediction distribution is more even among 4 classes. However, the accuracy rate on high and very high labels reduced to less than 50%. From

Dresvyanskiy, D., Minker, W., & Karpov, A. (2021). Deep Learning Based Engagement Recognition in Highly Imbalanced Data. In Lecture Notes in Computer Science (pp. 175).

Appendix 3: The training loss based on different batch size.

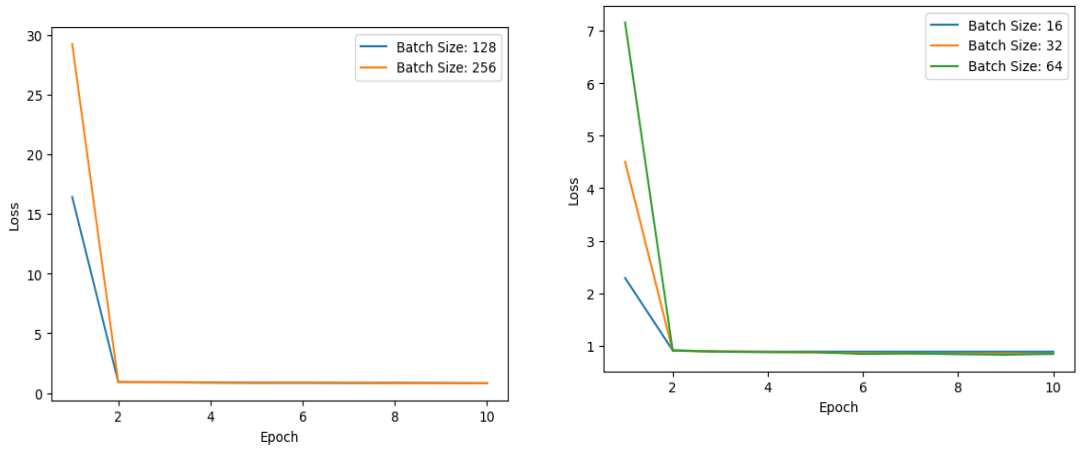


Figure 6 The change of training loss based on 40 x 22 resolution data. The training loss stop reducing after the 2nd epoch in 40 x 22 resolution experiment

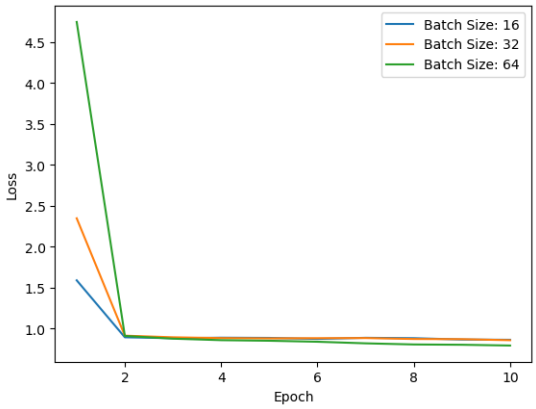


Figure 7 The change of training loss based on 20 x 11 resolution data. The training loss stop reducing after the 2nd epoch in 20 x 11 resolution experiment.

Appendix 4: The relationship between accuracy and batch size

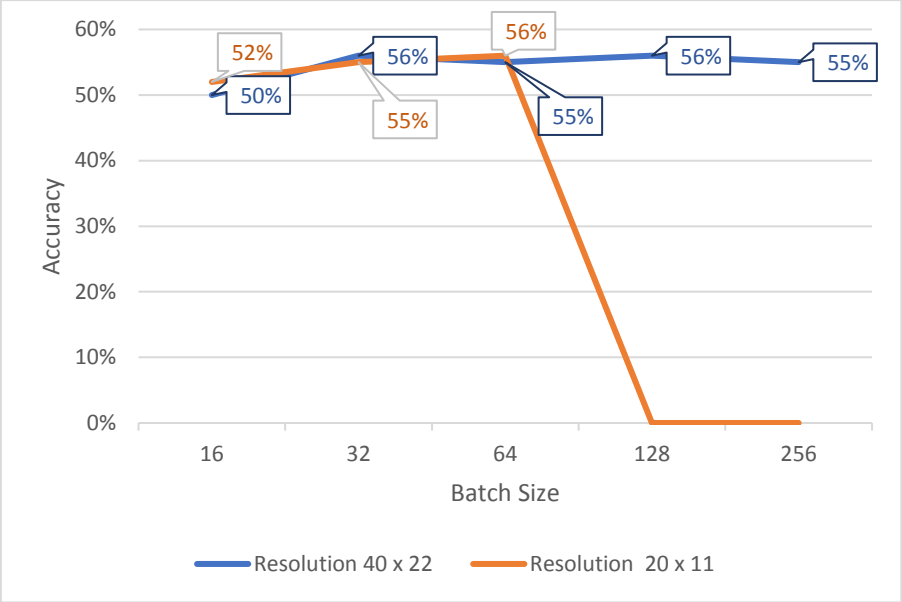


Figure 8 The relationship between accuracy and batch size. In the 20 x 11 resolution experiment, the accuracy presents the positive correlation with the batch size, while the 40 x 22 resolution increasing when the batch size increased to 32, then it fluctuated between 55% and 56%.