Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Applicability of Facial Video-Based Alertness Estimation Model in Real Online and In-Person Classrooms

Atsushi ASHIDA[a*], Ryosuke KAWAMURA[b], Hideaki HAYASHI[a] & Hajime NAGAHARA[a]
[a]*D3 Center, The University of Osaka, Japan*
[b]*Fujitsu Research of America, Inc., The United States*
*ashida@ids.osaka-u.ac.jp

Abstract: Accurately capturing learners' internal states is essential in modern educational environments to support effective teaching and the design of appropriate learning content. Various methods have been proposed for estimating such internal states, with recent approaches increasingly relying on machine learning techniques. However, models trained under specific conditions often fail to generalize to different instructional settings. To be practically useful, these models should be applicable across diverse learning environments, including e-learning, synchronous video-based instruction, and in-person classes. Nonetheless, few studies have evaluated internal state estimation models by transferring them from their training conditions to substantially different and operationally realistic educational settings. In this study, we investigate the effectiveness of our internal state estimation model by applying it in authentic classroom settings. The model estimates learners' alertness based on facial video, specifically focusing on the eye region, and was trained using data collected in a controlled e-learning environment. The evaluation was conducted using data obtained from real educational contexts, including both synchronous online classes and traditional in-person classroom sessions. This setting allowed us to assess the model's robustness across multiple instructional formats that reflect current hybrid learning environments. We compared the model's predictions to human-annotated labels indicating whether learners appeared to be asleep, using receiver operating characteristic (ROC) curves and area under the curve (AUC) scores. The results suggest that the model has the potential to function effectively even when applied to data collected in real-world instructional scenarios.

Keywords: Engagement, alertness, domain shift, computer vision, deep learning

1. Introduction

With societal shifts exemplified by the advent of Society 5.0, there is an increasing need for individualized learning environments that cater to learners' unique needs (Ghosh & Jermsittiparsert, 2024). To realize such personalized optimization, capturing learners' internal state with higher resolution than previously achieved is essential. Methods to assess these internal states typically involve analyzing behavioral and physiological data from learners.

In the field of learning analytics (LA), it is common practice to utilize behavioral data, such as learning logs derived from clickstreams, as they provide clear and quantifiable insights into cognitive aspects of learner behavior (Khor & Mutthulakshmi, 2023). However, such behavioral indicators primarily reflect cognitive dimensions of the learner's internal state. To enhance educational effectiveness, it is critical not only to address these cognitive dimensions but also to incorporate the learner's affective states. By integrating affective states alongside cognitive measures, more adaptive and individualized instructional strategies can be realized, ultimately leading to improved learning outcomes (e.g., D'Mello & Graesser, 2012).

Recent advancements in machine learning and deep learning have led to significant progress in developing technologies capable of capturing the learner's internal affective state. Such technologies hold promises for personalized education, offering deeper insights into learner engagement and emotional responses. However, most existing approaches rely heavily on supervised learning, necessitating data collected from a specific, controlled environment. These supervised methods frequently encounter difficulties when deployed in real-world educational settings, as data obtained outside controlled conditions typically suffer from substantial variability due to external disturbances inherent to authentic educational settings. Consequently, models trained within a controlled environment often fail to generalize adequately when transferred directly to different, more dynamic educational contexts. A common approach to addressing this issue is to retrain models from scratch, requiring extensive data collection and annotation efforts in each new setting. This practice is not only resource-intensive but also exacerbates the cold-start problem, limiting the immediate applicability and scalability of affective state estimation technologies across diverse learning environments.

This paper aims to investigate the applicability of the learner's internal state estimation model, particularly focusing on the model's behavior when applied to data collected in actual classroom settings. Understanding which aspects of these models generalize effectively across varying educational contexts could provide critical insights for enhancing the reusability and applicability of machine learning models, potentially mitigating the challenges posed by the cold-start issue. The research question is “Can a learner's internal state estimation model trained in a controlled environment be feasibly applied and deployed in real-world educational settings?”

To address this research question, we first trained a model using data collected in a controlled e-learning environment. This model utilizes features extracted from the learner's eye region and estimates the learner's alertness levels. We then applied this model to data collected in an actual classroom setting. The data was collected from a synchronous online class conducted at a vocational school in collaboration with our research group. The model's predictions were evaluated by comparison with human annotations to determine its effectiveness in estimating the learner's alertness level in the wild. As an evaluation metric, we used the receiver operating characteristic (ROC) curve and the area under the curve (AUC) to assess the model's performance. The results indicate that the model can accurately estimate alertness levels, provided that the learner's face is properly captured by the camera. This suggests that the model has potential for real-world applications in educational settings.

Additionally, this study was approved by the ethics review committee of the University of Osaka. Data collection was conducted after explaining the purpose and methods to the participants and obtaining their informed consent.

2. Related Research

Prior research has mainly focused on cognitive aspects of learner modeling, such as knowledge representation or problem-solving behaviors. In personalized learning environments, considerable attention has been given to intelligent tutoring systems (ITSs) that monitor learners' states and adapt content accordingly (Wenger, 1987). These systems typically incorporate learner models such as the overlay model or buggy model (Carr et al., 1977; Greer et al., 2010), which aim to represent what the learner knows correctly (overlay model) or misunderstands (buggy model) about domain knowledge. These models are often domain-specific and not easily generalizable across subjects.

The field of learning analytics (LA) has offered more general, domain-independent insights by analyzing large-scale behavioral data. For example, Ma et al. (2022) investigated “jump back” behavior in digital reading, revealing patterns linked to cognitive engagement. However, the learner's internal state encompasses not only cognition but also affective responses, such as alertness or engagement (Kaur et al., 2021).

To address this, researchers have increasingly incorporated multimodal learning analytics (MMLA), combining video, audio, and physiological signals to capture emotional states (Blikstein & Worsley, 2016). Applications include EEG-based impasse detection (Yamamoto et al., 2021) and collaborative learning support (Sugimoto et al., 2017; Spikol et al., 2017), detection of comprehension or non-comprehension (Holmes et al., 2018), and identifying learners' thinking patterns for neurodivergent students. Our research group also contributed to this area with models for alertness estimation based on facial video (Kawamura et al., 2021; Ashida et al., 2024). Despite their promise, MMLA methods often rely on data collected under controlled conditions for training data collection.

While MMLA methods have successfully captured the learner's internal states, most of these studies assume consistent training and deployment conditions. This assumption limits their ability to generalize across different instructional settings. More importantly, despite advancements in sensing and modeling technologies, relatively few studies have attempted to introduce such systems into actual classroom settings. This lack of in-situ implementation highlights a critical gap between controlled experimental validation and practical deployment. Our work aims to bridge this gap by applying a model trained in a controlled environment to real-world classroom data without retraining.

3. Alertness Estimation Model

3.1 Structure of The Model

In this section, the architecture of our developed alertness estimation model is described. This model focuses on learners' behavioral engagement, which is operationalized here as the degree of alertness (Fredricks et al., 2004). This model receives the facial video data of learners as input and outputs the estimated alertness level along with the time sequence. The alertness level is treated as a content-independent indicator of the learner's internal state, enabling its applicability across instructional contexts. Research studies revealed that the information from facial images is effective for recognizing the learner's inner state.

The structure of the alertness estimation model is illustrated in Figure 1. In our model, we utilize eye regions for estimating the alertness level of learners. In the preprocessing, first, the model extracts the frames from the input video data. The model estimates alertness levels based on 10 image frames extracted over a period of one second. Assuming an input video frame rate of 30 fps, frames are extracted from the video at intervals of three frames. After the frame extraction, the model extracts the eye region from the original image. Cropping the eye region consists of two phases. The former step is face detection using Dlib and OpenCV. The latter step is cropping the eye region based on facial landmarks by using Dlib. The eye region is cropped from the original image, and the cropped image is resized to a fixed size.

Let $S_{i:i+10}$ denote video segments from the i-th to $(i+10)$-th extracted frames, these 10 frames correspond to the one second of the input video and $y_{i:i+10}$be the alertness state during the period starting from the i-th frame and ending at $(i+10)$-th frame. In the recognition task, the alertness state during a period of the input sequence is the target of prediction. Therefore, we build a model that takes $S_{i:i+10}$ as input and predicts $y_{i:i+10}$.

The model focuses on the learner's alertness. The model consists of a convolutional neural network (CNN) for image feature extraction and a bidirectional gated recurrent unit (BiGRU) for learning time-directional dependencies. After preprocessing, the eye region images are then fed to a CNN-GRU model to extract features and estimate alertness state for each second. For the CNN component, we employed EfficientNet-B3 pre-trained on the ImageNet dataset (Tan & Le, 2019). Since the CNN-GRU processes ten frames simultaneously, the CNN features are concatenated and fed into fully connected (FC) layers consisting of 2560 and 256 units, respectively. The BiGRU module contains 256 hidden units. The final output is a probability score representing the likelihood of the learner alert.

3.2 Training Data Collection Environment

The training data were collected in a controlled e-learning environment. Data collection was conducted using class video lectures for undergraduate students. We used data from 53 undergraduate students, obtained with informed consent, while they were watching video lectures (Kawamura et al., 2021).

The built-in webcam of a laptop is used for face video data collection. The face video and screen capture are recorded by the software program Bandicam. The frame rate of the

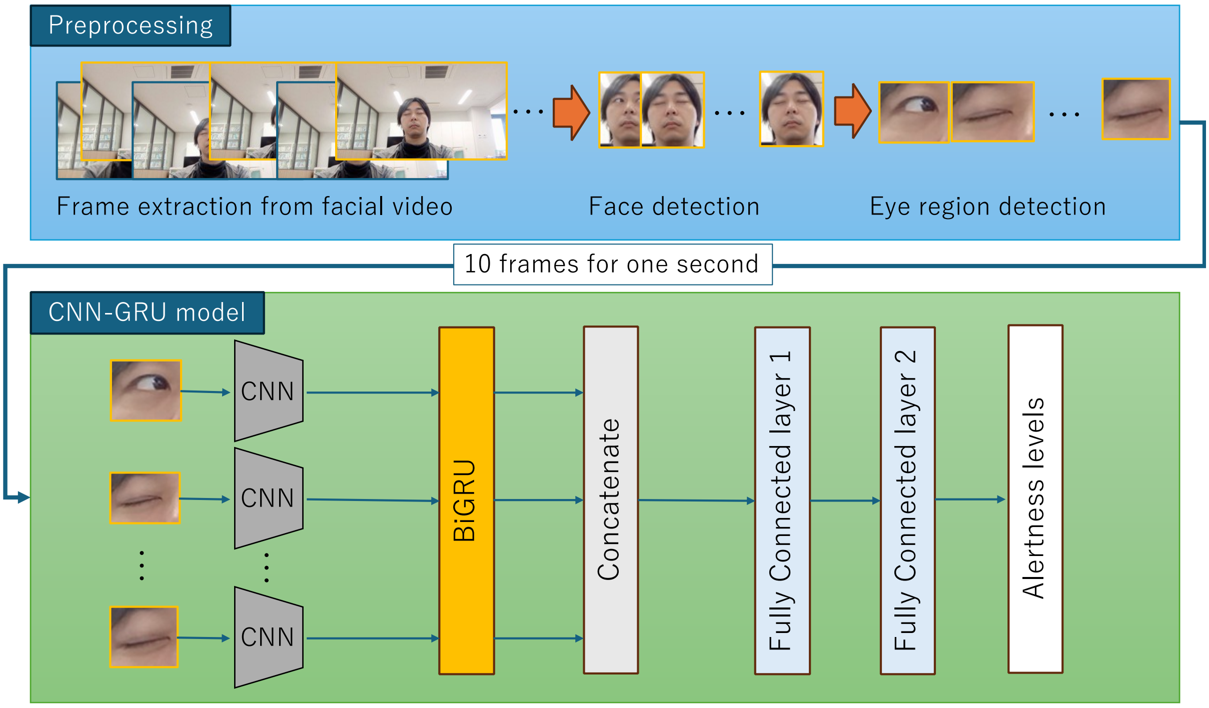
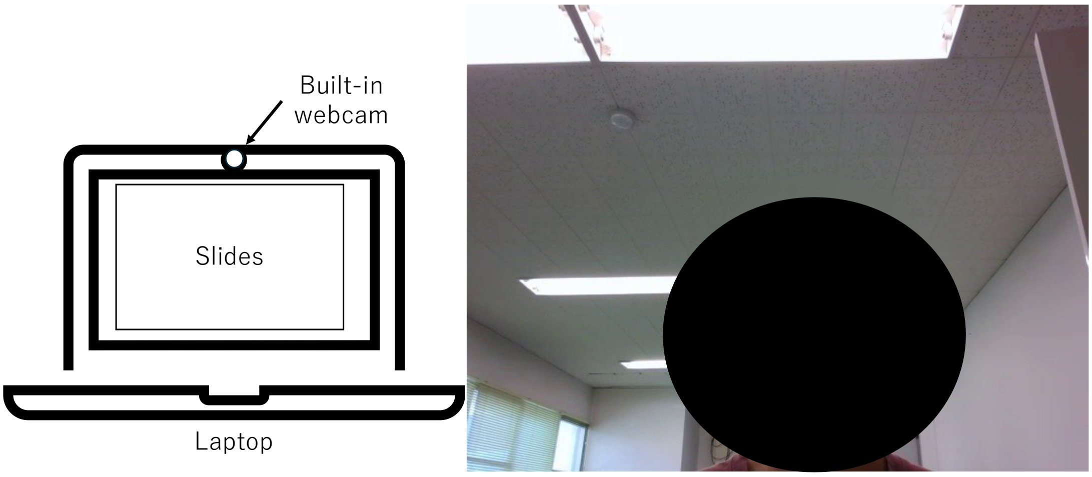


Figure 1. The structure of the model



[redacted]

Figure 2. The image of face videos collection environment and actual collected data in an e-learning setting. The student's face is captured from the front.

facial video is 30 fps, and the resolution of the web-camera is 1280 x 720 pixels, and the screen capture is used to determine whether learners were engaged in learning. The collected facial video data were coded by an annotator. The code is the binary value (0: Asleep, 1: Awake) for every second through human observation. Data from four subjects were annotated by two additional annotators to investigate the reliability of annotations. The annotations among the three annotators matched in 93.8% of selected instances (Ashida et al., 2024).

Lighting conditions for capturing facial video data are bright, and the web camera built into the notebook computer used by the students to attend the video lectures was used, therefore, the students' faces were captured from the front. The sample of captured facial video data is shown in Figure 2. However, in consideration of privacy, individuals are not identified.

4. Data Collection in the Real Class Setting

In-the-wild data were collected at “OCA, Osaka College of Design and IT Technology,” a private vocational school. Informed consent was obtained from all participants. The target class name was “Outline of e-sports.” The class lasted 90 minutes. This class is a lecture-based session in which the instructor explains the current situation surrounding e-sports using slides. We conducted data collection four times. Three sessions were delivered online, and one session was delivered in person. The in-person session was conducted in a different physical classroom from that used for the online sessions. The images of the data collection environments are shown in Figures 3 and 4. The online setting environment is shown in Figure 3. The online lectures were conducted as synchronous classes using Zoom, with the instructor's video and slides displayed on the screen in front of the learners. Smartphones were set up below the display to record a video of the learners' faces. In the case of in-person classes (Figure 4), slides were projected onto a screen at the front of the classroom, and the instructor delivered the lecture next to it. Smartphones were placed on the desk in front of the learners to record a video of their faces.

Each setting of the lighting conditions differed from those in the training data (Figure 2). Table 1 shows the number of participants for each class. All sessions were delivered by the same instructor. The total number of participants was 17, and the total number of video data was 50.

We asked the learners who participated in the data collection to record facial videos during the classes using smartphones. Motorola G13 devices were employed as recording equipment. The videos were captured using the front-facing camera at a resolution of 1920 × 1080 pixels and an average frame rate of ~30 fps. The smartphones were mounted on clamp-style stands affixed to the desk in landscape orientation, and learners were instructed to adjust the position so that their faces remained within the field of view throughout the class (Figure 5). The orientation of the face and the angle of the camera differed from the training data collection situation, with the recording being taken from a lower oblique angle.

5. Experiment

5.1 Experimental Setting

As a preliminary evaluation of the model's applicability to real-class data, we employed the ROC curve and its AUC. To plot the ROC curve, ground-truth labels were required. After recording, the videos were annotated by human annotators who watched each recording and added alertness labels. The annotation task was divided among three annotators, and each video was annotated by a single annotator. The annotation format, consistent with that used for the training data, was binary (1 = awake, 0 = asleep). When an annotator was unable to determine the state, they marked the segment as “undecidable.”

We generated a dataset containing the model outputs and the corresponding human annotations and then created an ROC curve by plotting the true positive rate (TPR) against the false positive rate (FPR) while varying the threshold used to classify the model outputs as either awake or asleep. Data points for which the model failed to output an alertness level, owing to the absence of a detected face during the one-second interval, were excluded from the evaluation. Likewise, time-steps that annotators marked as “undecidable” were omitted from the ground-truth labels. Consequently, the ROC analysis was performed only on intervals for which both a model prediction and a definitive human annotation were available.

Table 1. The participants of the real data collection

Class	No. 1	No. 2	No. 3	No. 4	Total
Type	Online	Online	Online	In-person	-
Number of participants	12	14	12	12	50

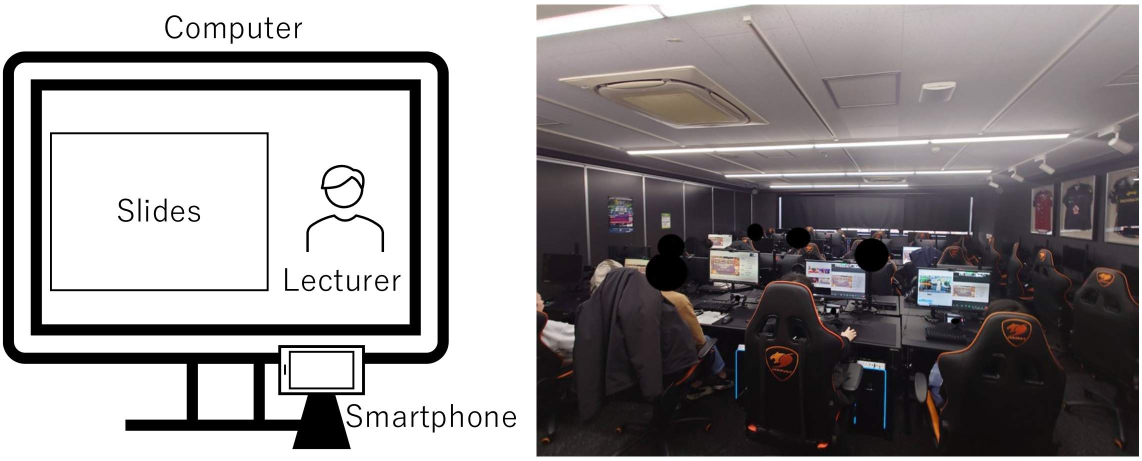


Figure 3. The image and the photo of the facial video data collection environment in online settings. The learners attended the online classes and watched the display in front of them. Many of them set the smartphone under the monitor.

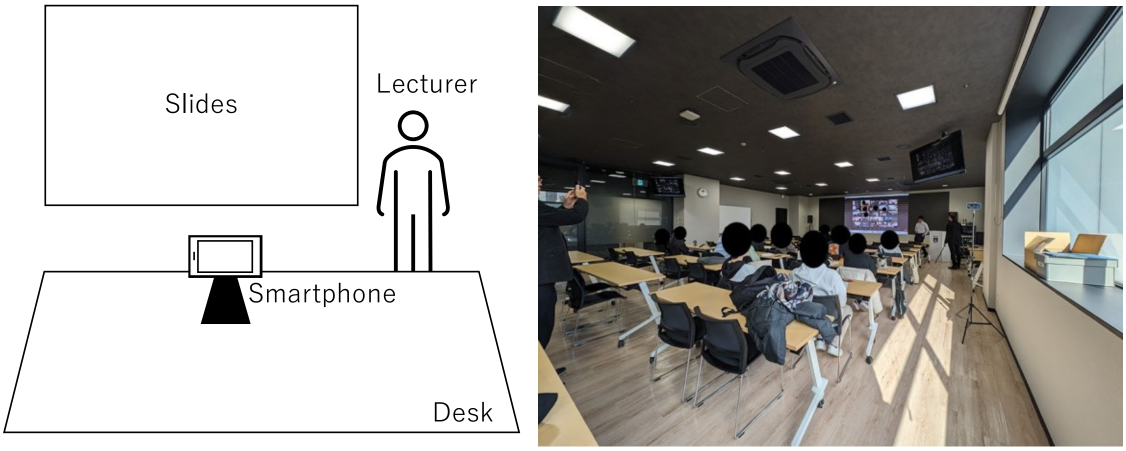


Figure 4. The image and the photo of the facial video data collection environment in an in-person setting. The classroom is different from Figure 3. Learners looked at the lecturer in front of them, and the angle of the face was also different from other settings.

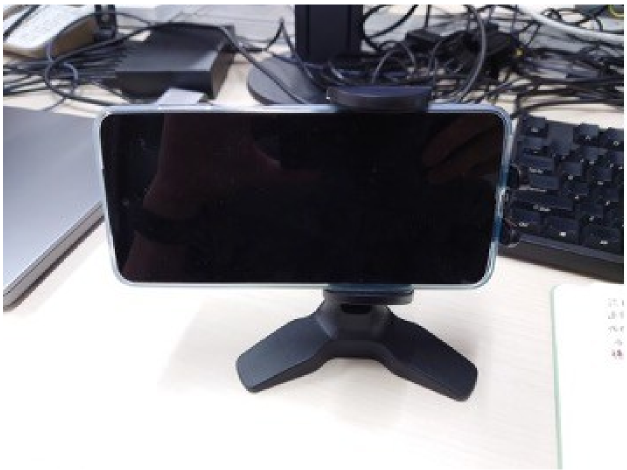


Figure 5. The smartphone setting for data collection, the vertical angle is adjustable. The learners were asked to set the device in front of them and adjust the angle to be able to capture their faces.

5.2 Results and Discussion

We created ROC curves using the data from each of the four individual classes, as well as an ROC curve using the data combined across all classes. Figures 6 and 7 show the ROC curves for the overall dataset that includes data from all classes and each individual class. Corresponding AUC values are shown in Table 2.

In the figures, the orange continuous line indicates the ROC curve obtained from our alertness estimation model. The broken line indicates the 1:1 ratio between TPR and FPR. In other words, this line shows the result of the random classifier. Also, the closer the orange lines are to the top-left corner, the better the performance it indicates. The AUCs indicate the area under the ROC curves. The AUC takes a value between 0 and 1, with values closer to 1 indicating better classifier performance.

From figures 6 and 7, the AUCs were over 0.75 for all classes, and 0.84 for over all data. Various interpretations of the AUC exist (de Hond et al., 2022). In this paper, we adopted the criteria of Polo et al. They suggested that AUC values can be interpreted as follows: 0.5–0.6 (failed), 0.6–0.7 (worthless), 0.7–0.8 (poor), 0.8–0.9 (good), and > 0.9 (excellent) (Polo et. al., 2020). When applying this criterion to our results, the AUC for the overall dataset can be interpreted as good. These findings suggest that our model, which was trained on data collected in a controlled e-learning environment, has the potential to estimate learners' alertness levels in a real educational setting.

When evaluating the AUC for each class, Class No. 1 is categorized as "poor", Classes 2 and 3 as "excellent", and Class No. 4 as "good" based on the aforementioned AUC interpretation criteria. This indicates that the alertness estimation model has the potential to estimate the learner's alertness level regardless of the form of the class, online or in-person. Higher AUC values are expected for online classes compared to the in-person class, as the data collection conditions in the online setting are more similar to those used during training. Indeed, Classes 2 and 3, which were conducted online, show higher AUC values than Class 4, which was conducted in person. However, Class 1, also an online class, exhibits a lower AUC value than Class 4. Since Class 1 was the first dataset collected, it may have been influenced by factors such as participant unfamiliarity with the system or environmental inconsistencies, which could have affected model estimation accuracy.

Also, there are scenes in which the model failed to estimate the learner's alertness levels. For example, there were scenes where the learners concentrated on the class, but the model outputs a low alertness level. In this case, the model fails to estimate the correct alertness level. We need to investigate the model performance of the data in the e-learning

Table 2. The value of the AUCs

Class	No. 1	No. 2	No. 3	No. 4	Total
AUC	0.76	0.95	0.91	0.81	0.84

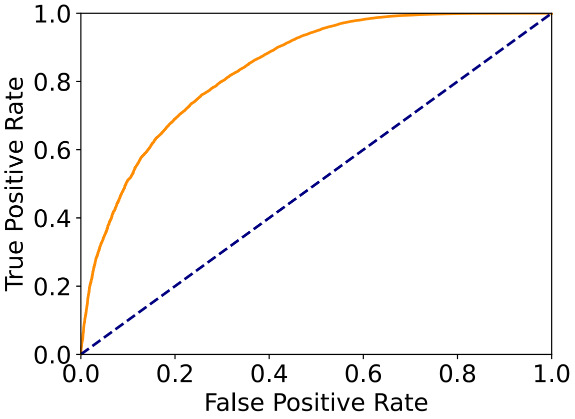
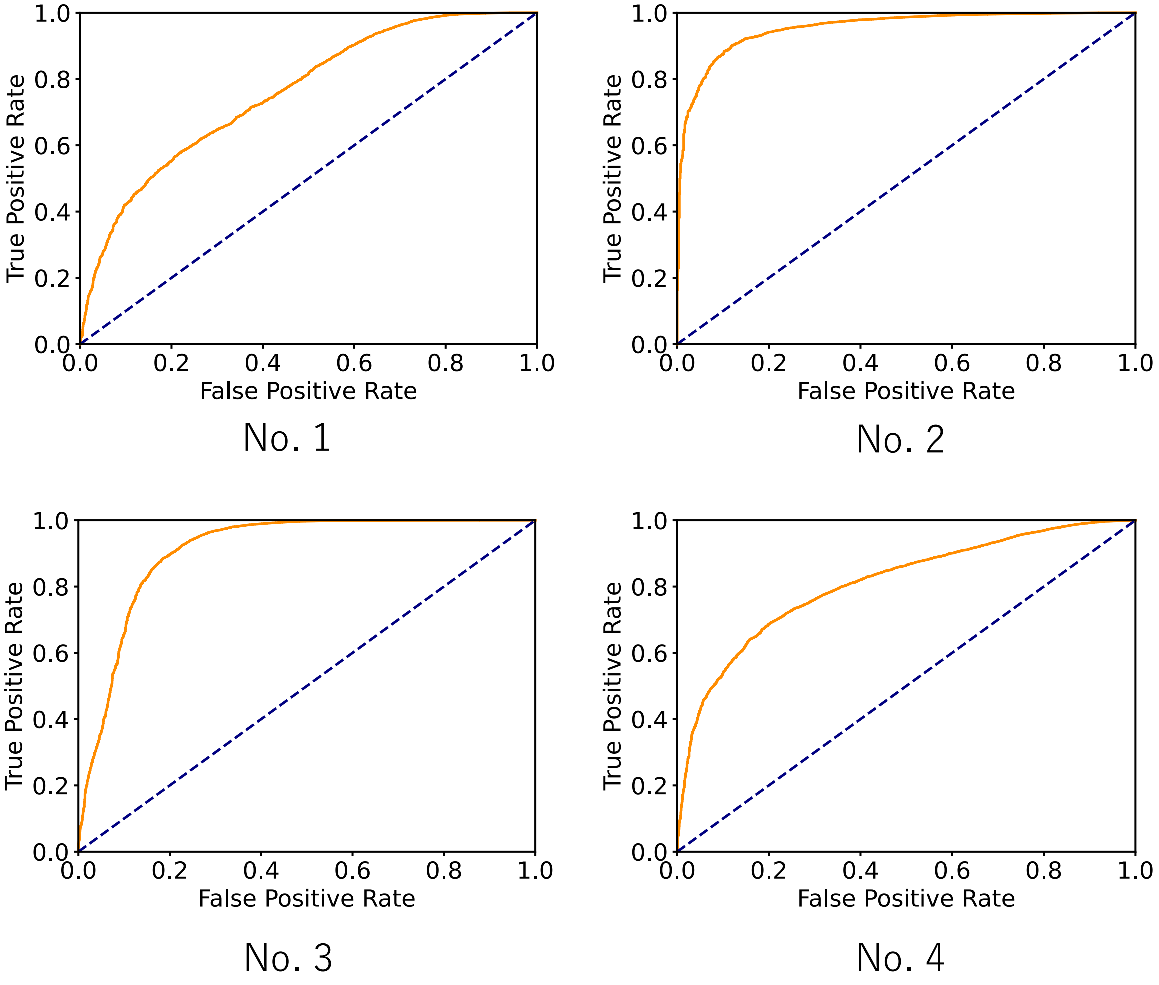


Figure 6. The ROC curve and AUC for overall data.



No. 1

No. 2

No. 3

No. 4

Figure 7. The ROC curves in class No. 1, No. 2, No. 3, and No. 4.

environment and compare it with data in the real educational environment to clarify whether the performance varies among environments. Also, we would like to clarify whether the additional training of the model using the data collected in this research improves the model's performance.

In the practical deployment of an alertness detection system, potential applications include promoting learners' metacognitive awareness when their alertness decreases or providing feedback to instructors by identifying students who require attention. In such cases, it is crucial to evaluate the system from the perspective of whether it can accurately detect learners with low alertness, which typically represents the minority class. Therefore, future evaluations should focus on metrics suitable for real-world implementation, such as plotting precision–recall (PR) curves and calculating PR-AUC values by treating the minority class, i.e., low-alertness learners, as the positive class.

6. Conclusion

This study investigated the applicability of a pre-trained alertness level estimation model to actual educational settings, which represent different environments from where the model was trained. The model used crop images of the area around the eyes from learners' facial videos as input and outputs the learners' alertness levels. It was trained on training data collected from learners taking e-learning courses in a controlled environment.

In the real-world environment, learners' facial videos were recorded during a total of four classes, including synchronous online classes and face-to-face classes. The alertness levels of the recorded videos were estimated by the alertness level estimation model, and the ground truth was annotated by human annotators. For model evaluation, ROC-AUC was used as the evaluation metric. The evaluation results showed that the alertness estimation model achieved a reasonably good AUC score, suggesting that alertness estimation using images of the eye region, as employed by the model, has the potential to be used in actual classroom environments.

As future work, by fine-tuning the existing model with data collected in real-world environments and examining how much the model's performance improves, we can determine whether learning in a new environment is necessary. Furthermore, the usability of the data output by this model differs in actual classes. For instance, when identifying students with low alertness levels to provide feedback, higher performance in correctly classifying low-alertness data is preferable. Therefore, evaluating the model using metrics that emphasize the correct identification of positive instances, such as PR-AUC, can be considered. Additionally, as a future development, we aim to contribute to the improvement of actual learning environments by developing a dashboard that allows for the evaluation of learners and the classes themselves based on the metrics obtained from this model.

Acknowledgement

The authors would like to thank the staff and students of OCA, Osaka College of Design and IT Technology, for their participation and cooperation in the experiments.

This paper is partially supported by the Innovation Platform for Society 5.0 from the Japan Ministry of Education, Culture, Sports, Science and Technology, and JSPS KAKENHI Grant Number 24K16753.

References

Ghosh, U. K., & Jermsittiparsert, K. (2024). Personalised learning systems and the human touch in society 5.0. In Practice, Progress, and Proficiency in Sustainability, IGI Global, 211–232.

Khor, E. T., & Mutthulakshmi, K. (2023). A systematic review of the role of learning analytics in supporting personalized learning. *Education Sciences, 14*(1), 51.

D'Mello, S., & Graesser, A. (2012). AutoTutor and affective autotutor: learning by talking with cognitively and emotionally intelligent computers that talk back. ACM Transactions on Interactive Intelligent Systems, 2(4), 1–39.

Wenger, E. (1987). *Artificial intelligence and tutoring systems: Computational and cognitive approaches to the communication of knowledge.* Morgan Kaufmann Publishers Inc.

Carr, B., & Goldstein, I. P. (1977). *Overlays: A theory of modelling for computer aided instruction*.

Greer, J. E., & McCalla, G. I. (Eds.). (2010). Student modelling: The key to individualized knowledge-based instruction. Springer.

Ma, B., Lu, M., Taniguchi, Y., & Konomi, S. (2022). Exploring jump back behavior patterns and reasons in e-book system. Smart Learning Environments, 9(1), 1–23.

Kaur, P., Kumar, H., & Kaushal, S. (2021). Affective state and learning environment based analysis of students' performance in online assessment. *International Journal of Cognitive Computing in Engineering, 2*, 12–20.

Blikstein, P., & Worsley, M. (2016). Multimodal Learning Analytics and Education Data Mining: using computational technologies to measure complex learning tasks. Journal of Learning Analytics, 3(2), 220–238.

Yamamoto, S., Tobe, Y., Tawatsuji, Y., & Hirashima, T. (2021). In-process Feedback by Detecting Deadlock based on EEG Data in Exercise of Learning by Problem-posing. Proceedings of the 29th International Conference on Computers in Education, 21–30.

Sugimoto, A., Hayashi, Y., & Seta, K. (2017). Multimodal interaction aware platform for collaborative learning. Proceedings of the 25th International Conference on Computers in Education, 316–325

Spikol, D., Ruffaldi, E., Landolfi, L., & Cukurova, M. (2017). Estimation of success in collaborative learning based on multimodal learning analytics features. 2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT), 269–273.

Holmes, M., Latham, A., Crockett, K., & O'Shea, J. D. (2018). Near Real-Time Comprehension Classification with Artificial Neural Networks: Decoding e-Learner Non-Verbal Behavior. IEEE Transactions on Learning Technologies, 11(1), 5–12.

Wong, A. Y., Bryck, R. L., Baker, R. S., Hutt, S., & Mills, C. (2023). Using a Webcam Based Eye-tracker to Understand Students' Thought Patterns and Reading Behaviors in Neurodivergent Classrooms. LAK23: 13th International Learning Analytics and Knowledge Conference, 453–463.

Kawamura, R., Shirai, S., Takemura, N., Alizadeh, M., Cukurova, M., Takemura, H., & Nagahara, H. (2021). Detecting drowsy learners at the wheel of e-learning platforms with multimodal learning analytics. IEEE Access, 9, 115165–115174.

Ashida, A., Kawamura, R., Shirai, S., Takemura, N., Alizadeh, M., Hayashi, H., & Nagahara, H. (2024). Is internal state feedback in an e-learning environment acceptable to people? In *Proceedings of the 32nd International Conference on Computers in Education*, 89–94.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School Engagement: Potential of the Concept, State of the Evidence. Review of Educational Research, 74(1), 59-109.

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning, 6105–6114.

de Hond, A. A. H., Steyerberg, E. W., & van Calster, B. (2022). Interpreting area under the receiver operating characteristic curve. *The Lancet Digital Health*, 4(12), e853–e855.

Polo, T. C. F., & Miot, H. A. (2020). Use of ROC curves in clinical and experimental studies. Jornal Vascular Brasileiro, 19, e20200186.