International Conference on Artificial Reality and Telexistence
Eurographics Symposium on Virtual Environments (2021)
J. Orlosky, D. Reiners, and B. Weyers (Editors)

# Deep Learning on Eye Gaze Data to Classify Student Distraction Level in an Educational VR Environment

Sarker M. Asish, Ekram Hossain, Arun K. Kulshreshth, and Christoph W. Borst
University of Louisiana at Lafayette

## Abstract

*Educational VR may increase engagement and retention compared to traditional learning, for some topics or students. However, a student could still get distracted and disengaged due to stress, mind-wandering, unwanted noise, external alerts, etc. Student eye gaze can be useful for detecting distraction. For example, we previously considered gaze visualizations to help teachers understand student attention to better identify or guide distracted students. However, it is not practical for a teacher to monitor a large numbers of student indicators while teaching. To help filter students based on distraction level, we consider a deep learning approach to detect distraction from gaze data. The key aspects are: (1) we created a labeled eye gaze dataset (3.4M data points) from an educational VR environment, (2) we propose an automatic system to gauge a student's distraction level from gaze data, and (3) we apply and compare three deep neural classifiers for this purpose. A proposed CNN-LSTM classifier achieved an accuracy of 89.8% for classifying distraction, per educational activity section, into one of three levels.*

**CCS Concepts**
• ***Computing methodologies*** → *Deep learning;* ***Virtual reality;*** • ***Applied computing*** → *Education;*

## 1 Introduction

Recent consumer devices can provide immersive virtual reality experiences with sufficient quality and affordability for home or school use. Potential benefits of VR for education include increased engagement and motivation of students, better communication of size and spatial relationships of modeled objects, and stronger memories of the experience. In a real classroom, teachers have a sense of the audience's engagement and actions from cues such as body movements, eye gaze, and facial expressions. This awareness is significantly reduced in a VR environment because a teacher can't see students directly. Additionally, students get distracted in VR due a variety of reasons such as noise in the real environment around the student, distractions from other avatars, or checking external tools [YB21].

We previously explored gaze visualizations to help teachers monitor students' attention when guiding VR field trips [RAF*20]. However, continual visualization of gaze from many students is not practical because a teacher would monitor many cues in a VR classroom while teaching. A solution is to automatically filter students based on attention level and visualize details only for students who may need extra consideration, allowing a teacher to monitor a large class with less effort. Broussard et al. [BRKB21] proposed a teacher interface, for a remote VR class, to show information about student actions, attention, and temperament. Its information display could sort or filter students based on student importance derived from attention level. It incorporated attention detection based only on gaze angle to target objects. Improved automatic distraction detection is needed for such interfaces.

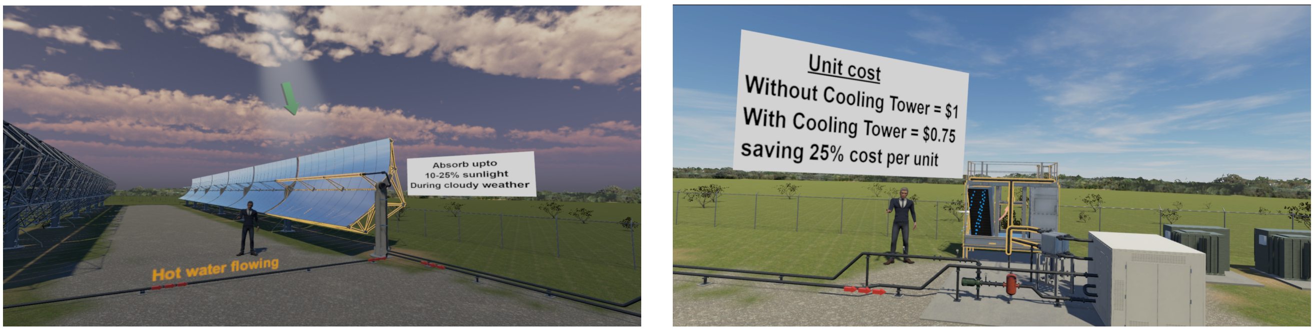
Gaze-data has been used in the past for detecting engagement levels [DOWH12,NI10], stress [JA18], confusion [SC20], and cognitive abilities [BLG*20] in non-VR educational applications. A few other previous studies [BMTM20, APGVG10, Ayr06] support the hypothesis of an existing relationship between gaze features and distraction. Most of the previous VR research has not examined the level of distraction during a class environment. The relationship between gaze features and distraction is complex due to individual variability. Therefore, the traditional statistical methods of data analysis are not suitable to handle such complex data.

We propose a deep learning system that identifies the distraction level of a student based on gaze data in VR. We designed an educational VR environment with various components (avatar, audio, text slides, and animations) to assist learning. We collected gaze data of participants using this VR environment, to train three machine learning models to detect distraction level (low, mid or high). We tested the resulting classification accuracy. Our system could detect distraction level of a student on a per-session basis and is a step towards developing a realtime distraction detection system.

## 2 Related Work

Educational VR has been mostly used for procedural motor skill training in fields such as aviation and medicine [GC04, OD17]. In

DOI: 10.2312/egve.20211326



(a) An avatar describing a solar panel.

(b) An avatar explaining the cooling process.

**Figure 1:** *Educational VR environment to explain how a solar field generates power. An avatar explains different components using audio, animations and text slides.*

the last decade, immersive VR has been studied in other educational contexts, such as safety training [BC17], and training public security personnel [BMC15]. VR has provided new opportunities for visualizing and interacting with abstract learning content (e.g., molecular structures [WMT*19]) as well as simulation applications that would be hazardous to practice in real life (e.g., hazardous situation) [MN11].

Recent research, specifically in the field of psychology and human-computer interaction, suggests that text and audio based learning is effective depending on the task. According to Modality Principle, on-screen speech is superior to on-screen text for learning [But14] in terms of complex graphic representations that include dual-channel processing in working memory. Sarune et al. [BMTM20] found that reading text from a virtual book is superior to listening for learning, specifically for knowledge retention, but found no significant differences for knowledge transfer. In some cases, VR leads to a higher sense of presence and keeps users engaged with educational content [MTM19,MOM19,RBD18]. However, text-based presentation could lead to higher cognitive load and less learning in VR [MTM19].

Psychological research found that many students use their cellphones to browse the internet or shop online while attending a class [MPL*18]. Students may also use a cellphone for social media or other non-academic activities while learning in the classroom, likely reducing knowledge retention. Research suggests that in complex or multitasking environments, attention can be diminished by shifting from one activity to another [DBL*20,SM12,RSG*15]. Additionally, students could easily be distracted in a VR environment as the entire space is open to look at and there may be many interesting objects that catch a student's attention [GBMT13].

Eye gaze has been studied for decades for a wide range of applications [Duc02] such as medical (e.g., eye surgery [MEK*01]) and business (e.g., analysis of shopping trends [KLD15] ). D'Mello et al. [DOWH12] studied student engagement levels with eye tracking data, using gaze pattern to identify engagement levels of a student and to re-engage them by directing attention towards an animated tutoring agent. Gaze has also been used to improve user satisfaction with assistive AI agents by detecting affective states like stress [JA18], engagement [NI10], confusion [SC20], and cognitive abilities [BLG*20]. Rahman et al. [RAF*20] suggested various gaze visualizations for monitoring distracted students. Their results show that the accuracy of detecting distracted students was significantly lower for multiple students compared to when only one student was present in the class. This suggests that manual monitoring of student gaze data in a class is a challenging task for a teacher. Although eye tracking in VR has been used successfully to measure attention, most of the previous VR research did not examine the level of distraction during a class environment. Many educational VR studies fail to capture run-time processes that occur during a VR educational session as they mainly focus on evaluating post immersion learning with few isolated measures [BMTM20, APGVG10, Ayr06]. These studies supported the hypothesis of an existing relationship between EEG or gaze features and distraction. However, the use of gaze features and their relation to distraction are complex due to individual variability. Therefore, traditional statistical methods of analysis are not suitable to handle such complex. The use of deep learning techniques has been applied in recent years, e.g., [Hea21].

In our study, we present multiple information sources in a VR field trip by combining audio to explain objects, an avatar to point at objects, a slideshow to highlight key terms, and graphical animations to visualize device operations. We examined self-reported data on user's impression of the experience and applied deep learning to detect distraction level in this environment.

## 3 Educational VR Environment

Our VR environment was a Virtual Energy Center [BRC16] (see Fig. 1) used for virtual field trips. we used it as a VR class to explain the functionality of components necessary for the power production . An avatar explained the process and components using pre-recorded audio instructions, slides, and animations. All these components work synchronously to explain the subject matter. Additionally, relevant solar field components were highlighted to help students focus on the component being discussed.

The environment presented several informational cues (avatar, animations, audio, and slides) simultaneously that have been found to improve learning. Liang-Yi [Lia11] found that avatars boost students' learning. Our environment has a teacher avatar to point at objects and animations that help students look at the component being explained. Such animations have been used in the past to visualize the internal components of an object [RMFW20]. In our environment, animations were used to visualize internal operations of solar devices. Audio cues explained several aspects of the solar panel. Baceviciute et al. [BMTM20] found that audio is not superior to reading text in terms of knowledge retention. However, that study did not use the combination of the audio with other educational assets like slides, avatars, or animations to present the information. In our study, text slides were used to capture key terms of a particular component and mathematical concepts/equations. Our preliminary tests suggested that these slides were helpful for knowledge retention since mathematical concepts/equations are not easy to follow if just explained verbally.

## 4 Method Overview

As described by the following sections, we collected gaze data from our VR environment to test machine learning models.

### 4.1 Participants and Apparatus

We recruited 21 study participants (16 male and 5 female) from the university. Their ages ranged from 19 to 35 years (mean 25.9). 10 had prior experience with a VR device.

The experiment used a Vive Pro Eye connected to a desktop computer (Core i7 6700K, NVIDIA GeForce GTX 1080, 16 GB RAM, Microsoft Windows 10 Pro). We used Unity 3D v2018.2.21f1 software to implement the VR experience. Data was logged at 120hz, synchronized to eye tracker reports. Deep learning classification scripts were written in Python 3.8.8 with sklearn, TensorFlow and Keras libraries.

### 4.2 Experiment Design

Distractions can be internal or external. Internal distractions may be psychological or emotional. External distractions include auditory, visual, or physical noise. It is difficult to control internal distractions in an experimental setup. So, we focused mainly on external distractions. Social media notifications, mobile ringtones, and external conversations/sounds are three major student distractions [DKB*15,ASD17]. We simulated these distractions in our experiment. We also considered that tapping a VR user's body could be a relevant external distraction for VR. However, due to strict COVID protocols, contact was excluded from the experiment. Regarding internal distractions, we relied on participant self report (see Table 4 described later).

In the distractions phase, external distractions appeared randomly and are described below:

- **Social Media**: We requested the participants to turn on all social media (Facebook, Twitter, Instagram etc.) notifications as the sounds could create distraction [MPL*18]. We did not control this distraction. Participants got these notifications from their own social media accounts.
- **External Conversations/Sounds**: We produced external conversations in three ways. First, we played a conversation between two people from a YouTube video. Second, a dialogue unrelated to the educational content played randomly (picked from Table 1) with an intent to shift attention. Prior research found that such dialogues create distractions of up to 15 seconds [KM19]. Third, we played door closing and opening sounds similar to a real class door sound. For each session containing distractions, these distractions appear every 45 seconds.
- **Mobile Ringtone**: We played a pre-recorded mobile ringtone (through the headset speakers) and we also called the participant's mobile phone once.

**Table 1:** *Dialogues used to shift the attention to an unrelated task to create a distraction*

| Dialogues to shift attention | |
|---|---|
| Q1 | Think about your last conversation with your family. |
| Q2 | Think about a current work challenge you are facing. |
| Q3 | Think about a bird you saw recently. |
| Q4 | Think about anything that crosses your mind. |

The labeling of data points [HDH*20, MMP*20] with ground-truth is an important step for training a machine learning model. Some cybersickness-related studies [MMP*20, ILJ*20] had participants report a sickness level every 30, 45 or 60 seconds. However, these did not validate the levels, leading to human errors that could affect training data quality. For detecting distractions, asking for feedback every 30, 45 or 60 seconds would undesirably distract participants beyond the intended distractions. To avoid this, we divided our VR tutorial into several logical sessions (ranging from 100 seconds to 282 seconds) that could have different distraction levels. A participant may also have a different distraction level at the beginning and the end of a session. For this, each session was divided into two sections: the beginning section (first half) and the ending section (later half). At the end of each session, participants were asked to report, for both the sections, their distraction level (low, mid or high) and if they were drowsy.

The experiment had two phases with the same educational content. Each phase was divided in four sessions, each covering a small topic. In phase-I, there were no external distractions. In phase-II, we created the three external distractions. Participants, in the role of students, tried both phases in random order. Each session ended with 2 educational quiz questions and each phase (with same educational content) had a different set of quiz questions. Thus, the participant answered a total of 16 quiz questions (2 phases x 4 sessions x 2 questions per session). Because the participants were not experts on solar panels, the quiz questions were designed to be easy to answer by attentive students. The purpose of the quiz questions was to help gauge if the participant was distracted, under the assumption of some correlation between correct quiz answers and attention. This was considered in data point labeling.

Our experiment had three questionnaires: a pre-questionnaire, a post-session-questionnaire and a post-questionnaire. The pre-questionnaire consisted of distractability questions from a cognitive failure questionnaire (Table 2) to assess general distraction level in the last six months [WKS02], based on regular activities. Participants answered these questions as 5 point Likert items. The post-session questionnaire (Table 3) was filled out at the end of every session to assess the distraction level (for beginning and end sections of each session), engagement level, and drowsiness.

**Table 2:** *Pre-Questionnaire. Participants answered Q1-Q7 as 5-point Likert-like items. Q8 and Q9 were short text type.*

| Pre-Questionnaire Questions | |
|---|---|
| Q1 | Do you say something and realize afterwards that it might be taken as insulting? |
| Q2 | Do you fail to hear people speaking to you when you are doing something else? |
| Q3 | Do you lose your temper and regret it? |
| Q4 | Do you leave important letters/emails unanswered for days? |
| Q5 | Do you find yourself suddenly wondering whether you've used a word correctly? |
| Q6 | Do you daydream when you ought to be listening to something? |
| Q7 | Do you start doing one thing at home and get distracted into doing something else (unintentionally)? |
| Q8 | Do you check your mobile in a regular classroom? If yes, how often, provide an approximate time interval like every 5 or 10 minutes? |
| Q9 | What are the common distractions for you in a regular classroom? |

**Table 3:** *Post-Session Questionnaire. It was filled out at the end of every session in each phase*

| Post-Session Questionnaire | |
|---|---|
| How distracted were you while watching this lesson at the beginning of the session? | low/mid/high |
| How distracted were you while watching this lesson at the end of the session? | low/mid/high |
| Were you feeling any drowsiness during the task? | yes/no |

**Table 4:** *Post-Questionnaire. Participants answered Q1-Q11 as 7-point Likert-like items. Q12-Q15 were multiple choice questions.*

| Post-Questionnaire Questions | |
|---|---|
| Q1 | To what extent did the VR class hold your attention? |
| Q2 | How much effort did you put into attending the VR class and quiz? |
| Q3 | Did you feel you were trying your best? |
| Q4 | To what extent did you lose attention? |
| Q5 | Did you feel the urge to see what was happening around you? |
| Q6 | To what extent you enjoyed the VR class and quiz exam, rather than something you were just doing? |
| Q7 | To what extent did you find the VR class challenging? |
| Q8 | How much knowledge you could retain after VR class over solar panels? |
| Q9 | To what extent did you enjoy the graphics and the animation? |
| Q10 | How much would you say you enjoyed the VR class? |
| Q11 | To what extent did you feel drowsiness? |
| Q12 | Which one helped you to understand the lessons?<br>a) audio b) slides c) avatar d) animations |
| Q13 | Which one helped you to recall information to answer quizzes?<br>a) audio b) slides c) avatar d) animations |
| Q14 | Which component(s) distracted you except our simulated distractions?<br>a) audio b) slides c) avatar d) animations |
| Q15 | Did you feel any other distraction during VR class except our created distraction?<br>a) Mind Wandering b) Internal Stress c) Others |

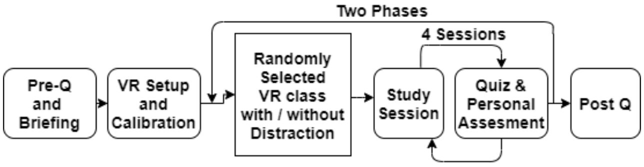


**Figure 2:** *Experiment Workflow*

Upon completion of all the sessions, participants filled out a post-questionnaire (Table 4), modified from [JCC*08], to gauge their overall experience. The total experiment duration was 45 to 60 minutes, but the VR portion including quizes lasted 29 to 45 minutes.

### 4.3 Data Collection Procedure

Due to COVID-19 risks, participants wore lower face masks in combination with disposable VR masks. Headsets were disinfected per participant. Participants were briefed about the study process and they provided signed consent. Subsequently, the participant was seated at the station, 2 meters away from the moderator. Participants filled out the pre-questionnaire. They then put on the VR headset and the integrated eye tracker was calibrated by software. Participants went through the two phases, each consisting of 4 sessions of the VR tutorial, in random order. They answered quiz questions and post-session questions (Table 3) after each session in each phase (session duration from 100 seconds to 282 seconds). After the end of the two phases, they filled out the post-questionnaire (see Table 4) about their experience. Our experimental workflow is summarized in Fig 2. We also asked our participants if they have any feedback about our VR tutorial and which components of the presentation distracted them or helped them for learning.

Raw gaze data collected throughout the sessions included timestamps, eye diameter, eye openness, eye wideness, gaze position, and gaze direction. The gaze sampling rate was 120Hz. Each frame included a flag used to discard readings reported as invalid by the tracker. For example, closing the eyes results in invalid gaze direction. Invalid data points were discarded for training the machine learning model. Eye diameter and eye openness were used to estimate drowsiness. We assumed that if a participant closed their eyes for more than two seconds continuously, they were drowsy. Additionally, we recorded a distance value, calculated as the distance between the Vive Eye's reported gaze origin and the highlighted object's position. This was intended to indicate how far from the highlighted object or avatar the participant was looking (see limitation in 6). This would give an indication of how attentive they were to relevant environment content.

### 4.4 Ground-Truth Construction and Validation

We considered three distraction levels for classification: low, mid and high. The participant's feedback at the end of each session was used in combination with quiz answers for labeling the data points associated with each section (beginning or ending) of a session. Our data labeling algorithm is described in Figure 3. If they answered both quiz questions correctly and rated their distraction level as low, associated data points were labeled as low distraction. If the quiz answers were not both correct and they rated distraction as high, associated points were labeled as high. If they answered both quiz questions correctly and rated their distraction as mid or high, drowsiness was considered. Reported drowsiness resulted in a "high" label and, otherwise, the label was "mid". If the quiz in-

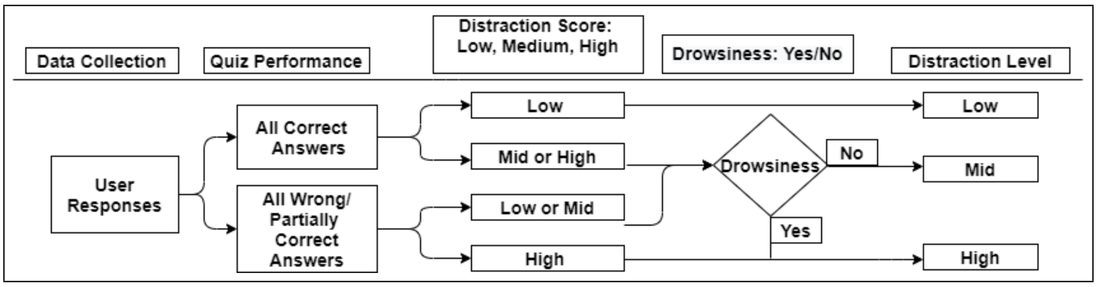


**Figure 3:** *Data Labeling Algorithm*

cluded one or two wrong answers, and reported distraction was low or mid, the label was again assigned as mid or high depending on reported drowsiness. Based on this method, the data distribution for both phases is shown in Figures 4 and 5. These figures show that we were successfully able to create distractions, since there were notably more distracted points in phase-II.

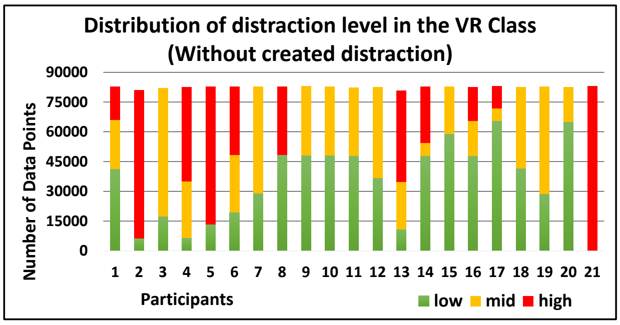


**Figure 4:** *Data distribution for Phase I (no external distractions).*

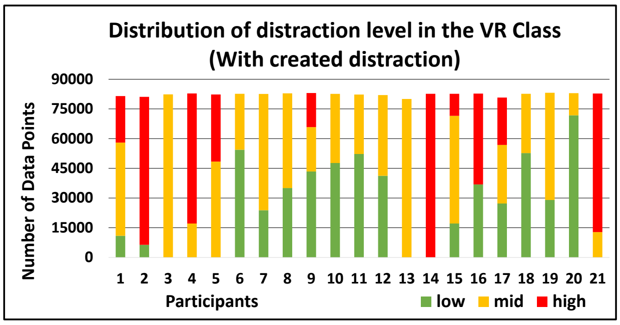


**Figure 5:** *Data distribution for Phase II (with external distractions). We counted mid and high level data points for each participant and noticed that 12 participants (out of 21) reported significantly higher level of distraction in this phase (indicated by yellow and red color in the Figure).*

### 4.5 Data Pre-Processing

The earlier-described eye tracker data was used for machine learning classifiers (e.g., CNN, LSTM). We split the dataset into training (70%) and test (30%) sets. Training sets are used to train classifiers and test sets are used to test classifier accuracy.

Before training, we pre-processed the data to potentially improve classifier accuracy. The data was first cleaned by replacing all invalid (“NaN”) values with zeros. For distraction classes (low, mid, and high labels), we found that the number of data points associated with each class was vastly different. The data was biased more towards low distraction. This skewed data would bias a classifier toward the low class. To avoid the bias and provide the same number of points per label, we up-sampled the data [DZW*14, PS17] for mid and high distraction classes by randomly creating duplicate copies of the data points within those classes. After this, we had 2831274 data points in the training set with 943758 data points for each class. Our test set had 1038331 data points. If we instead down-sampled our data to creating an equal count per class, some useful classification data could be lost.

We normalized data with min-max normalization and standardization. Min-max normalizes the data range to [0, 1] as follows:

$$Data_n = \frac{Data_i - Data_{min}}{Data_{max} - Data_{min}}$$

and data standardization is computed as:

$$Data_n = \frac{Data_i - Data_{avg}}{standard\ deviation}$$

We tried each technique separately for the entire dataset of all participants. We found that classifiers had a better accuracy with standardization. So, we chose standardization for our analysis.

### 4.6 Feature selection

We used the chi-squared test [TKA19] to identify the best features from our dataset. This gave the 9 most important features as: timestamp, left eye diameter, right eye diameter, distance value (as in 4.3), left eye openness, right eye openness, left eye wideness (another type of openness measure), right eye wideness, and drowsiness. A correlation matrix for these features is shown in Figure 6). We found that eye diameter, eye openness, and eye "wide" features are highly correlated with each other. We used the Extra Tree (ET) algorithm for feature extraction [KS20]. It gave a low score for drowsiness, and only three participants had detected drowsiness (for a short time). So, we did not use this feature.

### 4.7 Distraction Classification Models

We considered three deep learning models for our system: CNN, LSTM and CNN-LSTM. The CNN-LSTM model is our proposed model to combine the best features of the other two models.

**CNN**: We used the CNN model [ZLC*17] because it can learn

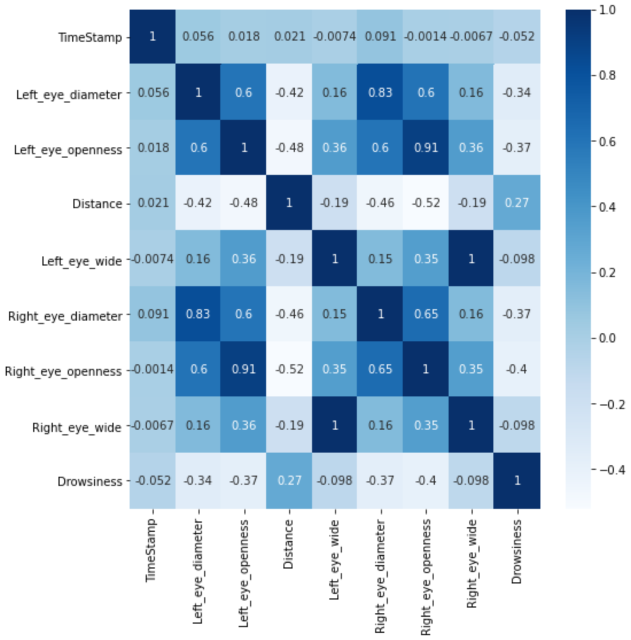


**Figure 6:** *Correlation matrix with heatmap indicates which features are most related to others*

**Table 5:** *Proposed CNN-LSTM architecture to classify the distraction level of students*

| Layer | Type | Output shape | ≠ param | Drop out | Activation |
|---|---|---|---|---|---|
| 1 | Conv1D | (8, 128) | 512 | - | ReLU |
| 2 | Batch Normalization | (8, 128) | 512 | - | - |
| 3 | MaxPool | (4, 128) | 0 | - | - |
| 4 | Conv1D | (4, 128) | 49280 | - | ReLU |
| 5 | Batch Normalization | (4, 128) | 512 | - | - |
| 6 | MaxPool | (2, 128) | 0 | - | - |
| 7 | LSTM | (128) | 131584 | | ReLU |
| 8 | Dropout | 128 | 0 | 0.2 | - |
| 9 | Flatten | ( 128) | 0 | - | - |
| 10 | Dense | 64 | 8256 | - | ReLU |
| 11 | Dense | 32 | 2080 | - | ReLU |
| 12 | Dense | 3 | 99 | - | Softmax |

to extract features from a sequence of observations and can classify raw time series data. The convolution kernel size [AM20] was 3, the batch size was 512, and the number of filter maps for the CNN was 128 (see Table 5 except the LSTM layer-7).

**LSTM**: We used LSTM because it would capture both temporal and spatial features of the gaze data. We set the batch size to 512 with hyper-parameter tuning. The model iterated over 200 epochs during training. After the first LSTM layer, we used a dropout layer of 50% to deal with overfitting. We used ReLU as the activation function for the first LSTM layer and the third dense layer. The last dense layer had three outputs for the three classes of distracted students whereas the activation function was softmax.

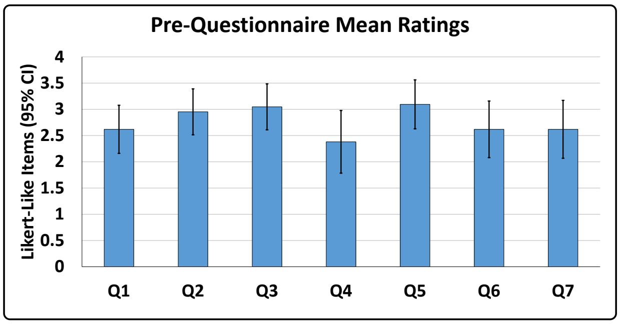


**Figure 7:** *Mean ratings for pre-questionnaire items.*

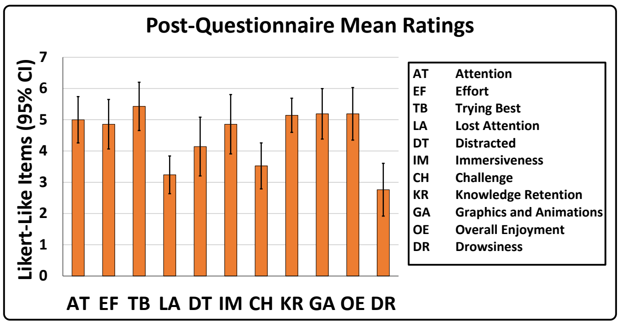


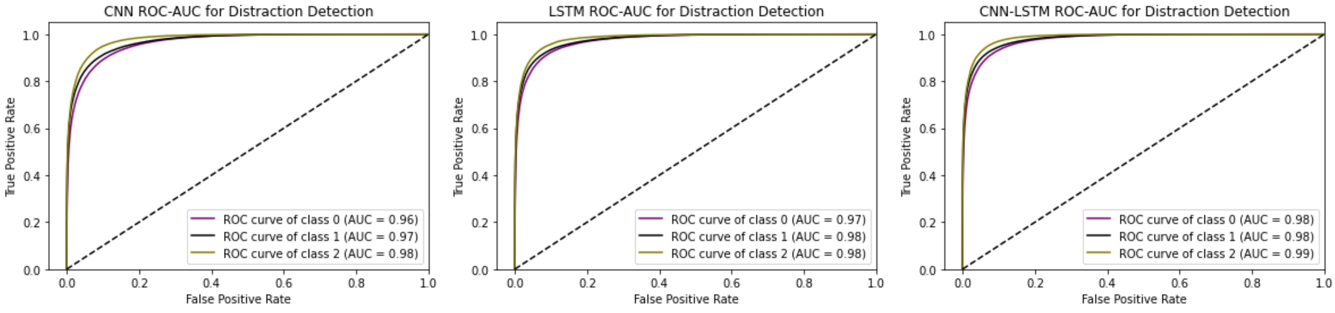
**Figure 8:** *Mean ratings for the post-questionnaire questions*

**CNN-LSTM**: We propose an improved model by merging layers from CNN and LSTM [SVSS15]. As the CNN layers are used for feature extraction from gaze data, the LSTM layer is used for temporal feature learning. The proposed model comprises of two Conv1D layers, one LSTM layer, and two fully connected dense layers (Table 5). The number of filters was 128 for the first two Conv1D layers, with the kernel size of 3. We used max pooling as the pooling operation with pool size 2. After the max pool operation, the output shape was reduced to (2, 128) and then the next LSTM layer is used for feature learning. We used the Adam optimizer [KB14] with a learning rate of $1 \times 10^{-3}$ and categorical cross-entropy as the loss function.

## 5 Results

Mean ratings for pre-questionnaire (Table 2 ) are plotted in Figure 7. We noticed that the majority of participants report distractibility in social situations. Similarly, mean ratings for the post-questionnaire (Table 4) are summarized in Figure 8. Most participants report trying their best to be attentive in VR but they got somewhat distracted. Moreover, most of them enjoyed the experience and were happy with the graphics/animations.

The accuracy and loss for the three models are summarized in Table 6. The CNN model had a lower accuracy and higher loss than the other models. The LSTM model had a significant improvement over the CNN model in terms of accuracy and loss. The CNN-LSTM model had the highest accuracy of 89.8% with a loss of 26.27%, an improvement over both the CNN and LSTM models.



(a) CNN (b) LSTM (c) CNN-LSTM

**Figure 9:** *The ROC-AUC curves for the three classification models. The class numbers 0, 1 and 2 corresponds to the three distraction classes, low, mid, and high, respectively.*

**Table 6:** *Average accuracy and loss of CNN, LSTM and CNN-LSTM models on Test Data*

| Name | Accuracy % | Loss % |
|---|---|---|
| CNN | 86.90 | 32.49 |
| LSTM | 88.40 | 29.58 |
| CNN-LSTM | 89.81 | 26.37 |

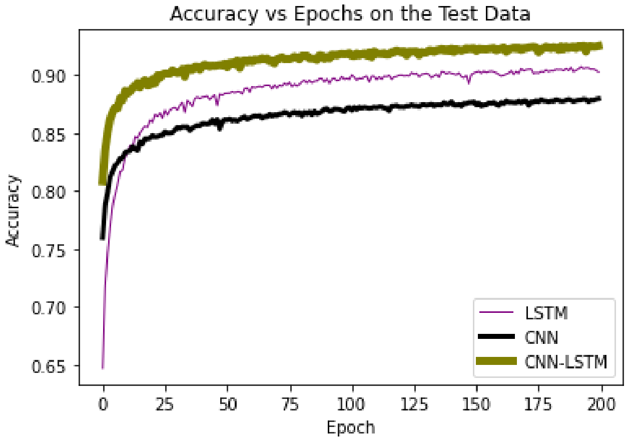


**Figure 10:** *Accuracy vs Epoch on the test data for classification*

The learning history on the test samples shows that CNN-LSTM converges to higher accuracy and lower loss faster than the other models (Figure 10 and 11).

The ROC-AUC curves for the three models are shown in Figure 9. The CNN model had an AUC of 98% for the high distraction class, which signifies that, 98% of the time, the model was able to distinguish between the high and other two classes (low and mid). The ROC-AUC curve for the LSTM model shows small improvement over the CNN model in the AUC score for the low and mid distraction classes. The CNN-LSTM model had the best performance for the three classes. This result suggests that the proposed CNN-LSTM model was able to distinguish between all three classes effectively.

The precision, recall and F1-scores for the three models are reported in Table 7. With an F1-score of 90%, the CNN-LSTM model performed best of the three models.

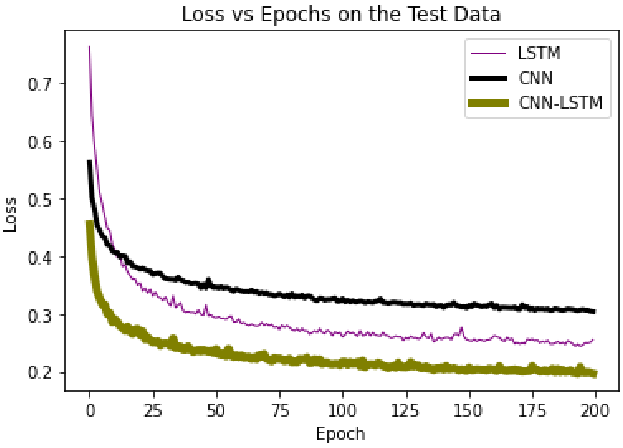


**Figure 11:** *Loss vs Epoch on the test data for classification*

**Table 7:** *Precision, recall and F1-score of the CNN, LSTM and CNN-LSTM models for the classification of distraction label*

| Name | Class | precision % | recall % | F1-score % |
|---|---|---|---|---|
| CNN | low | 0.88 | 0.85 | 0.86 |
| | mid | 0.87 | 0.88 | 0.87 |
| | high | 0.85 | 0.89 | 0.87 |
| LSTM | low | 0.91 | 0.85 | 0.88 |
| | mid | 0.88 | 0.90 | 0.89 |
| | high | 0.85 | 0.91 | 0.88 |
| CNN-LSTM | low | 0.90 | 0.89 | 0.90 |
| | mid | 0.91 | 0.89 | 0.90 |
| | high | 0.88 | 0.91 | 0.90 |

Testing was also conducted on the generalizability of our model to new variations of the educational environment. For this, we trained the model on data from three sessions and then tested classifier accuracy on data from the separate fourth session. Because each session had a different duration, the percentage of data points used for the test set was different for each case (Session 1: 26%,

**Table 8:** *Precision, recall and F1-score of the CNN-LSTM model for the classification of distraction label using 3 sessions for training and the remaining session for testing. The session used for testing is shown in column 1.*

| Session | Class | precision % | recall % | F1-score % |
|---|---|---|---|---|
| 1 | low | 0.66 | 0.62 | 0.64 |
| | mid | 0.51 | 0.64 | 0.57 |
| | high | 0.66 | 0.54 | 0.59 |
| 2 | low | 0.58 | 0.54 | 0.56 |
| | mid | 0.58 | 0.73 | 0.65 |
| | high | 0.58 | 0.40 | 0.47 |
| 3 | low | 0.62 | 0.74 | 0.67 |
| | mid | 0.58 | 0.52 | 0.55 |
| | high | 0.64 | 0.50 | 0.56 |
| 4 | low | 0.48 | 0.52 | 0.50 |
| | mid | 0.63 | 0.53 | 0.57 |
| | high | 0.60 | 0.66 | 0.63 |

Session 2: 15%, Session 3: 16%, and Session 4: 41%). The results are shown in Table 8. It is not surprising that the accuracy was lower (ranging from 48% to 66%) when the test data was completely new to the model.

We asked participants for comments or suggestions about the VR tutorial, which component(s) distracted them, and which component(s) helped them learn. Out of 21 participants, 18 indicated that audio helped them learn, 16 indicated slides as helpful, 15 indicated animations as helpful, and only 7 indicated the avatar as helpful. Surprisingly, 5 participants mentioned that the avatar distracted them, even though most participants mentioned that all these components work in sync and helped them to learn.

## 6 Discussion

Our results show that the CNN-LSTM model provides the best accuracy (Figure 10) and lower loss (Figure 11). We also measured the AUC and ROC values of the three classifiers to evaluate how good they were in distinguishing between the three distraction classes (Figure 9). The results suggested that the proposed CNN-LSTM model was able to distinguish between the three distraction classes more effectively than the other two models. Our work is a step towards an automatic real-time distraction level detection system for educational VR. We believe that such an automatic system could help manage a large guided class (30-50 students). For inattentive students, the system could trigger some action (such as pointing towards the object of interest [YKB19]) to bring their attention back without any manual intervention from the teacher.

Our experiment had some limitations. For detecting distraction level, ground-truth construction in an educational setup is challenging. Usually, educational sessions are long (more than 5 minutes). Frequently asking participants for their distraction level is not desirable due to its additional distracting effect. So, we divided our VR tutorial into several smaller sessions and asked the participant, at the end of each session, to rate their distraction level at the beginning and at the end of the session. This provided coarse granularity: in a 2-minute session, this gives more than 7000 data points per label. This could have affected our results. An alternative method for data labeling is to use known timing of controlled distraction events that last for a short duration (5-20 seconds for example). This would provide finer granularity for labeling and could potentially improve the accuracy of our system. Another limitation is the size of our dataset and type of participants. Due to COVID-19 protocols, we could not invite many participants or types of participants (we had 21 participants). Our test for generalizability of the model (Table 8) showed that our current model had a lower accuracy when tested on a new data set from a different session. We found that the computed distance feature (see 4.3), which was intended to be the distance between the looked-at point and the target/highlighted object, was miscalculated throughout our studies and was similar to a local gaze displacement magnitude based on Vive Eye's reported gaze origin. Nonetheless, it provided some value (see 4.6). We expect that the corrected distance or relative angle to target objects would likely improve results. Additionally, we could consider features characterizing fixations and saccades from eye tracking data [GR16]. Further research is needed to test this.

Student privacy is an important concern when sharing eye-gaze data of students with the teacher. In our study, eye-tracking data was collected from participants who gave permission to use their data within a standard informed consent model. The recorded data was anonymized. However, given that demographic information may be discerned from gaze data [LP14], great caution must be taken when handling it, especially if it has been gathered from minors (school students). If such a VR-based system is used for a real classroom, one must ensure that the students understand the meaning of eye tracking (perhaps by having them review example visualizations) and get permission from the students (and their parents, for minors) to track or record their eye gaze. Special care has to be taken for any longer-term storage to provide security, address legal requirements, and avoid any misuse of gaze data.

## 7 Conclusions and Future Work

We proposed a deep learning system to automatically detect the distraction level of students in a VR classroom. We tested three classification models (CNN, LSTM and CNN-LSTM) and found that the CNN-LSTM model had a better accuracy (89.81%) in classifying three distraction levels (low, mid and high). Here, we considered only eye-tracker data for detecting the distraction level. However, distraction level cannot be measured merely from eye gaze, as there are other factors involved (like physical and mental well being) that could affect distraction level. A student could be listening attentively even when not looking at certain objects, or vice versa. In the future, we would like to consider more metrics and sensor data (EEG, heart rate, skin conductance, etc.) for detecting distraction. Additionally, it is important to develop real-time detection methods and train/test models to work in a wider range of VR environments.

## 8 Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1815976 and by the Louisiana Board of Regents under contract No. LEQSF(2019-20)-ENH-DE-22. We thank Dr. Miao Jin for useful comments on this paper.

## References

[AM20] Agrawal A., Mittal N.: Using cnn for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy. *The Visual Computer 36*, 2 (2020), 405–412. 6

[APGVG10] Antonenko P., Paas F., Grabner R., van Gog T.: Using electroencephalography to measure cognitive load. *Educational Psychology Review 22*, 4 (2010), 425–438. 1, 2

[ASD17] Agrawal P., Sahana H., De' R.: Digital distraction. In *Proceedings of the 10th International Conference on Theory and Practice of Electronic Governance* (2017), pp. 191–194. 3

[Ayr06] Ayres P.: Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and instruction 16*, 5 (2006), 389–400. 1, 2

[BC17] Buttussi F., Chittaro L.: Effects of different types of virtual reality display on presence and learning in a safety training scenario. *IEEE transactions on visualization and computer graphics 24*, 2 (2017), 1063–1076. 2

[BLG*20] Barral O., Lallé S., Guz G., Iranpour A., Conati C.: Eye-tracking to predict user cognitive abilities and performance for user-adaptive narrative visualizations. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (2020), pp. 163–173. 1, 2

[BMC15] Bertram J., Moskaliuk J., Cress U.: Virtual training: Making reality work? *Computers in Human Behavior 43* (2015), 284–292. 2

[BMTM20] Baceviciute S., Mottelson A., Terkildsen T., Makransky G.: Investigating representation of text and audio in educational vr using learning outcomes and eeg. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), pp. 1–13. 1, 2, 3

[BRC16] Borst C. W., Ritter K. A., Chambers T. L.: Virtual energy center for teaching alternative energy technologies. In *2016 IEEE Virtual Reality (VR)* (2016), IEEE, pp. 157–158. 2

[BRKB21] Broussard D. M., Rahman Y., Kulshreshth A. K., Borst C. W.: An interface for enhanced teacher awareness of student actions and attention in a vr classroom. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (2021), pp. 284–290. doi:10.1109/VRW52623.2021.00058. 1

[But14] Butcher K. R.: The multimedia principle. *The Cambridge handbook of multimedia learning 2* (2014), 174–205. 2

[DBL*20] Dumoulin S., Bouchard S., Loranger C., Quintana P., Gougeon V., Lavoie K. L.: Are cognitive load and focus of attention differentially involved in pain management: an experimental study using a cold pressor test and virtual reality. *Journal of Pain Research 13* (2020), 2213. 2

[DKB*15] David P., Kim J.-H., Brickman J. S., Ran W., Curtis C. M.: Mobile phone distraction while studying. *New media & society 17*, 10 (2015), 1661–1679. 3

[DOWH12] D'Mello S., Olney A., Williams C., Hays P.: Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of human-computer studies 70*, 5 (2012), 377–398. 1, 2

[Duc02] Duchowski A. T.: A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers 34*, 4 (2002), 455–470. 2

[DZW*14] Dubey R., Zhou J., Wang Y., Thompson P. M., Ye J., Initiative A. D. N., et al.: Analysis of sampling techniques for imbalanced data: An n= 648 adni study. *NeuroImage 87* (2014), 220–241. 5

[GBMT13] Gardony A. L., Brunyé T. T., Mahoney C. R., Taylor H. A.: How navigational aids impair spatial memory: Evidence for divided attention. *Spatial Cognition & Computation 13*, 4 (2013), 319–350. 2

[GC04] Gallagher A. G., Cates C. U.: Virtual reality training for the operating room and cardiac catheterisation laboratory. *The Lancet 364*, 9444 (2004), 1538–1540. 1

[GR16] George A., Routray A.: A score level fusion method for eye movement biometrics. *Pattern Recognition Letters 82* (2016), 207–215. 8

[HDH*20] Herbig N., Düwel T., Helali M., Eckhart L., Schuck P., Choudhury S., Krüger A.: Investigating multi-modal measures for cognitive load detection in e-learning. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (2020), pp. 88–97. 3

[Hea21] Healy B. C.: Machine and deep learning in ms research are just powerful statistics–no. *Multiple Sclerosis Journal 27*, 5 (2021), 663–664. 2

[ILJ*20] Islam R., Lee Y., Jaloli M., Muhammad I., Zhu D., Rad P., Huang Y., Quarles J.: Automatic detection and prediction of cybersickness severity using deep neural networks from user's physiological signals. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (2020), IEEE, pp. 400–411. 3

[JA18] Jyotsna C., Amudha J.: Eye gaze as an indicator for stress level analysis in students. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (2018), IEEE, pp. 1588–1593. 1, 2

[JCC*08] Jennett C., Cox A. L., Cairns P., Dhoparee S., Epps A., Tijs T., Walton A.: Measuring and defining the experience of immersion in games. *International Journal of Human-Computer Studies 66,9* (2008), 641–661. 4

[KB14] Kingma D. P., Ba J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 6

[KLD15] Kim M., Lee M. K., Dabbish L.: Shop-i: Gaze based interaction in the physical world for in-store social shopping experience. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (2015), pp. 1253–1258. 2

[KM19] Kosmyna N., Maes P.: Attentivu: an eeg-based closed-loop biofeedback system for real-time monitoring and improvement of engagement for personalized learning. *Sensors 19*, 23 (2019), 5200. 3

[KS20] Kasongo S. M., Sun Y.: A deep learning method with wrapper based feature extraction for wireless intrusion detection system. *Computers & Security 92* (2020), 101752. 5

[Lia11] Liang-Yi Chung: Using avatars to enhance active learning: Integration of virtual reality tools into college english curriculum. In *The 16th North-East Asia Symposium on Nano, Information Technology and Reliability* (2011), pp. 29–33. doi:10.1109/NASNIT.2011.6111116. 2

[LP14] Liebling D. J., Preibusch S.: Privacy considerations for a pervasive eye tracking world. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct Publication - UbiComp '14 Adjunct* (2014), ACM Press. doi:10.1145/2638728.2641688. 8

[MEK*01] Mrochen M., Eldine M. S., Kaemmerer M., Seiler T., Hütz W.: Improvement in photorefractive corneal laser surgery results using an active eye-tracking system. *Journal of Cataract & Refractive Surgery 27*, 7 (2001), 1000–1006. 2

[MMP*20] Martin N., Mathieu N., Pallamin N., Ragot M., Diverrez J.-M.: Virtual reality sickness detection: an approach based on physiological signals and machine learning. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (2020), IEEE, pp. 387–399. 3

[MN11] Mikropoulos T. A., Natsis A.: Educational virtual environments: A ten-year review of empirical research (1999–2009). *Computers & Education 56*, 3 (2011), 769–780. 2

[MOM19] Meyer O. A., Omdahl M. K., Makransky G.: Investigating the effect of pre-training when learning through immersive virtual reality and video: A media and methods experiment. *Computers & Education 140* (oct 2019), 103603. doi:10.1016/j.compedu.2019.103603. 2

[MPL*18] MENDOZA J. S., PODY B. C., LEE S., KIM M., MCDONOUGH I. M.: The effect of cellphones on attention and learning: The influences of time, distraction, and nomophobia. *Computers in Human Behavior 86* (2018), 52–60. 2, 3

[MTM19] MAKRANSKY G., TERKILDSEN T. S., MAYER R. E.: Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learning and Instruction 60* (apr 2019), 225–236. doi:10.1016/j.learninstruc.2017.12.007. 2

[NI10] NAKANO Y. I., ISHII R.: Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proceedings of the 15th international conference on Intelligent user interfaces* (2010), pp. 139–148. 1, 2

[OD17] OBERHAUSER M., DREYER D.: A virtual reality flight simulator for human factors engineering. *Cognition, Technology & Work 19*, 2-3 (2017), 263–277. 1

[PS17] PATIL S. S., SONAVANE S. P.: Improved classification of large imbalanced data sets using rationalized technique: Updated class purity maximization over_sampling technique (ucpmot). *Journal of Big Data 4*, 1 (2017), 1–32. 5

[RAF*20] RAHMAN Y., ASISH S. M., FISHER N. P., BRUCE E. C., KULSHRESHTH A. K., BORST C. W.: Exploring eye gaze visualization techniques for identifying distracted students in educational vr. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (2020), IEEE, pp. 868–877. 1, 2

[RBD18] RUCINSKI C. L., BROWN J. L., DOWNER J. T.: Teacher–child relationships, classroom climate, and children's social-emotional and academic development. *Journal of Educational Psychology 110*, 7 (2018), 992. 2

[RMFW20] RADIANTI J., MAJCHRZAK T. A., FROMM J., WOHLGENANNT I.: A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & Education 147* (2020), 103778. 3

[RSG*15] RODRIGUE M., SON J., GIESBRECHT B., TURK M., HÖLLERER T.: Spatio-temporal detection of divided attention in reading applications using eeg and eye tracking. In *Proceedings of the 20th international conference on intelligent user interfaces* (2015), pp. 121–125. 2

[SC20] SIMS S. D., CONATI C.: A neural architecture for detecting user confusion in eye-tracking data. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (2020), pp. 15–23. 1, 2

[SM12] SZAFIR D., MUTLU B.: Pay attention! designing adaptive agents that monitor and improve user engagement. In *Proceedings of the SIGCHI conference on human factors in computing systems* (2012), pp. 11–20. 2

[SVSS15] SAINATH T. N., VINYALS O., SENIOR A., SAK H.: Convolutional, long short-term memory, fully connected deep neural networks. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (2015), IEEE, pp. 4580–4584. 6

[TKA19] THASEEN I. S., KUMAR C. A., AHMAD A.: Integrated intrusion detection model using chi-square feature selection and ensemble of classifiers. *Arabian Journal for Science and Engineering 44*, 4 (2019), 3357–3368. 5

[WKS02] WALLACE J. C., KASS S. J., STANNY C. J.: The cognitive failures questionnaire revisited: dimensions and correlates. *The Journal of general psychology 129*, 3 (2002), 238–256. 3

[WMT*19] WON M., MOCERINO M., TANG K.-S., TREAGUST D. F., TASKER R.: Interactive immersive virtual reality to enhance students' visualisation of complex molecules. In *Research and Practice in Chemistry Education*. Springer, 2019, pp. 51–64. 2

[YB21] YOSHIMURA A., BORST C. W.: A study of class meetings in vr: Student experiences of attending lectures and of giving a project presentation. *Frontiers in Virtual Reality 2* (2021), 34. URL: https://www.frontiersin.org/article/10.3389/frvir.2021.648619, doi:10.3389/frvir.2021.648619. 1

[YKB19] YOSHIMURA A., KHOKHAR A., BORST C. W.: Eye-gaze-triggered visual cues to restore attention in educational VR. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (mar 2019), IEEE. doi:10.1109/vr.2019.8798327. 8

[ZLC*17] ZHAO B., LU H., CHEN S., LIU J., WU D.: Convolutional neural networks for time series classification. *Journal of Systems Engineering and Electronics 28*, 1 (2017), 162–169. 5