IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 April 2025, accepted 4 June 2025, date of publication 9 June 2025, date of current version 18 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3578185

RESEARCH ARTICLE

EyeUnderstand: Dashboard for Gaze and Deep-Learning Driven Comprehension Estimation in Online Lectures

KO WATANABE[1], GITESH GUND[2], JAYASANKAR SANTHOSH[1,2], HARUKA SAKAGAMI[3], YUKI MATSUDA[4], (Member, IEEE), ANDREAS DENGEL[1,2], AND SHOYA ISHIMARU[5], (Member, IEEE)

[1]Smart Data and Knowledge Services, German Research Center for Artificial Intelligence (DFKI) GmbH, 67663 Kaiserslautern, Germany
[2]Department of Computer Science, RPTU Kaiserslautern-Landau, 67663 Kaiserslautern, Germany
[3]Graduate School of Science and Technology, Nara Institute of Science and Technology (NAIST), Nara 630-0192, Japan
[4]Faculty of Environmental, Life, Natural Science, and Technology, Okayama University, Okayama 700-0082, Japan
[5]Graduate School of Informatics, Osaka Metropolitan University, Sakai 599-8531, Japan

Corresponding author: Ko Watanabe (ko.watanabe@dfki.de)

This work was supported in part by Japan Society for the Promotion of Science under Grant 23KK0188.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Graduate School of Informatics at Osaka Metropolitan University.

ABSTRACT Online videos are a potent tool for educators to disseminate knowledge widely to diverse student audiences. However, collecting student feedback remains a significant challenge for lecturers, particularly in the absence of feedback. Understanding students' subjective comprehension levels during online video lectures with sensor technology is yet to be thoroughly researched. This study uses eye-tracking technology to predict self-reported comprehension levels during video lectures. We recruited 20 participants from Germany and Japan who were invited to watch 50-minute lecture videos in three domains. The participants self-annotate the time segment in each lecture video where they dropout using open-source *LabelStudio* and answer the survey. We applied Long-Short-Term Memory (LSTM) to the preprocessed dataset and achieved an F1 Score of 0.886 for predicting binary self-annotated comprehension levels. We also introduce *EyeUnderstand*, the web-based application for visualizing the results of the comprehension estimation. We recruited 28 participants for the user study. As a result, 89.3% of the students and 92.9% of the lecturers confirmed that our application is practical.

INDEX TERMS Eye-tracking, hybrid education, online lecture, subjective comprehension estimation, dropout, deep-learning, web application, education.

I. INTRODUCTION

Eyes speak more eloquently than lips. This is a well-known saying in Japan that the eyes express honest thoughts better than what a person says. As explained in this proverb, much research has been done to understand people through their eyes [1], [2], [3], [4], [5]. Hence, understanding the eyes is important in human-computer interaction (HCI).

Remote sensing is becoming increasingly crucial after the devastating COVID-19 pandemic. Education has undergone a profound transformation, rapidly shifting towards hybrid learning models. Online video lectures have become a powerful tool for disseminating knowledge to a vast audience asynchronously. However, compared to in-person lectures, the lack of monitoring students' capabilities is a significant challenge for teachers, leading to an urgent need for a deeper understanding of non-verbal cues from remote audiences. Compelling evidence from previous research underscores the limited attention span of audiences, with studies suggesting a mere 10–20 minutes of sustained focus before attention wanes [6]. Moreover, dropout rates escalate proportionately to the duration of video content [7]. Understanding content

The associate editor coordinating the review of this manuscript and approving it for publication was Wenming Cao.

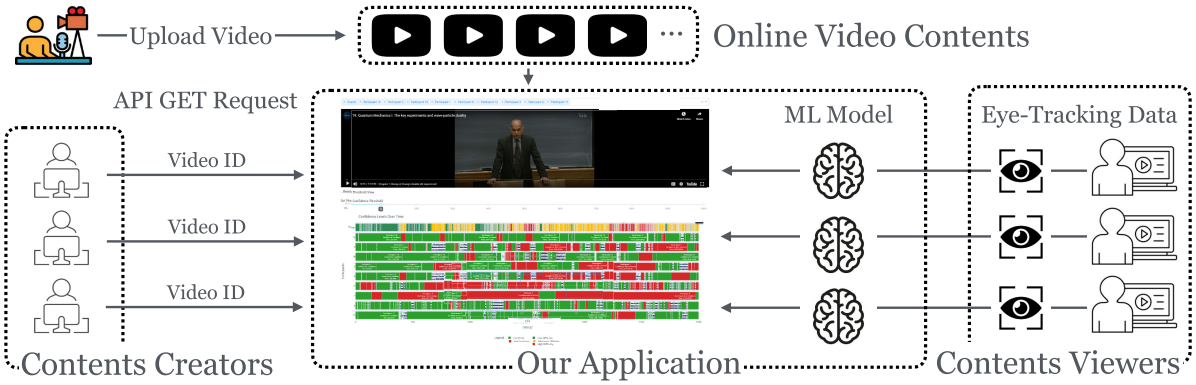


FIGURE 1. Overview of our application *EyeUnderstand* use case scenario. Our application allows for the importation of eye-tracking log data and the estimation of self-reported comprehension levels with the machine learning model. The result will be visualized as a bar graph, and video content creators (such as lecturers) can check where the content viewers (such as students) drop while watching the online video.

viewers (students) comprehension while watching online video lectures will help content creators (lecturers) and content viewers gain feedback and higher-quality videos.

Comprehension can be viewed from both objective and subjective perspectives. The objective perspective involves estimating understanding through test scores or other quantifiable assessments [8], [9], [10]. In contrast, the subjective perspective refers to how learners perceive their level of understanding, which is often a key factor when students fall behind or drop out. While many experiments and studies have focused on objectively measuring comprehension [8], [9], [10], this research aims to investigate the subjective aspect of comprehension in real-time during class.

In this study, we aim to understand human subjective comprehension levels while watching the online video lecture from their eyes. To do so, we recruited an eye-tracker to collect participants' eye movements while watching the video and gain post-self-annotation of subjective comprehension level from themselves. We instructed participants to complete the survey to ensure that they remained focused on the lecture. Figure 1 shows our application overview of the whole application concept. We introduce the approach of estimating subjective comprehension levels from the eyes and aim to combine multiple students' overviews to understand the quality of online lecture videos better. We implemented a pilot application, *EyeUnderstand*, which can import eye-tracking raw data and visualize the estimation outcome of subjective comprehension level as graphs. The application, *EyeUnderstand*, allows video content creators, especially lecturers, to gain valuable feedback on when and which time segments of the video students find dropout against the video. This insight can be instrumental in improving the quality of online lecture videos and enhancing the learning experiences.

Our findings and implementation make significant strides toward education and providing a lifecycle for video content creators and viewers, especially feedback for remote learning environments. Our contributions (C1-C3) to this paper are as follows:

- C1 **Subjective Comprehension Level Estimation**: We investigate prediction of subjective comprehension state from the participant gaze data.
- C2 **Cultural Comparison**: We compared the gaze behaviors of participants in Germany and Japan.
- C3 **Implementation of Web Application**: We present our application *EyeUnderstand* and conduct a user-study of the usability.

This paper begins by introducing relevant related works. Following that, we explain the methodologies of our study. Then, we explain our data collection approach, results, and discussion. Subsequently, we will introduce our web application *EyeUnderstand* in detail. Finally, we address the limitations of our dataset and outline future directions for this research.

II. BACKGROUND AND RELATED WORK

This section presents inspired related work. The section begins with various sensors used for activity recognition. Then, we introduce works on cognitive understanding using various sensors. Lastly, eye-tracking works, focusing on understanding lecture attendees.

A. ACTIVITY SENSING WITH SENSORS

Zhai et al. [11] presents a pioneering framework centered around the design of a real-time video stream-oriented behavior recognition platform [11]. Their groundbreaking work harnesses the power of edge computing within the camera, facilitating robust activity recognition under diverse conditions. Dimiccoli et al. [12] has used egocentric camera to collect 21 classes of activities [12]. This work achieved a recognition accuracy of 79%. *EduSense* project study focuses on using a camera to collect students' activities in offline classes [13]. The project uses a Microsoft Kinect one-depth camera and Intel NUC to discover student activities such as raising a hand, sitting, standing, smiling, speaking, attention, class gaze, or head orientation. There are more

TABLE 1. Comparison of previous work against our work. OC stands for objective comprehension, SC stands for subjective comprehension, and ET stands for eye-tracking. The implementation of the App (visualization application), and the study detail is also mentioned.

Author	OC	SC	ET	App	Study Detail
Augereau et al. [9]	✓	✗	✓	✗	Estimate TOEIC score using mobile eye-tracker with an error of 36.3 points.
Sauter et al. [20]	✓	✗	✓	✗	The study aim to predict test quiz performance using the synchronicity of eye movements.
de-la Peña [21]	✓	✗	✓	✗	Implicit comprehension improves with shorter eye movements, larger vocabulary, and faster processing.
Huang et al. [22]	✓	✗	✓	✗	L2 comprehension links to FFD on unfamiliar words, moderated by WM.
Srivastava et al. [3] [23]	✗	✓	✓	✗	They study collect 100 participants subjective comprehension while watching online video lecture.
Sanches et al. [8]	✓	✓	✓	✗	Estimate correct answer percentage with 5.27% error (eye gaze) vs. 9.04% error (self-assessment)
Santhosh et al. [24]	✓	✓	✓	✗	Adaptive Interventions significantly improving subjective and objective comprehension
Ours	✗	✓	✓	✓	Model and dashboard implementation of subjective comprehension visualization with an eye-tracker.

works on sensing activities like nodding [14], [15], [16] or postures [17], [18]. Zhu et al. [19] present work on comparing mouse (cursor movement) activity and gaze behavior in e-learning condition [19]. The work found that both mouse and gaze movement correlates with activity. The work is influential in replacing estimating gaze movement but does not include pupil diameter. These projects aim to measure participants' offline and online activity using sensors.

B. COGNITIVE STATE SENSING WITH SENSORS

Kawamura et al. [25] put forth a compelling approach to estimate audience wakefulness during video lectures, harnessing the synergistic fusion of multi-modal data. By leveraging facial expressions captured via webcams and seat pressure measurements, they achieved a commendable F1-macro score of 0.70, signifying a substantial advance in wakefulness estimation during remote lectures. Abdelrahman et al. [26] used thermal sensor to measure the cognitive load. The work achieved in measuring nose temperature to reduce when having a higher cognitive load. Meanwhile, forehead temperature is the inverse relationship of the nose. Other then above, engagement detection [27], [28], [29], [30], or attention levels estimation [23], [31], [32] are done by several researchers.

Kar et al. [33] have made significant strides in estimating audience attention during lectures, combining presentation slides, eye gaze, and gaze gestures to capture a learner's attention level, achieving an impressive average absolute error rate of 8.68%. Abdelrahman et al. [34] mentioned that attention has various classes. To understand participants' attention states, they collected 22 participants and achieved four different attention state classifications by AUC of 75.7% for user conditions.

Burch [35] focuses on providing real-time slide feedback, leveraging the students' gaze as a vital indicator of engagement and confidence. Meanwhile, Bixler and D'Mello [36] have pioneered the development of a supervised classification cross-domain model for detecting mind-wandering during lectures, leveraging an extensive dataset of 132 users' mind-wandering reports. With the domain dependency, AUROCs scored from 0.57 to 0.72 for estimating mind-wondering. Similarly, Zermiani et al. [37] have embarked on a pilot study that analyzes gaze patterns associated with tendencies towards mind-wandering during lecture viewing. While their findings highlighted notable trends in off-screen fixation behavior, challenges on the variation in mind-wandering scenarios and participant-specific responses call for further nuanced investigations and novel methodological approaches to unlock the full potential of lecture-based learning environments.

C. COMPREHENSION ESTIMATION USING EYE-TRACKING

Table 1 shows the overall comparison of the previous work and our study. Augereau et al. [9] work on estimating TOEIC [1] score using a mobile eye-tracker. After calculating fixation and saccade, the study achieved an absolute mean error of 36.3 points, with a standard deviation of 46.5 points.

Sauter et al. [20] have aimed to predict test quiz performance using the synchronicity of eye movements. While their study did not establish a direct correlation between eye movement synchronicity and test results, it underscored the pivotal role played by the teacher's presentation style in shaping student engagement.

de-la Peña [21] aims to understand the role of eye-tracking in implicit reading comprehension and intervening skills such as vocabulary, rapid automatized naming, and processing speed. Their findings implicate that better performance in reading comprehension (literal, inferential, and total) is related to and explained by shorter eye movement, more extensive vocabulary, and processing speed.

Huang et al. [22] combines online (eye-tracking) and offline (reading comprehension test) measures to investigate the relationships among word processing, working memory (WM), and second language (L2) reading comprehension performance, and their results expand the understanding of the role of WM in unfamiliar word processing during L2 reading comprehension.

Srivastava et al. [3] pioneered contactless sensors to gauge learning difficulties in digital learning environments. They collected 100 participants' real-time difficulty self-reports with eye-tracking and thermal cameras. Srivastava et al. [23] then convert gaze information into the area of interest (AOI)

[1] http://www.toeic.or.jp/library/toeic data/toeic en/pdf/data/ TOEIC Program DAA.pdf

and check the synchronization with the screen. The work aims to determine whether the student is on track with the lecturer's explanation.

Sanches et al. [8] proposed a method to estimate the objective understanding of a learner by analyzing eye movements while reading. Their findings reported an error of 5.27% in the number of correct answers estimation by using eye gaze features, while a comparison using the reader's self-assessment understanding leads to a 9.04% error.

Santhosh et al. [24] utilized real-time engagement-based ChatGPT-generated summaries to enhance reader comprehension and learning outcomes. The results revealed that AI-driven interventions exhibited significantly better learning outcomes, higher engagement, and better objective comprehension results.

In conclusion, various research studies aim to use eye-tracking to estimate objective and subjective comprehension. However, the visualization of subjective comprehension through dashboard implementation has yet to be discovered. Hence, this study's most significant contribution aligns here.

III. METHODOLOGY

This section explains experimental settings, materials, data acquisition tools, data processing, and feature engineering processes.

A. DATA ACQUISITION

In this study, we used Tobii eye-tracker,[2] Label Studio [38] for self subjective comprehension annotation, and Google Form[3] used for the survey for collecting subjective feedback.

1) EYE-TRACKING DEVICE

The Tobii eye-tracker, which has a sampling rate 90Hz and records eye movements, is mounted in the Microsoft Surface Studio 1. It is a remote device with an academic license. All participants used Microsoft Surface Studio 1 desktop computers. The screen size is 637.35 × 438.90 mm, ensuring each lecture video can be displayed on a screen. This device is widely utilized in human-computer interaction research due to its high precision in capturing gaze behavior and pupil diameter changes.

2) LABEL STUDIO

Microsoft Surface Studio, with a screen resolution of 4500 × 3000 pixels. The large screen size ensured clear visibility of the lecture videos. The eye-tracker collects timestamps, pupil diameters, and x and y locations of the gaze. We also collect self-annotation of the labeling of participant *dropout* time segment using *LabelStudio* [38]. The application is an open-source graphical user interface designed for ease of use. It empowers anyone to make subjective comprehension labels easily.

[2]https://www.tobii.com/products/eye-trackers/screen-based
[3]https://workspace.google.com/products/forms/

TABLE 2. The list of online lecture videos. LID corresponds to Lecture ID, and CS corresponds to Computer Science.

LID	Category	Language	Online Video Lecture Link
L1	Music	English	https://youtu.be/uDVr0GaD7gI
L2	Physics	English	https://youtu.be/uK2eFv7ne_Q
L3	CS	English	https://youtu.be/xv0MnQhVWjI
L4	Music	Japanese	https://youtu.be/4ZeyWopr1dE
L5	Physics	Japanese	https://youtu.be/jwQY0vOAiOQ
L6	CS	Japanese	https://youtu.be/-j1hoCubiyE

3) MCQ TOOL

Google Forms presented the participants with the multiple-choice questions (MCQs). This tool streamlined the process of assessing participants' comprehension and provided structured data for analysis.

B. MATERIALS: ONLINE VIDEO LECTURES

Table 2 shows the details of lectures viewed by English and Japanese speakers. As shown, we prepared six lecture videos (L1-L6), three types for participants in Germany and another three for Japan. We prepared lectures in English and Japanese for participants in Germany and Japan.

We select three domain-specific lectures to collect a variety of behaviors while watching each university-level lecture. Also, we collect participants from two different language domains to verify their robustness or versatility. A post-survey asked two questions: "What did you find easy (difficult) to understand about the lectures?". This perspective supports understanding why the participant felt *dropout* about the video lecture.

C. DATA PREPROCESSING

This section explains the major steps of the preprocessing pipeline. In the initial stage of our study, the dataset comprised 13,445,116 entries across 40 columns. Through a meticulous preprocessing process, we meticulously refined the dataset to 9,845,886 entries distributed among 17 key columns, ensuring the utmost reliability of our data.

To ensure consistency and comparability of data points across different scales, we standardized the numerical features. This process involved normalizing these features based on the mean and standard deviation, effectively rescaling them with a mean of zero and a standard deviation of one. Such standardization is important and vital for many machine learning algorithms, which assume data is normally distributed and scales uniformly. It ensures the uniformity and reliability of our data.

D. FEATURE ENGINEERING

Table 3 shows the entire feature engineering of the raw dataset for estimating participant subjective comprehension. This reduction was primarily due to eliminating rows and columns that contained missing values or were deemed irrelevant for our analysis, enhancing the quality and reliability of our data. The selection of features was strategically focused on

TABLE 3. Dataset after preprocessing and feature engineering for subjective comprehension estimation.

Dataset	Entries	Columns	Keys
Raw data	13,445,116	40	*#timestamp*, *tobii_system_timestamp*, *tobii_device_timestamp*, *second*, *uid*, *cid*, *subjective_comprehension*, *gaze_x*, *gaze_y*, *pupil_diameter*, *left_gaze_x*, *left_gaze_y*, *left_pupil*, *left_gaze_origin_in_user_coordinate_system_x*, *left_gaze_origin_in_user_coordinate_system_y*, *left_gaze_origin_in_user_coordinate_system_z*, *left_gaze_point_in_user_coordinate_system_x*, *left_gaze_point_in_user_coordinate_system_y*, *left_gaze_origin_in_trackbox_coordinate_system_x*, *left_gaze_origin_in_trackbox_coordinate_system_y*, *left_pupil_validity*, *left_gaze_origin_validity*, *right_gaze_x*, *right_gaze_y*, *right_pupil* *right_gaze_origin_in_user_coordinate_system_x*, *right_gaze_origin_in_user_coordinate_system_y*, *right_gaze_origin_in_user_coordinate_system_z*, *right_gaze_point_in_user_coordinate_system_x*, *right_gaze_point_in_user_coordinate_system_y*, *right_gaze_origin_in_trackbox_coordinate_system_x*, *right_gaze_origin_in_trackbox_coordinate_system_y*, *right_pupil_validity*, *right_gaze_origin_validity*
Preprocessed	9,845,886	17	*#timestamp*, *second*, *uid*, *cid*, *subjective_comprehension*, *gaze_x*, *gaze_y*, *left_gaze_x*, *left_gaze_y*, *right_gaze_x*, *right_gaze_y*, *pupil_diameter*, *left_pupil*, *right_pupil*, *pupil_diameter_r_l_difference*

TABLE 4. An example of the segmentation annotations created by the authors. One round is split into three segments.

Location	Amount	Gender	Age (Mean)	Major of the study in the university
Germany	10 Participants	6 Male & 4 Female	24 - 38 (27.3)	7 Computer Science, 2 Psychology, 1 Physics
Japan	10 Participants	8 Male & 2 Female	23 - 25 (23.5)	10 Computer Science
Total	20 Participants	14 Male & 6 Female	23 - 38 (25.3)	17 Computer Science, 2 Psychology, 1 Physics

those directly influencing the model's predictive capabilities, instilling comprehension in the robustness of our model. System-dependent features are specific to the Tobii eye-tracker, and other irrelevant attributes, such as duplicated timestamps and raw system-related attributes, were removed.

1) FIXATION

Fixations occur when the gaze remains relatively stable over a specific area of interest for a significant duration, suggesting that the viewer focuses intently on that screen segment. This stability often correlates with deeper cognitive processing and engagement, making it valuable for assessing confidence and attention levels. Fixations were quantified by setting a minimum duration threshold of 150 milliseconds and a maximum dispersion threshold within a 100-pixel radius [39]. This approach ensures that only meaningful gaze data, indicative of cognitive engagement, are considered, excluding random or fleeting eye movements.

While using fixation as the selected features, the rest of the features presented in Table 3 used the mean value of each feature. Annotations such as *uid*, *cid*, and *subjective_comprehension* were kept the same. For *#timestamp*, we keep the fixation's start and end value of the fixation. Lastly, *second*, which is the time-lapse from the start *#timestamp*, is also keeping the first and the last *second* when fixation ends.

2) RIGHT AND LEFT PUPIL DIAMETER DIFFERENCE

Kucewicz et al. [40] has mentioned that pupil diameter increases when a human tries to memorize knowledge in a brain [40]. Nobukawa et al. [41] mention that calculating right and left pupil diameter measures the cognitive state accurately [41]. Having an accurate pupil diameter as a feature requires differences between right and left pupil diameters.

Right and left pupil diameter difference indicates cognitive load and emotional response, as variations in pupil size can reflect changes in mental effort and emotional state. A key feature of pupillometry is the difference between the pupils, which offers a nuanced view of physiological responses. Intricately linked to cognitive processes during learning activities, these responses can be accurately measured using pupillometry, highlighting their potential in this domain.

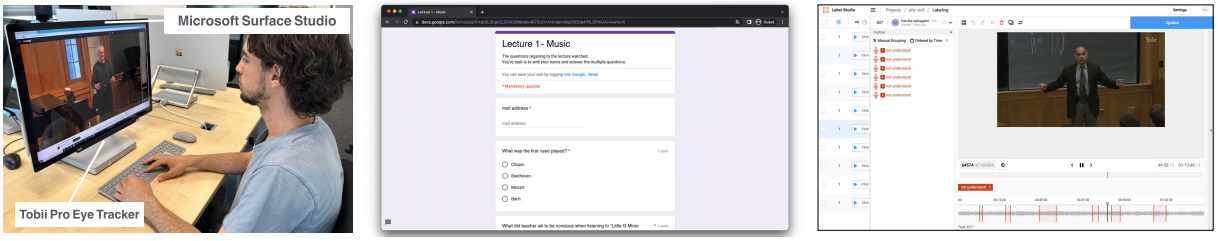


FIGURE 2. Experiment condition. Tobii eye-tracker collects gaze data, and the Windows Surface Studio webcam collects facial recordings. Participants answer the survey using Google Forms and use *LabelStudio* to annotate their subjective comprehension state.

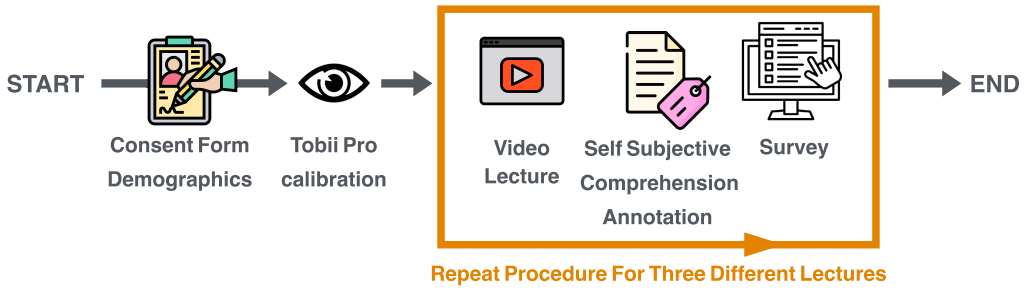


FIGURE 3. Experiment workflow. Participants watched three lectures and then conducted subjective comprehension annotations and answered a survey.

E. MODEL ARCHITECTURE

Models were trained using features such as pupil diameter, gaze coordinates, and fixation durations. The following architectures were employed:

1) GAUSSIAN NAIVE BAYES (GNB), DECISION TREE (DT), AND RANDOM FOREST (RF)

For classification, we used GNB, DT, and RF models. The RF model was configured with 100 trees, Gini impurity as the criterion, and a maximum of seven features per split.

2) LONG SHORT-TERM MEMORY (LSTM)

The LSTM model consisted of four layers (128, 64, 32, and 16 units) with a 30% dropout rate after each layer to prevent overfitting. The final dense layer used a sigmoid activation function to predict comprehension levels. Training employed the Adam optimizer with an adaptive learning rate scheduler, and early stopping was applied to halt training when validation performance plateaued.

3) TRANSFORMER

We employed a Transformer encoder for temporal analysis. Input data consisted of short time-series segments, with positional encoding applied to capture temporal order. Each segment was processed through Transformer encoder blocks featuring multi-head self-attention to assess the importance of different time steps. A global average pooling layer condensed temporal data into a feature vector, which was passed through dense layers to predict comprehension using a sigmoid activation function.

IV. DATA COLLECTION

The data collection process is a crucial component of this research, as it forms the foundation for developing and evaluating the EyeUnderstand application. The study focuses on capturing participants' eye-tracking data while they engage with educational video content and subsequently predicting their subjective comprehension levels through deep learning models.

A. PARTICIPANTS

In this study, we recruited 20 participants in Germany and Japan. Table 4 shows the demographic information of participants. We recruit participants in Germany who use English as their primary language at university or in their company. For participants in Japan, we recruit those who study or work mainly in Japanese. Participants in Japan and Germany completed consent forms before the study, which allowed them to opt out at any time during the study. Consideration of General Data Protection Regulation (GDPR) in the consent form for participants in Germany. Participants in both countries can opt out of the experiment at any time.

B. EXPERIMENTAL SETTINGS

Figure 2 shows the detail of the experimental settings. Participants get an explanation of the experiment's purpose and the use of the collected data. Once they confirm, participants sign the consent form and write demographic information such as gender, age, and significance of the study. Then, participants do a calibration on the eye-tracker. After calibration, the participant starts watching a lecture

video. Once the participant finishes watching, follow the post-process, which is writing a survey and making a time segment annotation of when the user feels *dropout* in the video lecture. Participants repeat this procedure with two more video lectures.

We asked participants to watch each video on different days to avoid fatigue during continuous trials. The lectures are in music, physics, and computer science. In this study, we conducted experiments in the same room in a controlled manner in each country.

C. EXPERIMENT WORKFLOW

Figure 3 illustrates the experiment workflow. The experiment workflow is the following: (1) Participants were briefed on the study's purpose and the intended use of the data. (2) After confirming their understanding, they signed the consent form and provided demographic information (e.g., gender, age, and field of study). (3) A calibration was performed using the eye-tracker. (4) Participants watched a lecture video, completed a post-survey, and annotated time segments where they felt dropout. (5) This process was repeated for two additional videos on separate days to avoid fatigue. The lecture videos covered three domains: music, physics, and computer science. Experiments were conducted in controlled environments in Germany and Japan to ensure consistency.

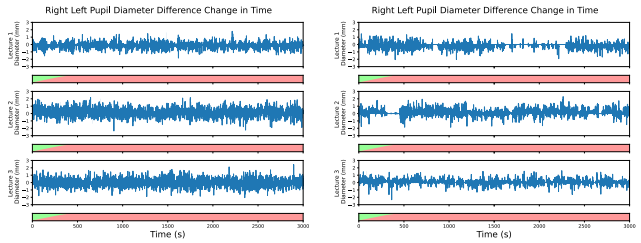
The experiments were conducted in a controlled laboratory environment to ensure the consistency and accuracy of the eye-tracking data collected. The eye-tracking data was collected while participants watched educational videos, allowing us to map fluctuations in gaze and pupil response to varying levels of self-reported subjective comprehension. The room was uniformly lit with neutral luminance to reduce the impact of external lighting variations on pupil dilation, a factor known to influence eye-tracking measurements [42]. The importance of controlling environmental factors has been emphasized in eye-tracking literature, as lighting and screen brightness variations can significantly affect pupil size and gaze stability [43].

V. RESULT AND DISCUSSION

This section will describe the dataset balance and comprehension estimation results in detail.

A. DATASET BALANCE

In this section, we summarize the collected dataset. This section is designed to provide valuable support for future researchers contemplating the utilization of our publicly available datasets. Figure 4 shows an output of gaze data (right and left pupil diameter difference) and self-reported *subjective_comprehension* state compared with time. The red-colored rectangle indicates a time duration when the participant self-reported as *dropout* in the video lecture. The time duration of the *dropout* state differs between the example of *Participant 5* and *Participant 8* as an example shows.



(a) Participant 5 **(b)** Participant 8

FIGURE 4. Sample of different participants' gaze data (right and left pupil diameter difference) with self-reported comprehension annotation.The green rectangular parts represent the confident segment, and the red ones represent the dropout segment.As the sample shows, there are significant differences in the count of self-report comprehension annotations.

Furthermore, we identified distinct characteristics associated with the annotation labels. Specifically, we investigated the minimum, maximum, and total duration of *dropout* time within the first 50 minutes of the video. Remarkably, participant P5 exhibited the minimum total *dropout* time during lecture 2, while participants P12, P14, and P17 in lecture 6 all recorded 0.0 seconds within a total time-frame of 3000 seconds (50 minutes). In contrast, the maximum total *dropout* time among all participants was recorded by participant P8 in lecture 2, amounting to 2358.5 seconds within the 3000-second (50-minute) time frame. A post-survey analysis indicated that the complexity of the terminology was cited as a prominent difficulty during the lecture. In contrast, providing comprehensible examples and diagrams was attributed to heightened participant subjective comprehension.

B. SUBJECTIVE COMPREHENSION STATE ESTIMATION

In this section, we present the results of self-reported subjective comprehension estimation using gaze data. Comprehension estimation is binary, classified as *True* or *False*. We applied several machine learning models to the dataset, employing leave-one-participant-out cross-validation (LOPOCV). The results are summarized in Table 5. Additionally, we compared the effectiveness of various input features across machine-learning and deep-learning models.

Initially, pupil diameter and gaze coordinates were used as baseline features since the Tobii eye-tracker directly collects these data points (see Table 3). Using these baseline features, LOPOCV yielded an F1 score of 0.515 with the Random Forest classifier, the best machine learning model. The deep-learning model (LSTM) achieved a significantly higher F1 score of 0.865.

Next, we incorporated the difference between right and left pupil diameters, which is known to measure cognitive state [41] accurately. Adding this feature improved the LSTM model's performance, achieving an average LOPOCV F1 score of 0.886. This result indicates that the pupil diameter difference is valuable for estimating subjective comprehension levels.

TABLE 5. Performance Metrics of comprehension estimation for Gaussian Naive Bayes, Decision Tree, Random Forest, LSTM, and Transformer against leave-one-participant-out cross-validation (LOPOCV).

Dataset Features	Model	Accuracy	Precision	Recall	F1 Score
Pupil Diameter, Gaze Coordinates	Gaussian Naive Bayes	0.518	0.522	0.519	0.500
	Decision Tree	0.510	0.510	0.510	0.505
	Random Forest	0.519	0.519	0.519	0.515
	LSTM	0.778	0.805	0.958	0.865
	Transformer	0.793	0.804	0.981	0.875
Pupil Diameter, Gaze Coordinates, Right-Left Pupil Diameter Difference	Gaussian Naive Bayes	0.502	0.517	0.502	0.359
	Decision Tree	0.514	0.514	0.514	0.513
	Random Forest	0.517	0.517	0.517	0.517
	LSTM	0.754	0.818	0.903	**0.886**
	Transformer	0.786	0.810	0.960	0.869
Pupil Diameter, Gaze Coordinates, Right-Left Pupil Diameter Difference, Fixation Duration	Gaussian Naive Bayes	0.509	0.515	0.509	0.457
	Decision Tree	0.524	0.524	0.524	0.524
	Random Forest	0.537	0.537	0.537	0.537
	LSTM	0.724	0.810	0.860	0.818
	Transformer	0.705	0.810	0.866	0.808

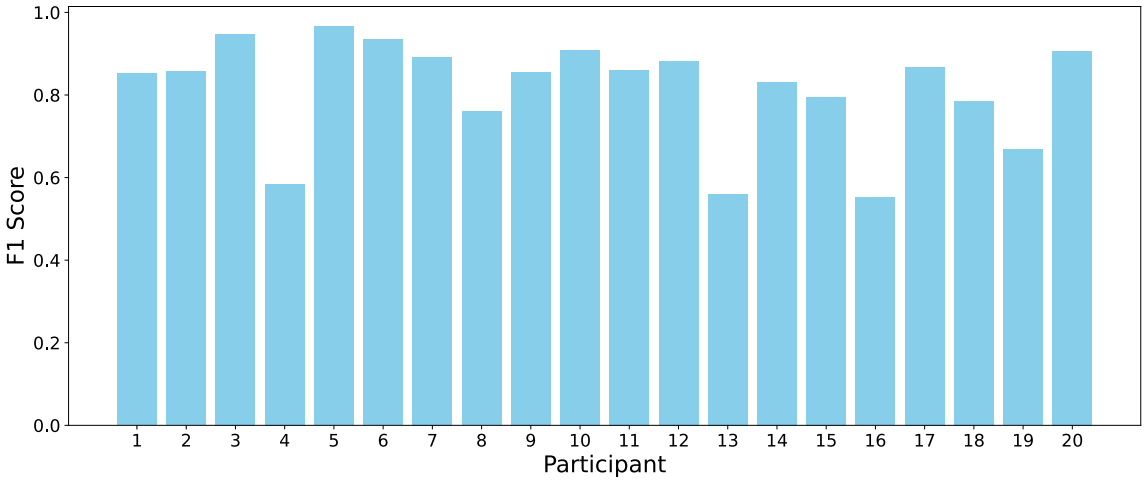


FIGURE 5. Best perform LOPOCV (leave-one-participant-out cross-validation) result of estimating comprehension using LSTM.

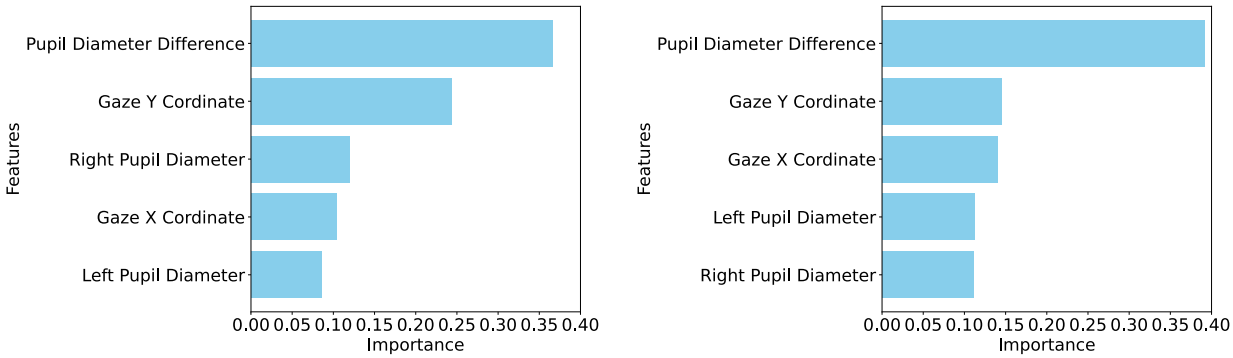
We then included fixation duration as an additional feature, as fixation is closely associated with participants' attention during video viewing. However, adding fixation duration reduced the LOPOCV F1 score for the LSTM model to 0.818. This suggests that fixation duration may introduce ambiguity in subjective comprehension state classification. Prolonged fixation can indicate either high interest or difficulty in following content.

Figure 5 displays individual participant subjective comprehension estimation results for the feature combination yielding the highest average F1 score. Some participants (e.g., P4, P14, and P16) exhibited lower scores. Reviewing the webcam recordings from the Windows Surface Studio, we observed these participants frequently moving closer or further from the screen, resulting in misalignment with the calibrated eye-tracker. This movement impacted the data's consistency.

In conclusion, the difference between right and left pupil diameters proved helpful for estimating subjective comprehension. However, frequent upper-body movement among some participants can reduce measurement accuracy with the mounted Tobii eye-tracker.

C. CULTURAL COMPARISON

To understand the influence of cultural differences on gaze behavior and feature importance during online video lectures, we analyzed data separately for German and Japanese participants. Feature importance was determined using a machine learning model trained on various eye-tracking



(a) German Participants **(b)** Japanese Participants

FIGURE 6. Top five important feature for predicting subjective comprehension level among German and Japanese participants.

TABLE 6. Comparison between participants in Germany and Japan for leave-one-participant-out cross-validation using LSTM.

Participant Group	Mean Accuracy	Mean Precision	Mean Recall	Mean F1 Score
Germany	0.644	0.810	0.785	0.757
Japan	0.730	0.826	0.841	0.820

and pupillometry features to predict participants' subjective comprehension.

The feature importance graphs (FIGURE 6) illustrate the relative significance of each feature in the model's predictions for both groups. Each bar represents the contribution of a specific feature to the model's predictive accuracy, with the y-axis showing relative importance and the x-axis listing the features.

Pupil Diameter Right-Left Difference holds notably higher importance for Japanese participants, where it is the second most critical feature, contributing over 0.2 to the model's predictions. This is less prominent in the German group. This difference may indicate a higher sensitivity to cognitive load or emotional response during learning among Japanese participants, potentially reflecting a holistic and integrative cognitive style prevalent in Japanese educational contexts. In such contexts, learners may actively balance emotional and cognitive responses, influencing their gaze behaviors during online lectures.

Gaze Coordinates show relatively similar importance across both groups, though they are slightly more influential in the Japanese group. This suggests that while the specific areas of the screen where participants focus their gaze affect comprehension, this influence is somewhat consistent across cultures, with only slight variations in emphasis.

Finally, *Pupil Diameter* and *Right Pupil/Left Pupil* features are of minimal importance in both groups, indicating that while pupillary responses are measured, they are less predictive of the outcomes in this educational context for both German and Japanese participants.

These cultural differences in feature importance may stem from underlying variations in educational practices and cognitive processing styles. For example, the emphasis on fixation duration in German participants could reflect a learning culture prioritizing sustained attention and deep engagement with content. Meanwhile, the significant role of pupil diameter differences in Japanese participants might suggest a learning style more attuned to balancing cognitive load and emotional processing, consistent with a holistic educational approach.

By understanding these differences, educators and developers of online educational tools can better tailor content and interaction methods to suit the learning styles of different cultural groups, potentially enhancing the effectiveness of online education across diverse audiences.

D. ENHANCEMENT WITH SLIDING WINDOWS TRAINING

Building on the results of our initial experiments, where various machine learning models were tested on gaze data for subjective comprehension estimation, we sought to optimize the training process further and improve model performance.

To achieve this, we implemented a sliding windows approach, a technique commonly used in time-series data analysis, to enhance the temporal resolution of our model inputs. The sliding windows approach allows the model to capture more nuanced temporal dependencies by training on overlapping segments of gaze data rather than treating the entire lecture as a single, static input. This method aligns well with the dynamic nature of gaze behavior, where student attention and comprehension can fluctuate significantly over short periods.

Table 6 shows a comparative analysis of the results obtained using the sliding windows approach on the datasets from Germany and Japan. These results demonstrate that the sliding windows method effectively captures the temporal dynamics of gaze behavior, leading to a more accurate

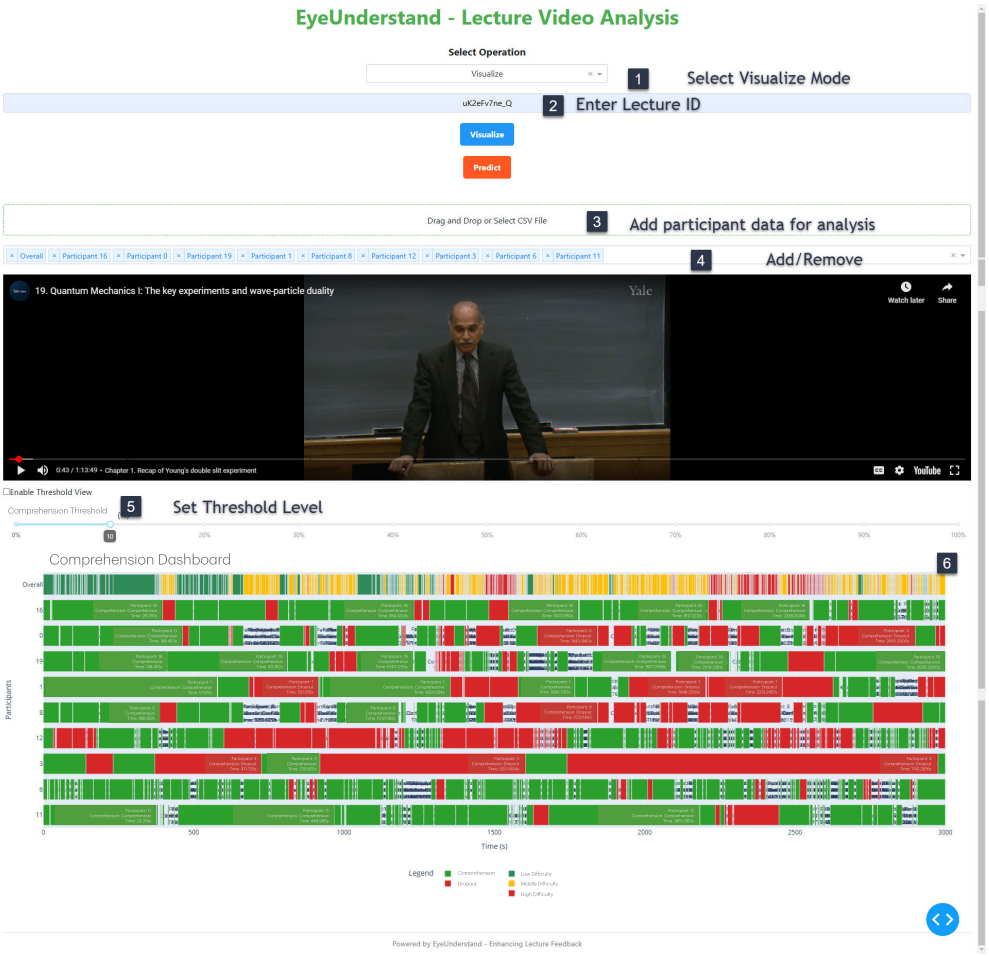


FIGURE 7. User Interface of an application: ***EyeUnderstand***.

prediction of subjective comprehension levels. Notably, the improvement in F1 scores suggests that the method helps balance precision and recall, reducing the risk of overfitting and enhancing the generalization of the models across different cultural contexts.

The higher accuracy observed in the Japan dataset could also reflect the method's adaptability to specific gaze patterns prevalent among Japanese participants. This adaptability is crucial for developing culturally responsive educational tools that cater to diverse student populations. By integrating the sliding windows approach, we improved model performance and gained deeper insights into the temporal patterns of student engagement during online lectures. This enhancement paves the way for more sophisticated and real-time feedback mechanisms for effective online education.

VI. END-USER APPLICATION

This section introduces our pilot application, *EyeUnderstand*,[4] which is shown in Figure 7. The application visualizes comprehension levels using Tobii eye-tracking data.

[4]https://anonymous.4open.science/r/eyeunderstand/

A. USER INTERFACE

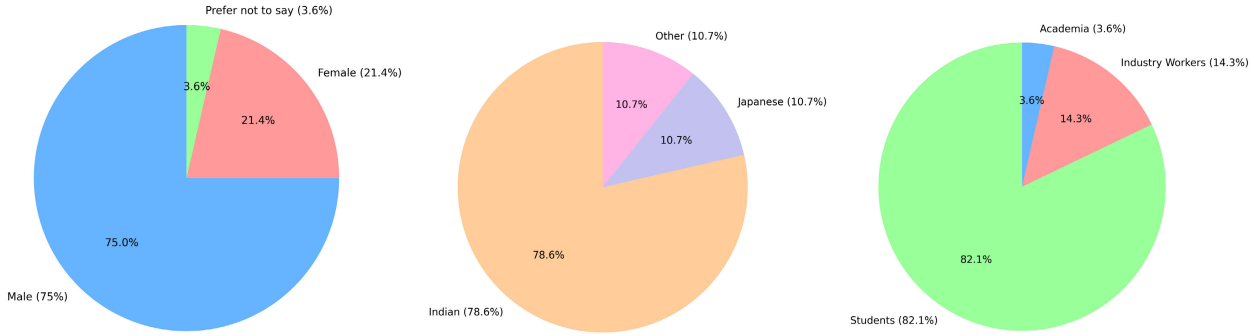
Figure 7 provides an overview of the web application which analyzes the lecture video viewers through eye-tracking. Participant first enter the lecture ID, which corresponds to the YouTube ID. Then, add the data of participant while watching the lecture collected by Tobii eye-tracker. Once inserted, the datas processing starts and visualize the comprehension levels. User can select participant IDs to select the target users and also threshold to make the comprehension level range.

B. USER STUDY

We conducted a feedback survey of the web application EyeUnderstand to gather insights about its usability, understandability, and usefulness. These insights will guide future improvements to enhance the tool's effectiveness for students and lecturers.

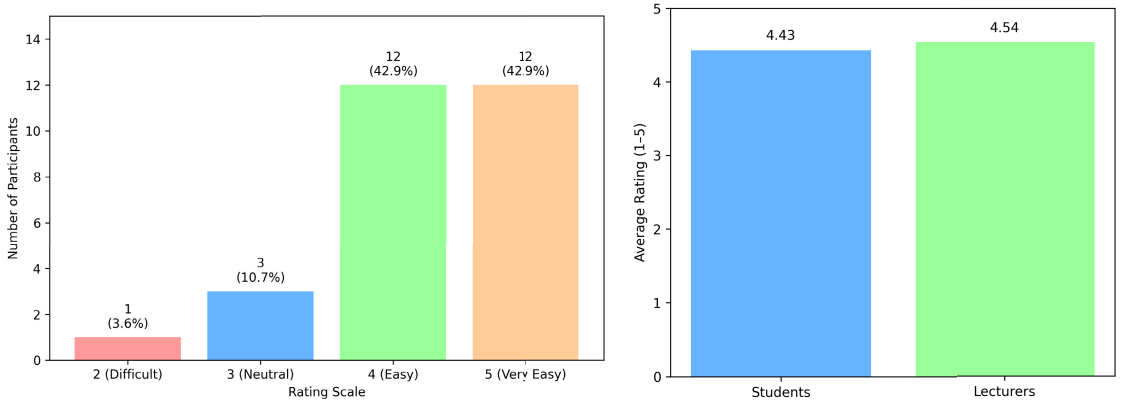
1) STATISTICS OF USER STUDY PARTICIPANT BACKGROUND

Figure 8 shows the statistics of the overall backgrounds of the participants. The survey collected responses from 28 participants, primarily students and professionals from



(a) Gender **(b)** Nationality **(c)** Occupation

FIGURE 8. **Participants distributions of the *EyeUnderstand* user study.**



(a) Difficulty level rating of the userability. **(b)** Usefulness ratings of students and lecturers.

FIGURE 9. **Overall rating of the user-study survey on usability of *EyeUnderstand*.**

academia and industry. Most participants were students aged 19 to 32 (mean age: 25.3). Key demographic distributions are as follows:

- **Gender:** Male (75%), Female (21.4%), Prefer not to say (3.6%).
- **Nationality:** Indian (78.6%), Japanese (10.7%), Other nationalities (10.7%, including Rwandan, Chilean, and Pakistani).
- **Occupation:** Students (82.1%), Industry professionals (14.3%), Academic researchers (3.6%).

Regarding familiarity with eye-tracking technology, 78.6% of participants reported familiar, 14.3% were unfamiliar, and 7.1% were unsure. These demographics suggest a predominantly young, student-oriented sample with considerable prior knowledge of eye-tracking methodologies.

2) USABILITY TESTING RESULT OF *EYEUNDERSTAND*

Figure 9 presents the overall results from the usability testing conducted during the user study. Approximately 92.9% of participants understood the application's concept, while 7.1% remained uncertain.

Participants gave navigation an average rating of 4.25 out of 5 (scale: 1 = Very Difficult, 5 = Very Easy). 42.9% (12 participants) rated navigation as Very Easy (5), 42.9% (12 participants) as Easy (4), 10.7% (3 participants) were Neutral (3), and 3.6% (1 participant) found it Difficult (2).

The average usability rating among students was 4.43 out of 5, with 89.3% rating between 4 and 5. Lecturers reported an average rating of 4.54 out of 5, with a 92.9% rating between 4 and 5. Participants specifically highlighted the following strengths:

User feedback (1): Pinpointing video timestamps where students lacked comprehension helps lecturers improve content delivery.

User feedback (2): Setting a none comprehensive threshold allows targeted review of problematic sections.

User feedback (3): The color-coded comprehension bars (green/yellow/red) are intuitive for identifying trends.

Based on this feedback, it is confirmed that *EyeUnderstand* performs effectively for its intended users.

VII. LIMITATIONS AND FUTURE WORK

This section explains limitations and future work on the comprehension estimation model and the performance of the *EyeUnderstand* application.

A. COMPREHENSION ESTIMATION MODEL

The first limitation of our study is its focus on only three lecture categories chosen to represent diverse instructional

styles. Although these categories provided a solid basis for analysis, they do not encompass the entire range of lecture methodologies. Future research should include a broader array of instructional modalities for a more comprehensive understanding.

Another limitation is the binary classification of comprehension. Comprehension inherently exists on a continuum, and a binary approach does not fully capture its varying degrees of certainty. Post-annotation also poses challenges because participants may struggle to recall uncertain moments after 50 minutes of viewing. While real-time annotation could mitigate this issue, it risks distracting participants during the task. Future work will explore human-in-the-loop annotation techniques and regression models to capture comprehension levels better.

Another limitation is the small sample size (20 participants), which may not capture the full range of individual variation. A more extensive and diverse participant pool would improve the generalizability of our findings. Due to this concern, the Transformer model underperformed in our study, but it may yield better results with more data. Expanding the dataset is thus a priority for future research.

Participant self-esteem may further influence self-annotation accuracy, as individuals with higher self-esteem might offer more nuanced responses. Future studies should investigate the influence of domain knowledge and demographic factors on comprehension estimation to account for these differences.

We plan to enhance learning outcomes by estimating comprehension levels in real-time, integrating features such as question prompts or varied lecture difficulty. Investigating the relationship between domain knowledge and eye movements will also be a key focus. Ultimately, we aim to develop an automated comprehension estimation application that delivers real-time feedback to lecture creators and provides summary insights for students.

B. IMPROVEMENTS TO OUR APPLICATION

Despite our application's current promise, its most pressing limitation is accessibility. Eye-tracking hardware remains prohibitively expensive for widespread, individual use, particularly in large classrooms or resource-constrained settings. To address this challenge, our future work will explore replacing specialized hardware with webcam-based eye tracking [44], [45], which can estimate comprehension levels [2], [46] using readily available, low-cost devices. This approach has the potential to significantly broaden our user base while lowering financial barriers and improving scalability.

Another key area for improvement lies in the user interface. Meaningful and intuitive visualization is essential to translate complex eye-tracking data into actionable insights for educators. By offering clear, time-aligned representations of student engagement and comprehension, instructors can more easily identify pivotal moments of uncertainty and success, tailoring their teaching strategies for maximum impact.

Finally, the lack of long-term evaluation remains an important gap. Future studies will examine whether this tool can become an integral part of an educator's daily workflow, not only for refining lectures but also for supporting students' study habits and deepening their comprehension. In-depth, longitudinal analyses will help us determine the true efficacy and sustainability of using eye-tracking data to enhance both teaching practices and learning outcomes.

VIII. CONCLUSION

This study collected eye-tracking data to predict comprehension levels during video lectures. Twenty participants from Germany and Japan watched 50-minute lecture videos covering three domains. Participants completed a post-survey and self-annotated their comprehension levels using the open-source tool *LabelStudio*. We applied an LSTM model to the preprocessed dataset, achieving an F1 score of 0.886 for predicting binary comprehension levels. The analysis identified pupil diameter as a significant feature for estimating comprehension in German and Japanese participants. We also introduce *EyeUnderstand*, the web-based application for visualizing the results of the comprehension estimation through eye-tracking. We recruited 28 participants for the user study. As a result, 89.3% of the students and 92.9% of the lecturers confirmed that our application is practical.

REFERENCES

[1] S. Berkovsky, R. Taib, I. Koprinska, E. Wang, Y. Zeng, J. Li, and S. Kleitman, "Detecting personality traits using eye-tracking data," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, May 2019, pp. 1–12, doi: 10.1145/3290605.3300451.

[2] X. Zhang, Y. Sugano, and A. Bulling, "Evaluation of appearance-based methods and implications for gaze-based applications," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, May 2019, pp. 1–13, doi: 10.1145/3290605.3300646.

[3] N. Srivastava, E. Velloso, J. M. Lodge, S. Erfani, and J. Bailey, "Continuous evaluation of video lectures from real-time difficulty self-report," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, May 2019, pp. 1–12, doi: 10.1145/3290605.3300816.

[4] D. Dembinsky, K. Watanabe, A. Dengel, and S. Ishimaru, "Eye movement in a controlled dialogue setting," in *Proc. Symp. Eye Tracking Res. Appl.*, New York, NY, USA, Jun. 2024, pp. 1–7, doi: 10.1145/3649902.3653337.

[5] R. Morita, K. Watanabe, J. Zhou, A. Dengel, and S. Ishimaru, "GenAIReading: Augmenting human cognition with interactive digital textbooks using large language models and image generation models," 2025, *arXiv:2503.07463*.

[6] D. M. Bunce, E. A. Flens, and K. Y. Neiles, "How long can students pay attention in class? A study of student attention decline using clickers," *J. Chem. Educ.*, vol. 87, no. 12, pp. 1438–1443, Dec. 2010.

[7] J. Kim, P. J. Guo, D. T. Seaton, P. Mitros, K. Z. Gajos, and R. C. Miller, "Understanding in-video dropouts and interaction peaks inonline lecture videos," in *Proc. 1st ACM Conf. Learn. Scale Conf.*, New York, NY, USA, Mar. 2014, pp. 31–40, doi: 10.1145/2556325.2566237.

[8] C. L. Sanches, K. Kise, and O. Augereau, "Japanese reading objective understanding estimation by eye gaze analysis," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2017, pp. 121–124.

[9] O. Augereau, K. Kunze, H. Fujiyoshi, and K. Kise, "Estimation of English skill with a mobile eye tracker," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., Adjunct*, Sep. 2016, pp. 1777–1781.

[10] S. Ishimaru, T. Maruichi, K. Kise, and A. Dengel, "Gaze-based self-confidence estimation on multiple-choice questions and its feedback," in *Proc. Symp. Emerg. Res. Asia Asian Contexts Cultures*, New York, NY, USA, Apr. 2020, p. 8, doi: 10.1145/3391203.3391227.
[11] Z. Zhai, X. Chen, Y. Zhao, L. Zhao, J. Qian, and J. Wu, "SmartCamera: Realtime video stream-oriented action recognition platform in edge environment," in *Proc. Adjunct ACM Int. Joint Conf. Pervasive Ubiquitous Comput. Proc. ACM Int. Symp. Wearable Comput.*, New York, NY, USA, Sep. 2021, pp. 88–89, doi: 10.1145/3460418.3479303.
[12] M. Dimiccoli, J. Marín, and E. Thomaz, "Mitigating bystander privacy concerns in egocentric activity recognition with deep learning and intentional image degradation," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 4, pp. 1–18, Jan. 2018, doi: 10.1145/3161190.
[13] K. Ahuja, D. Kim, F. Xhakaj, V. Varga, A. Xie, S. Zhang, J. E. Townsend, C. Harrison, A. Ogan, and Y. Agarwal, "EduSense: Practical classroom sensing at scale," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 3, pp. 1–26, Sep. 2019, doi: 10.1145/3351229.
[14] K. Watanabe, Y. Soneda, Y. Matsuda, Y. Nakamura, Y. Arakawa, A. Dengel, and S. Ishimaru, "DisCaaS: Micro behavior analysis on discussion by camera as a sensor," *Sensors*, vol. 21, no. 17, 2021, Art. no. 5719. [Online]. Available: https://www.mdpi.com/1424-8220/21/17/5719
[15] S. Kawato and J. Ohya, "Real-time detection of nodding and head-shaking by directly detecting and tracking the 'between-eyes,'" in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Jun. 2000, pp. 40–45.
[16] C. Chen, Y. Arakawa, K. Watanabe, and S. Ishimaru, "Quantitative evaluation system for online meetings based on multimodal microbehavior analysis," *Sensors Mater.*, vol. 34, no. 8, pp. 3017–3027, 2022.
[17] C. Liu, Z. Dong, L. Huang, W. Yan, X. Wang, D. Fang, and X. Chen, "TagSleep3D: RF-based 3D sleep posture skeleton recognition," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 8, no. 1, pp. 1–28, Mar. 2024, doi: 10.1145/3643512.
[18] R. H. Venkatnarayan and M. Shahzad, "Gesture recognition using ambient light," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 1–28, Mar. 2018, doi: 10.1145/3191772.
[19] R. Zhu, L. Shi, Y. Song, and Z. Cai, "Integrating gaze and mouse via joint cross-attention fusion net for students' activity recognition in E-learning," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 7, no. 3, pp. 1–35, Sep. 2023, doi: 10.1145/3610876.
[20] M. Sauter, T. Hirzle, T. Wagner, S. Hummel, E. Rukzio, and A. Huckauf, "Can eye movement synchronicity predict test performance with unreliably-sampled data in an online learning context?" in *Proc. Symp. Eye Tracking Res. Appl.*, New York, NY, USA, Jun. 2022, pp. 1–5, doi: 10.1145/3517031.3529239.
[21] C. De-La-Peña, "Eye-tracking contribution on processing of (implicit) reading comprehension," *J. New Approaches Educ. Res.*, vol. 13, no. 1, p. 13, Aug. 2024.
[22] L. Huang, J. Ouyang, and J. Jiang, "The relationship of word processing with L2 reading comprehension and working memory: Insights from eye-tracking," *Learn. Individual Differences*, vol. 95, Apr. 2022, Art. no. 102143.
[23] N. Srivastava, S. Nawaz, J. Newn, J. Lodge, E. Velloso, S. M. Erfani, D. Gasevic, and J. Bailey, "Are you with me? Measurement of learners' video-watching attention with eye tracking," in *Proc. 11th Int. Learn. Anal. Knowl. Conf.*, New York, NY, USA, Apr. 2021, pp. 88–98, doi: 10.1145/3448139.3448148.
[24] J. Santhosh, A. Dengel, and S. Ishimaru, "Gaze-driven adaptive learning system with ChatGPT-generated summaries," *IEEE Access*, vol. 12, pp. 173714–173733, 2024.
[25] R. Kawamura, S. Shirai, M. Aizadeh, N. Takemura, and H. Nagahara, "Estimation of wakefulness in video-based lectures based on multimodal data fusion," in *Adjunct Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. Proc. ACM Int. Symp. Wearable Comput.*, New York, NY, USA, Sep. 2020, pp. 50–53, doi: 10.1145/3410530.3414386.
[26] Y. Abdelrahman, E. Velloso, T. Dingler, A. Schmidt, and F. Vetere, "Cognitive heat: Exploring the usage of thermal imaging to unobtrusively estimate cognitive load," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 3, pp. 1–20, Sep. 2017, doi: 10.1145/3130898.
[27] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "DAiSEE: Towards user engagement recognition in the wild," 2016, *arXiv:1609.01885*.
[28] M. Singh, X. Hoque, D. Zeng, Y. Wang, K. Ikeda, and A. Dhall, "Do I have your attention: A large scale engagement prediction dataset and baselines," 2023, *arXiv:2302.00431*.
[29] K. Watanabe, T. Sathyanarayana, A. Dengel, and S. Ishimaru, "EnGauge: Engagement gauge of meeting participants estimated by facial expression and deep neural network," *IEEE Access*, vol. 11, pp. 52886–52898, 2023.
[30] K. Watanabe, A. Dengel, and S. Ishimaru, "Metacognition-EnGauge: Real-time augmentation of self-and-group engagement levels understanding by gauge interface in online meetings," in *Proc. Augmented Hum. Int. Conf.*, Apr. 2024, pp. 301–303.
[31] A. S. Sharma, M. R. Amin, and M. Fuad, "Augmenting online classes with an attention tracking tool may improve Student engagement," 2022, *arXiv:2210.07286*.
[32] N. Tanaka, K. Watanabe, S. Ishimaru, A. Dengel, S. Ata, and M. Fujimoto, "Concentration estimation in online video lecture using multimodal sensors," in *Proc. Companion ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, New York, NY, USA, Oct. 2024, pp. 71–75, doi: 10.1145/3675094.3677587.
[33] P. Kar, S. Chattopadhyay, and S. Chakraborty, "Gestatten: Estimation of user's attention in mobile MOOCs from eye gaze and gaze gesture tracking," *Proc. ACM Hum.-Comput. Interact.*, vol. 4, no. 1, pp. 1–32, Jun. 2020, doi: 10.1145/3394974.
[34] Y. Abdelrahman, A. A. Khan, J. Newn, E. Velloso, S. A. Safwat, J. Bailey, A. Bulling, F. Vetere, and A. Schmidt, "Classifying attention types with thermal imaging and eye tracking," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 3, pp. 1–27, Sep. 2019, doi: 10.1145/3351227.
[35] M. Burch, "Gaze-based monitoring in the classroom," in *Proc. Symp. Eye Tracking Res. Appl.*, New York, NY, USA, May 2023, pp. 1–3, doi: 10.1145/3588015.3589198.
[36] R. E. Bixler and S. K. D'Mello, "Crossed eyes: Domain adaptation for gaze-based mind wandering models," in *Proc. ACM Symp. Eye Tracking Res. Appl.*, New York, NY, USA, May 2021, pp. 1–12, doi: 10.1145/3448017.3457386.
[37] F. Zermiani, A. Bulling, and M. Wirzberger, "Mind wandering trait-level tendencies during lecture viewing: A pilot study," in *Proc. Symp. Eye Tracking Res. Appl.*, New York, NY, USA, Jun. 2022, pp. 1–7, doi: 10.1145/3517031.3529241.
[38] M. Tkachenko, M. Malyuk, A. Holmanyuk, and N. Liubimov. (2020). *Label Studio: Data Labeling Software*. [Online]. Available: https://github.com/heartexlabs/label-studio
[39] G. Buscher, E. Cutrell, and M. R. Morris, "What do you see when you're surfing?: Using eye tracking to predict salient regions of web pages," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, Apr. 2009, pp. 21–30, doi: 10.1145/1518701.1518705.
[40] M. T. Kucewicz, J. Dolezal, V. Kremen, B. M. Berry, L. R. Miller, A. L. Magee, V. Fabian, and G. A. Worrell, "Pupil size reflects successful encoding and recall of memory in humans," *Sci. Rep.*, vol. 8, no. 1, p. 4949, Mar. 2018.
[41] S. Nobukawa, A. Shirama, T. Takahashi, T. Takeda, H. Ohta, M. Kikuchi, A. Iwanami, N. Kato, and S. Toda, "Pupillometric complexity and symmetricity follow inverted-U curves against baseline diameter due to crossed locus coeruleus projections to the Edinger–Westphal nucleus," *Frontiers Physiol.*, vol. 12, Feb. 2021, Art. no. 614479.
[42] I. Daguet, D. Bouhassira, and C. Gronfier, "Baseline pupil diameter is not a reliable biomarker of subjective sleepiness," *Frontiers Neurol.*, vol. 10, p. 108, Feb. 2019.
[43] J. Xu, Y. Wang, F. Chen, and E. Choi, "Pupillary response based cognitive workload measurement under luminance changes," in *Proc. IFIP Conf. Hum.-Comput. Interact.*, 2011, pp. 178–185.
[44] V. Shah, K. Watanabe, B. B. Moser, and A. Dengel, "PupilSense: A novel application for webcam-based pupil diameter estimation," 2024, *arXiv:2407.11204*.
[45] V. Shah, B. Moser, K. Watanabe, and A. Dengel, "Webcam-based pupil diameter prediction benefits from upscaling," in *Proc. 17th Int. Conf. Agents Artif. Intell.*, 2025, pp. 376–385.
[46] A. Bhatt, K. Watanabe, A. Dengel, and S. Ishimaru, "Appearance-based gaze estimation with deep neural networks: From data collection to evaluation," *Int. J. Activity Behav. Comput.*, vol. 10, no. 1, pp. 1–15, 2024.

KO WATANABE was born in Hiroshima, Japan, in 1994. He received the B.E. degree in mechanical engineering from Tokyo University of Agricultural and Technology, Tokyo, Japan, in 2017, the M.E. degree from Nara Institute of Science and Technology, Japan, in 2019, and the Ph.D. degree in computer science from the University of Kaiserslautern-Landau. He was also a Software Engineer with DeNA, Tokyo. His current research focuses on the investigation of technologies that augment human intellect. His awards and honors include the Best Short Paper Award in ETRA 2025. He also enrolled as a Poster and the Demo Chair in Augmented Humans (AHs), in 2025 and 2026.

GITESH GUND was born in India, in 1995. He received the B.E. degree in computer engineering from Pune Institute of Computer Technology, Pune, India, in 2017. He is currently pursuing the M.Sc. degree in computer science with RPTU Kaiserslautern-Landau, Germany, specializing in intelligent systems and software engineering. He then was a Software Engineer with GS Laboratory, Pune, from 2017 to 2022, where he contributed to machine learning-based behavioral threat detection systems. He is an Application Security Engineer with Serviceware GmbH, Idstein, Germany. His current research focuses on integrating eye-tracking and deep learning to enhance lecturer feedback by estimating student comprehension in online education. His research interests include eye-tracking, deep learning, human-centered computing, and educational technologies.

JAYASANKAR SANTHOSH was born in Kerala, India, in 1992. He received the bachelor's degree in computer science from Mahatma Gandhi University, India, in 2014, and the master's degree in computer science from the Technical University of Kaiserslautern, Germany, in 2018. Since 2019, he has been a Ph.D. Researcher with the German Research Center for Artificial Intelligence (DFKI) GmbH, Kaiserslautern. He is currently a member of the Immersive Quantified Learning Laboratory (IQL Laboratory) and the Smart Data and Knowledge Services (SDS) Department, DFKI GmbH. He has been a Teaching Assistant with RPTU Kaiserslautern-Landau, since 2022. From 2017 to 2019, he was a Research Assistant with DFKI GmbH. He has published articles in prestigious venues, such as IEEE Access, *Activity and Behavior Computing*, Ubicomp, and IUI conferences. His research interests include deep learning-based affective state recognition, time series analysis, assessing student involvement in e-learning, feedback-based intervention, and adaptive learning interfaces. He is a professional member of the Association for Computing Machinery (ACM).

HARUKA SAKAGAMI was born in Osaka, Japan, in 1999. She received the B.E. degree in social information science from Kyoto Women's University, Kyoto, Japan, in 2022, and the M.E. degree from Nara Institute of Science and Technology, Nara, Japan, in 2024. During her M.E. degree, she studied the use of eye gaze to estimate a person's comprehension of learning as an Internship Student with the Immersive Quantified Learning Laboratory (IQL Laboratory), German Research Center for Artificial Intelligence (DFKI) GmbH, Germany, from 2023 to 2024. She is currently an Application Engineer in Japan.

YUKI MATSUDA (Member, IEEE) was born in 1993. He received the B.E. degree from the Advanced Course of Mechanical and Electronic System Engineering, National Institute of Technology, Akashi College, Japan, in 2015, and the M.E. and Ph.D. degrees from the Graduate School of Information Science, Nara Institute of Science and Technology, Japan, in 2016 and 2019, respectively. During his Ph.D. degree, he studied as a Visiting Researcher with Ulm University, Germany, from 2017 to 2018. His current research interests include urban sensing, civic computing, ubiquitous computing, and affective computing. He is a member of IPSJ. He received the IEEE PerCom Best Demonstration Award, in 2019.

ANDREAS DENGEL received the Diploma degree in CS from TUK and the Ph.D. degree from the University of Stuttgart. He is currently the Scientific Director of DFKI GmbH, Kaiserslautern. In 1993, he became a Professor of computer science with TUK, where he holds the Chair of Knowledge-Based Systems. Since 2009, he has been appointed as a Professor (Kyakuin) with the Department of Computer Science and Information Systems, Osaka Prefecture University. He also worked at IBM, Siemens, and Xerox Parc. He is the co-editor of international computer science journals and has written or edited 12 books. He is the author of more than 300 peer-reviewed scientific publications and supervised more than 170 master's and Ph.D. theses. His main scientific research interests include pattern recognition, document understanding, information retrieval, multimedia mining, semantic technologies, and social media. He is a fellow of IAPR and received many prominent international awards. He is a member of several international advisory boards, has chaired major international conferences, and founded several successful start-up companies.

SHOYA ISHIMARU (Member, IEEE) was born in Ehime, Japan, in 1991. He received the B.E. and M.E. degrees in electrical engineering and information science from Osaka Prefecture University, Japan, in 2014 and 2016, respectively, and the Ph.D. degree (summa cum laude) in engineering from RPTU Kaiserslautern-Landau, Germany, in 2019. He has been a Professor with the Research Institute of Innovation-Academy Co-Creation, Osaka Metropolitan University (OMU), Japan, since 2025. In addition, he has been a Project Professor with the Graduate School of Informatics, OMU, since 2023, and a Researcher with the Keio Media Design Research Institute, since 2014. He was a Junior Professor with RPTU Kaiserslautern-Landau, from 2021 to 2023, and a Senior Researcher with the German Research Center for Artificial Intelligence (DFKI) GmbH, Germany, from 2019 to 2023. His research investigates human–computer interaction, machine learning, and cognitive psychology with the aim of amplifying human intelligence. His awards and honors include the Best Presentation Award at the Asian CHI Symposium 2020, the Poster Track Honorable Mention at UbiComp/ISWC 2018, and the MITOU Super Creator which is a title given to outstanding software developers (around ten people per year) by the Ministry of Economy, Trade, and Industry, Japan.