Computers & Graphics 109 (2022) 75–87

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on ICAT-EGVE Highlights

Detecting distracted students in educational VR environments using machine learning on eye gaze data

Sarker Monojit Asish *, Arun K. Kulshreshth, Christoph W. Borst

University of Louisiana at Lafayette, LA, United States

ARTICLE INFO

Article history:
Received 18 February 2022
Received in revised form 25 August 2022
Accepted 21 October 2022
Available online 2 November 2022

Keywords:
Education
Virtual Reality
Eye Tracking
Machine Learning
Deep Learning
Distraction Detection

ABSTRACT

Virtual Reality (VR) has been found useful to improve engagement and retention level of students, for some topics, compared to traditional learning tools such as books, and videos. However, a student could still get distracted and disengaged due to a variety of factors including stress, mind-wandering, unwanted noise, and external alerts. Student eye gaze data could be useful for detecting these distracted students. Gaze data-based visualizations have been proposed in the past to help a teacher monitor distracted students. However, it is not practical for a teacher to monitor a large number of student indicators while teaching. To help filter students based on distraction level, we propose an automated system based on machine learning to classify students based on their distraction level. The key aspects are: (1) we created a labeled eye gaze dataset 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 several classifiers for this purpose. Each classifier classifies distraction, per educational activity section, into one of three levels (low, mid or high). Our results show that Random Forest (RF) classifier had the best accuracy (98.88%) compared to the other models we tested. Additionally, a personalized machine learning model using either RF, kNN, or Extreme Gradient Boosting (XGBoost) model was found to improve the classification accuracy significantly.

1. Introduction

Virtual reality (VR) has long been suggested as a way to enhance education [1]. Students can virtually take field trips to any place or learn about different machinery and how it works with reduced concern about safety and cost. VR can produce experiences that are vividly remembered, along with numerous other effects that seem to hinge on immersive or embodied experiences [2]. Furthermore, 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. However, there are certain challenges associated with VR-based education. 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 cannot see students directly. Additionally, students get distracted in VR due a variety of reasons such as noise in the real environment around the student (e.g., phone ringing, notification tones, etc.), distractions from other objects or features in the virtual room (e.g., an interesting object in the environment), distractions from other avatars, or checking external tools [3]. Thus, it is challenging for a teacher to help/guide students who are confused or distracted. These distractions direct the attention of students away from the educational content being presented and thus inhibit their abilities to learn in the classroom.

We previously explored gaze visualizations to help teachers monitor students' attention when guiding VR field trips [4]. However, continual visualization of gaze from many students is not practical, leading to a higher cognitive load, 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. [5] 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.

* Corresponding author.
E-mail addresses: asish.sust@gmail.com (S.M. Asish), arunkul@louisiana.edu (A.K. Kulshreshth), cwborst@gmail.com (C.W. Borst).

https://doi.org/10.1016/j.cag.2022.10.007

Gaze-data has been used in the past for detecting engagement levels [6,7], stress [8], confusion [9], and cognitive abilities [10] in non-VR educational applications. A few other previous studies [11–13] 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 (e.g., mean, standard deviation, minimum, maximum values, etc.) of data analysis are not suitable to handle eye gaze data to classify distraction level. The reason being the fact that the number of input features and possible associations among them increase [14].

We propose a system based on machine learning that identifies the distraction level of a student based on eye gaze data in VR [15]. 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 various 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 real-time distraction detection system. We had two experiments. In the first experiment, we compared three deep learning classification models (CNN, LSTM, and CNN-LSTM). In the second experiment, our follow-up experiment, we compared the best model(CNN-LSTM) from experiment 1 with two other machine learning models (Random Forest and Extreme Gradient Boosting). We also explored if a personalized machine learning model, where we train and test using the data from the same person, would improve the classification accuracy.

2. Related work

Educational VR has been mostly used for procedural motor skill training in fields such as aviation and medicine [16,17]. In the last decade, immersive VR has been studied in other educational contexts, such as safety training [18], and training public security personnel [19]. VR has provided new opportunities for visualizing and interacting with abstract learning content (e.g., molecular structures [20]) as well as simulation applications that would be hazardous to practice in real life (e.g., hazardous situation) [21].

Eye gaze has been studied for decades for a wide range of applications [22] such as medical (e.g., eye surgery [23]) and business (e.g., analysis of shopping trends [24]). D'Mello et al. [6] 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 data has been used to improve user satisfaction with assistive AI agents by detecting affective states like stress [8], engagement [7], confusion [9], and cognitive abilities [10]. Recently, Lengyel et al. [25] utilized gaze data for predicting future attention targets in a simulated VR meeting. Gaze data has also been used for task recognition [26] and evaluation of road safety education program in VR [27]. In a computer interface, researchers [28] have detected mind wandering by analyzing eye gaze features while reading text. Another study build a real-time mind-wandering detection and intervention system while reading comprehension [29].

There are a variety of activities that could distract students in an educational environment (VR or otherwise). Psychological research found that many students use their cellphones to browse the internet or shop online while attending a class [30]. 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 [31–33]. 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 [34].

Rahman et al. [4] 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 [11–13]. 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 (e.g., mean, standard deviation, minimum, maximum values, etc.) are not suitable to handle eye-gaze data. The use of deep learning techniques has been applied in recent years, e.g., [35].

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 [36] in terms of complex graphic representations that include dual-channel processing in working memory. Sarune et al. [11] 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. Han et al. [37] proposed some intervention strategies to improve students' attention and their findings suggest that instructions from real world teachers can be transferred to virtual classroom. In some cases, VR leads to a higher sense of presence and keeps users engaged with educational content [38–40]. However, text-based presentation could lead to higher cognitive load and less learning in VR [38].

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 machine learning to detect distraction level in this environment.

3. Educational VR environment

Our VR environment was a Virtual Energy Center [41] (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 [42] 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

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

past to visualize the internal components of an object [43]. 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. [11] 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. Makransky et al. [38] found that multimedia slides increases users' interest but creates less learning. However, in this study, we assume that combining all educational assets may improve learning.

4. Methodology

4.1. Method overview

As described by the following sections, we collected gaze data from our VR environment to test machine learning models. We had two experiments. In experiment 1, we trained and tested three deep learning machine models: CNN, LSTM, and CNN-LSTM. The CNN-LSTM is our proposed models which merges layers from CNN and LSTM. Both supervised and unsupervised learning approaches were tested. From the results of the experiment 1 (see the result section), we found then the CNN-LSTM model has the best accuracy and additionally we found that the models are not transferable to other educational session/scene since the features used were specific to a session/scene. Thus, we decided to explore this further and conducted a follow up experiment (experiment 2). In experiment 2, we added more gaze/head based features which were relative to the object of interest and the teacher avatar. This would allow the models to be more general and could potentially be used for classification in other educational environments without retraining. We compared the best model (CNN-LSTM) from experiment 1 with two other machine learning models (Random Forest and XGBoost). Additionally, we also used the data from this experiment 2 to explore if a personalized machine learning model would improve the classification accuracy.

4.2. Participants and apparatus

In experiment 1, we recruited 21 study participants (16 male and 5 female) from the university. Their ages ranged from 19 to 35 years (mean 25.9) and 10 of them had prior experience with a VR device. In experiment-2, we collected data from 20 participants (13 male and 7 female). Their age ranged from 18 to 38 years (mean age of 22.1) and 12 of them had prior VR experience. The experiment duration was 45 to 60 min, but the VR portion including follow-up questions (see Tables 3 and 4) lasted 29 to 45 min.

Both experiments used a Vive Pro Eye connected to a desktop computer (Core i7 6700 K, 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 120 Hz, synchronized to eye tracker reports. Deep learning classification scripts were written in Python 3.8.8 with sklearn, TensorFlow and Keras libraries.

4.3. Design of our experiments

In this section, we describe the design choices that we made for both our experiments (experiment 1 and experiment 2).

4.3.1. Distractions

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 [44,45]. 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).

Both experiments had two phases with the same educational content: one with no external distractions and one with external distractions. In the distractions phase, external distractions appeared randomly (counterbalanced using the Latin-square) 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 [30]. 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 s [46]. Third, we played door closing and opening sounds similar to a real class door sound. For each session containing distractions, these distractions appear every

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.

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?

45 s in experiment 1 and every 30 to 40 s randomly in experiment 2.

- **Mobile Ringtone**: We played a pre-recorded mobile ringtone (through the headset speakers) and we also called the participant's mobile phone once.

4.3.2. Data labeling

The labeling of data points [47,48] with ground-truth is an important step for training a machine learning model. Some cybersickness-related studies [48,49] had participants report a sickness level every 30, 45 or 60 s. 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 s would undesirably distract participants beyond the intended distractions. To avoid this, we divided our VR tutorial into several logical sessions (ranging from 100 s to 282 s) 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. Same approach was used for both experiments.

4.3.3. Experiment phases

Both experiments 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.

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? (a) audio (b) slides (c) avatar (d) animations
Q13	Which one helped you to recall information to answer quizzes? (a) audio (b) slides (c) avatar (d) animations
Q14	Which component(s) distracted you except our simulated distractions? (a) audio (b) slides (c) avatar (d) animations
Q15	Did you feel any other distraction during VR class except our created distraction? (a) Mind Wandering (b) Internal Stress (c) Others

4.3.4. Experiment questionnaires

Each experiment had three questionnaires: a pre-questionnaire, a post-session-questionnaire and a post-questionnaire. The pre-questionnaire consisted of distractibility questions from a cognitive failure questionnaire (Table 2) to assess general distraction level in the last six months [50], 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. Upon completion of all the sessions, participants filled out a post-questionnaire (Table 4), modified from [51], to gauge their overall experience. The total experiment duration was 45 to 60 min, but the VR portion including quizzes lasted 29 to 45 min.

Fig. 2. User Setup for the study.

4.4. Data collection procedure

Our study was approved by the University IRB committee (Approval# FA20-51 CACS). 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 (see Fig. 2), 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 s to 282 s). After the end of the two phases, they filled out the post-questionnaire (see Table 4) about their experience. Our experimental workflow, for both experiments, is summarized in Fig. 3. 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 with learning.

In experiment 1, 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 120 Hz. Each data frame received from the eye-tracking API included a flag which indicates if the data is valid. For example, closing the eyes results in a invalid gaze direction value. All these invalid data points were discarded from the training data. Eye diameter and eye openness were used to estimate drowsiness based on past research on detecting drowsiness for drivers [52–54]. 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.

In experiment 2, the data collected was same as experiment 1, with the exception of distance value, with some extra features. These extra features were eye-gaze angles, head gaze angle, Vive's reported gaze origin value (one 3d vector for each eye), head orientation, and a modified distance value, calculated as the distance between the head gaze and eye gaze point. There were two eye gaze angles, one relative to the highlighted object and one relative to the teacher avatar. The eye-gaze angle relative to the highlighted object is calculated as the angle between gaze direction and the vector joining the participant location and the highlighted object. The eye-gaze angle relative to the teacher avatar is calculated as the angle between gaze direction and the vector joining the participant location and the teacher avatar. The head-gaze angle was relative to the highlighted object and is calculated as the angle between head-gaze direction and the vector joining the participant location and the highlighted object.

4.5. 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 Fig. 4. 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 included 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 of experiment 1 is shown in Figs. 5 and 6. These figures show that we were successfully able to create distractions, since there were notably more distracted points in phase-II.

4.6. 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. We discarded all the data values which were reported as invalid by the eye-tracking API. For the extracted features (distance, and eye/head gaze angles), some calculated values were invalid (NaN: Not a number). We cleaned this data by replacing all these invalid (NaN: Not a Number) 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 in case of experiment 1 and more towards mid distraction in experiment 2. This skewed data would bias a classifier towards the low class or mid class. To avoid the bias and provide the same number of points per label, we up-sampled the data [55,56] for other distraction classes by randomly creating duplicate copies of the data points within those classes. After this, for experiment 1, we had 2,831,274 data points in the training set with 943,758 data points for each class and 1,038,331 data points in the test set. In case of experiment 2, we had 2,223,910 data points for training and 953,105 data points for testing. To avoid bias and overfitting with the training data, we applied a combination of under-sampling and Synthetic Minority Oversampling (SMOTE) [57] so that classifiers can learn on the dataset perturbed by "SMOTING" the minority class and under-sampling the majority class.

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

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

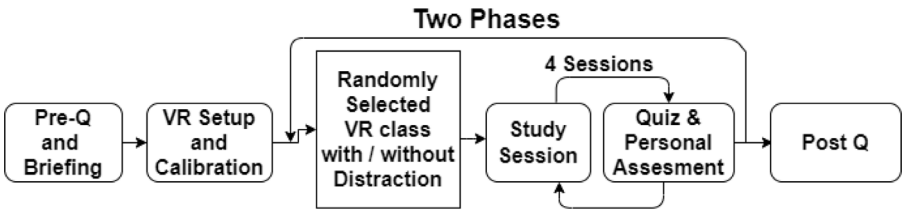

Fig. 3. Experiment Workflow for both experiments.

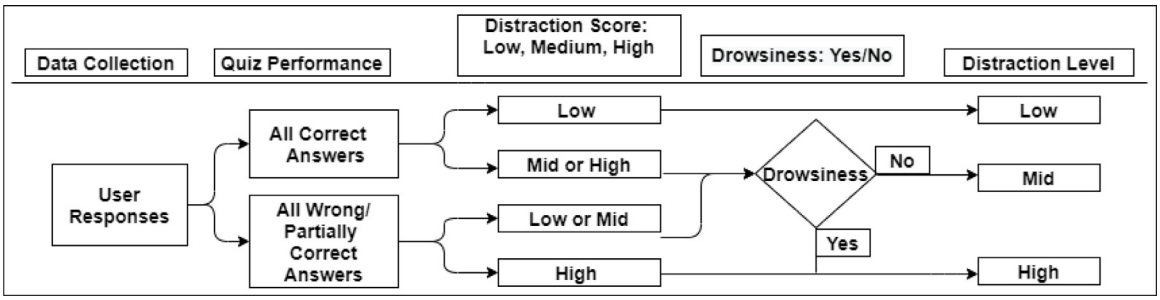

Fig. 4. Data Labeling Algorithm for Supervised Learning Models.

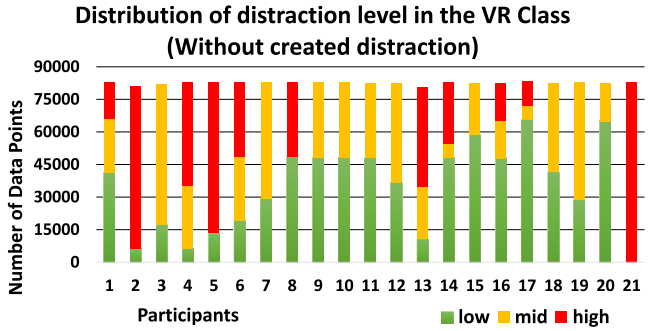

Fig. 5. Data distribution for Phase I (no external distractions) of experiment 1.

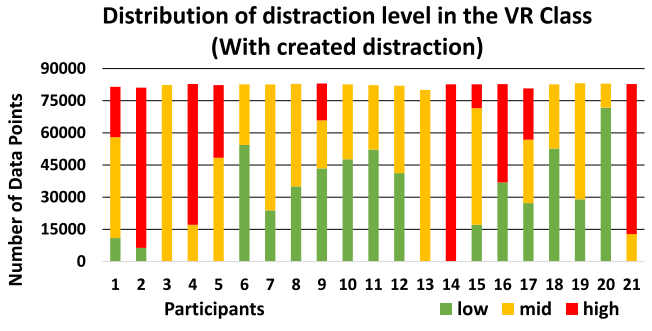

Fig. 6. Data distribution for Phase II (with external distractions) of experiment 1. We counted mid and high level data points for each participant and noticed that 12 participants (out of 21) reported significantly higher level of distraction in this phase (indicated by yellow and red color in the Figure).

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.

4.7. Feature selection

We used the chi-squared test [58] to identify the best features from our dataset obtained from experiment 1. This gave the 9 most important features as: timestamp, left eye diameter, right eye diameter, distance value (as in 4.4), 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 Fig. 7. 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 [59]. 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.

4.8. Distraction classification models

In case of experiment 1, 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. These models are described below:

CNN: We used the CNN model [60] because it can learn to extract features from a sequence of observations and can classify raw time series data. The convolution kernel size [61] was 3, the batch size was 512, and the number of filter maps for the CNN was 128 (see Table 6 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 (see Table 5), 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.

CNN-LSTM: We propose an improved model by merging layers from CNN and LSTM [62]. 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 6). The number of filters was 128 for the first two Conv1D layers, with a kernel size of 3. We used max pooling as the pooling

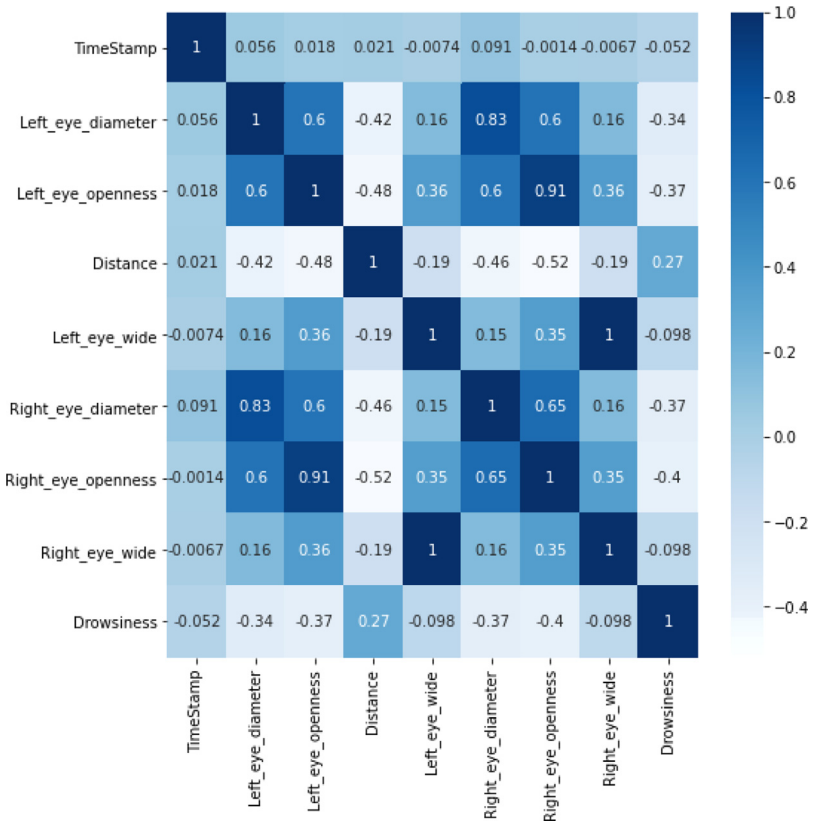

Fig. 7. Correlation matrix with heatmap indicates which features are most related to others in case of experiment 1.

Table 5
LSTM architecture for classification of student distraction.

Layer	Type	Output shape	≠ param	Drop out	Activation
1	LSTM	(128)	66560	–	ReLU
2	Dropout	(128)	0	0.50	–
3	Dense	64	8256	–	ReLU
4	Dense	32	2080	–	ReLU
5	Dense	3	99	–	Softmax

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 [63] with a learning rate of $1e^{-3}$ and categorical cross-entropy as the loss function.

In case of experiment 2, our follow-up experiment, we compared the CNN-LSTM, our best model based on results of experiment 1 (see the Results section), with two other popular models: Random Forest and XGBoost. For both experiment training and testing sets are person-independent.

We also wanted to test if a machine learning model would perform better (in terms of accuracy) if we use the training and test data from the same person. This approach is known as personalized machine learning. We compared three models for this approach: Random Forest, XGBoost, kNN and Linear Discriminant Analysis (LDA). These models are described below:

Random Forest: Random Forest is an ensemble learning method which construct multiple decision trees through different data subsets, and voting on the results of multiple decision trees to get the prediction as output of the model. We used "RandomizedSearchCV" library from sklearn to optimize our hyperparameters for Random Forest and we found the optimized parameter where estimator = 200, max depth = 460, and max features = 'sqrt'. We plugged these into the model and reported the results.

XGBoost: Extreme Gradient Boosting(XGBoost) is a refined and customized version of a gradient boosting decision tree system which is proposed by Chen and Guestrin [64] in 2016. It is an ensemble method since it combines multiple decision trees where each tree built based on the result of the previous developed tree. In contrast to Random Forest, in which trees are grown to their maximum extent, XGBoosting makes use of trees with fewer splits. It runs faster than other model used in our work as it drives fast learning through parallel and distributed computing along with efficient memory usage. We implemented it with default hyper-parameters using Python library.

kNN: k-nearest-neighbors (kNN) classifier implements learning based on the k nearest neighbors. The choice of the value of k is dependent on data. At low k values, there is overfitting (training error is low and test error is high) of data variance. We evaluated from 1 to 10 to choose k value and we found that it works best for k = 6 and the parameter metric is Minkowski by default.

Linear Discriminant Analysis (LDA): LDA is a linear classification model which uses statistical properties of the input data. For each input variable, it calculates the mean and the variance of the variable for each class. We used the default parameters (solver = 'svd' and shrinkage = None) and the estimated statistical properties from data were plugged into the LDA model to make prediction.

5. Results

5.1. Experiment 1: Comparing deep learning models

The accuracy and loss for the three models are summarized in Table 7. The CNN model had a lower accuracy and higher loss than the other models. The LSTM model had a significant improvement

Table 6
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

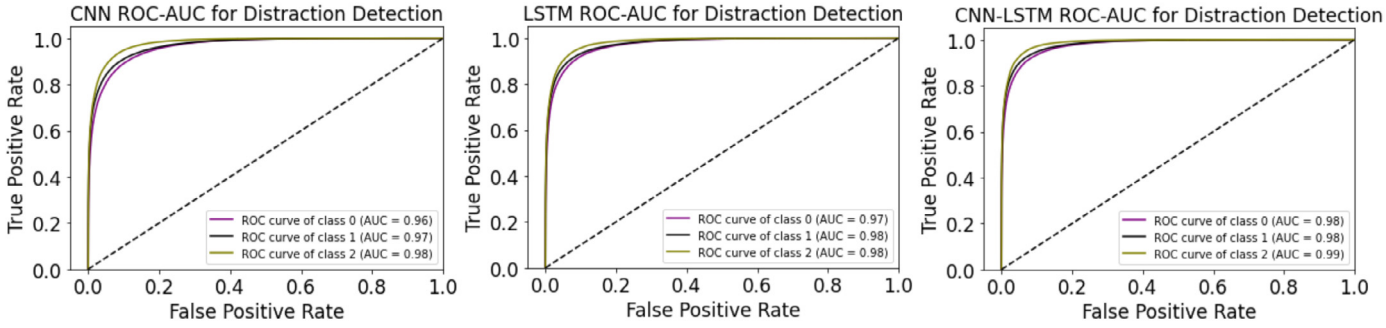

Fig. 8. The ROC-AUC curves for the three deep learning based classification models in experiment 1. The class numbers 0, 1 and 2 corresponds to the three distraction classes, low, mid, and high, respectively.

Table 7
Average accuracy and loss of CNN, LSTM and CNN-LSTM models on Test Data from experiment 1.

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

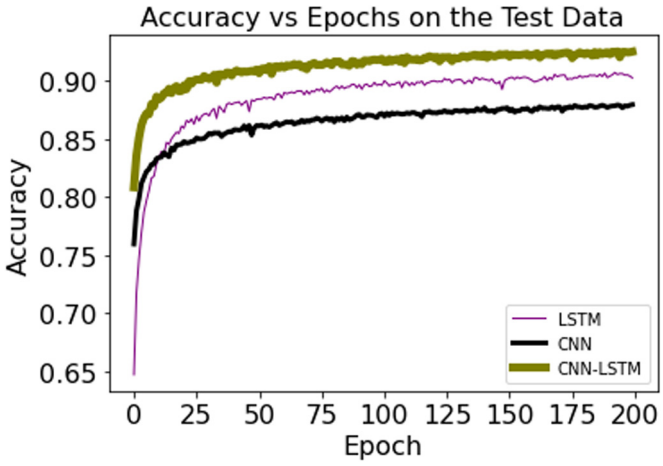

Fig. 9. Accuracy vs Epoch on the test data in experiment 1.

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. The learning history on the test samples shows that CNN-LSTM converges to higher accuracy and lower loss faster than the other models (Figs. 9 and 10).

The ROC-AUC curves for the three models are shown in Fig. 8. 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.

As the accuracy is not the only evaluation metrics for classification, precision and recall are measured to see individual class scores. Precision is defined as the ratio of correctly predicted positive observations to the total predicted positive observations. Recall is defined as the ratio of correctly predicted positive observations to the all observations in actual positive class. F1 Score is the weighted average of Precision and Recall. It is generally described as the harmonic mean of the two. Therefore, this score takes both false positives and false negatives into account. The precision, recall and F1-scores for the three models are reported in Table 8. With an F1-score of 90%, the CNN-LSTM model performed best of the three models.

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%, Session 2: 15%, Session 3: 16%, and Session 4: 41%). The results are shown in Table 9. It is not surprising that the accuracy was lower (ranging from 48% to 66%) when the test data was completely new to the model.

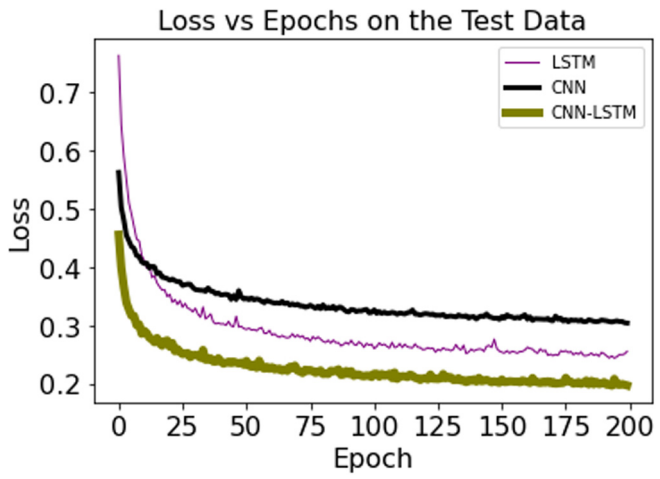

Fig. 10. Loss vs Epoch on the test data for classification in experiment 1.

Table 8
Precision, recall and F1-score of the CNN, LSTM and CNN-LSTM models in experiment 1 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

Table 9
Precision, recall and F1-score of the CNN-LSTM model for the classification of distraction label in experiment 1 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

Mean ratings for pre-questionnaire (Table 2) are plotted in Fig. 11. We noticed that the majority of participants report distractibility in social situations. Similarly, mean ratings for the post-questionnaire (Table 4) are summarized in Fig. 12. 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.

We asked participants for comments or suggestions about the VR tutorial, which component(s) distracted them, and which component(s) helped them learn. In experiment 1, 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

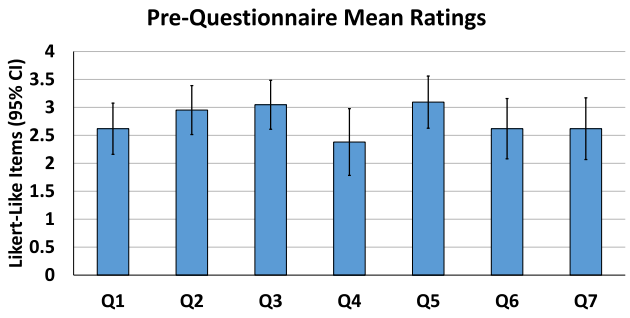

Fig. 11. Mean ratings for pre-questionnaire items in experiment 1.

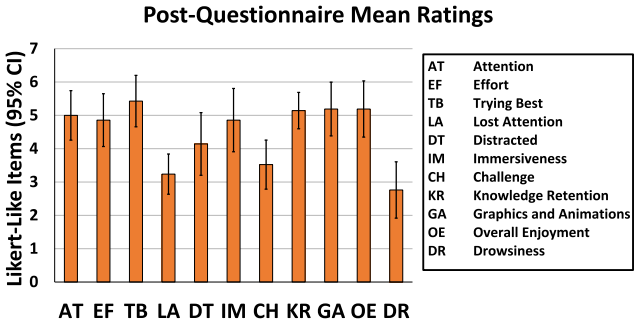

Fig. 12. Mean ratings for the post-questionnaire questions in experiment 1.

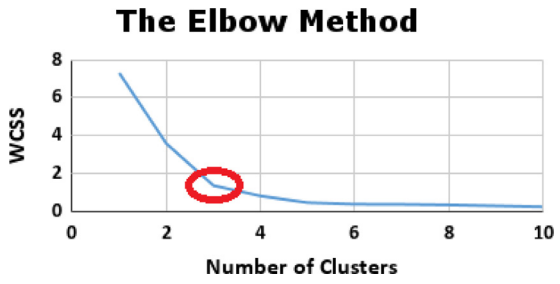

Fig. 13. The relationship between the number of clusters and Within Cluster Sum of Squares (WCSS)(elbow method).

participants mentioned that all these components work in sync and helped them to learn.

5.1.1. Supervised vs unsupervised learning

So far we designed a data labeling algorithm and analyzed results based on supervised classification system. However, as the data labeling and verification system is complex and time consuming, it is still unclear if the supervised data labeling is the best approach for labeling the gaze data. Instead of asking user's feedback for data labeling, we could use an unsupervised method, such as K-means clustering, to label our gaze data. Thus, we decided to compare these two approaches (supervised vs unsupervised). The elbow method on our data shows a kink at $k = 2$ and $k = 3$ (see Fig. 13), which indicates that we should consider two or three clusters. We chose three cluster to compare our results with our supervised models with three classes corresponding to low, medium and high distraction levels [65].

We split the dataset into training (70%) and test (30%) sets. The training set was used to train the classifiers and the test set was used to test a classifier's accuracy. Using the same training data, we trained the three deep learning models (CNN, LSTM and CNN-LSTM) for both supervised and unsupervised data labeling methods. The average accuracy for the three models for both

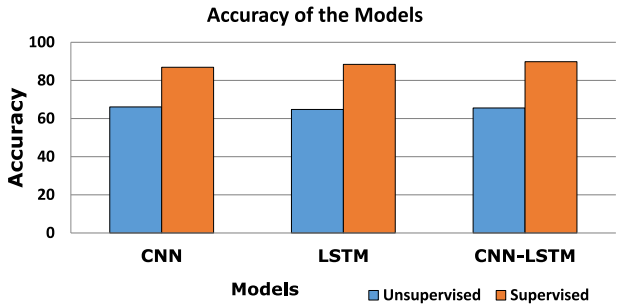

Fig. 14. Average accuracy of the models with unsupervised learning.

Table 10
Precision, recall and F1-score of the deep learning models with unsupervised learning.

Name	Class	Precision %	Recall %	F1-score %
CNN	Low	0.92	1.00	0.96
	Mid	0.38	0.51	0.43
	High	0.63	0.45	0.52
LSTM	Low	0.91	1.00	0.95
	Mid	0.38	0.49	0.43
	High	0.59	0.43	0.50
CNN-LSTM	Low	0.91	1.00	0.95
	Mid	0.36	0.50	0.42
	High	0.64	0.44	0.52

unsupervised and supervised learning is shown in Fig. 14. The overall accuracy for all models are very close to each other for both unsupervised and supervised learning models. However, the accuracy is significantly lower for the unsupervised models. The precision, recall and F1-scores are shown in Table 10 for the unsupervised models. We found that accuracy was better for the low distraction class with the unsupervised learning. However, it had significantly lower accuracy for medium and high distraction classes compared to the supervised learning models (see Tables 8 and 10 for comparison).

5.2. Experiment 2: Follow-up experiment

We had data from 20 participants for this follow up experiment. The comparison of average accuracy with the new features (as discussed in 4.4) is shown in Table 11 for the three models that we tested (CNN-LSTM, RF, and XGBoost). The column 2 shows the results when we used the old features from experiment 1 with the change that the distance feature was now changed (calculated as the distance between the head gaze and eye gaze point). The column 3 shows the results when the gaze angles corresponding to eye gaze and head gaze (two angles for each, one relative to the highlighted object and the other relative to the teacher avatar) were added to the feature set. The column 4 shows the results when both gaze angles and gaze origin (reported by the HTC Vive's API) were used. We found that adding gaze angles and gaze origin did improve the accuracy significantly. The Random Forest (RF) model had the best accuracy of 98.88% with the new features. We also collected head-tracking data during this experiment but we did not find any significant contribution by the quaternion values (x, y, z, and w) of head rotation. Therefore, we ignored these values from our selected features.

To test for generalizability, we trained the models with data from three session (out of 4 sessions) and used the remaining session data for testing. The results are shown in Fig. 15. According to the results, we see that the accuracy did not improved significantly compared to experiment 1 (see Table 9. The Random Forest model still performed the best with accuracy ranging from 60.27% to 69.11%. The precision, recall and F1-scores for the RF model are shown in Table 12.

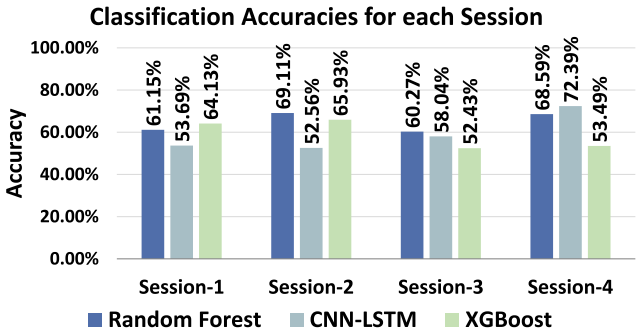

Fig. 15. Classification Accuracy when 3 sessions are used for training and the remaining 4th session is used for testing. The session used for testing is shown on the *x*-axis.

5.2.1. Personalized machine learning approach

Based on our previous analysis, the research question in this stage was how well a classifier would perform if the training and test data was from the same person. We tested four classifiers: Random Forest (RF), XGBoost, Linear Discriminant Analysis (LDA), and k-nearest-neighbors (kNN). We added the kNN and LDA models in this analysis since these models work better with smaller data sets such as in this case where we have data from a single person. We did not test CNN-LSTM since it is a deep learning model and its training requires a lot of data which is not feasible from data obtained from a single person.

Similar to our prior test, we tested two scenarios: one with training data from all the sessions and the second one where three sessions were used for training and the remaining fourth session was used for testing. In the first case, we trained the model with 60% data and tested the models with the remaining 40% data for each participant. In the second case, we wanted to test the classifier on new data for testing its generalizability. Thus, we stacked the data from three session for training and used the remaining fourth session data for testing. For both scenarios, the results were very similar. We have shown the result of our first test in Table 13. Overall, the best performance was achieved using RF, XGBoost and kNN models for all the participants. The LDA model performed worse for all participants. All of these ML models take only few seconds to train and test the data from the same person.

6. Discussion

According to our first experiment, the results show that the CNN-LSTM model provides the best accuracy (Fig. 9) and lower loss (Fig. 10). 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 (Fig. 8). 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. Additionally, we tested and compared the performance of these models with unsupervised (K-means clustering) learning to label data. Our results show that the unsupervised learning is not a good choice for classification of distraction based on the gaze data. This shows that the relationship between gaze features and distraction is more complex and we cannot use any statistical unsupervised methods, such as K-means, to label this data. We believe that this happens because gaze data has a lot of variability across individuals and thus it

Table 11
Average accuracy of CNN-LSTM, Random Forest (RF) and Extreme Gradient Boosting(XGBoost) models using the data from all the sessions.

Model name	Accuracy with the old features (though distance was calculated differently)	Accuracy with gaze angles added to the feature set	Accuracy with gaze angles and Vive's reported gaze origin added to the feature set
CNN-LSTM	77.26%	90.85%	97.50%
RF	96.80%	97.62%	98.88%
XGBoost	90.85%	97.43%	98.10%

Table 12
Precision, recall and F1-score of the Random Forest (RF) model for the classification of distraction label using 3 sessions for training and using the remaining session for testing. The session used for testing is shown in column 1.

Session	Class	Precision %	Recall %	F1-score %
1	Low	0.61	0.50	0.55
	Mid	0.71	0.68	0.69
	High	0.67	0.74	0.70
2	Low	0.47	0.75	0.58
	Mid	0.84	0.74	0.79
	High	0.83	0.53	0.69
3	Low	0.73	0.51	0.58
	Mid	0.60	0.64	0.62
	High	0.47	0.79	0.64
4	Low	0.46	0.81	0.59
	Mid	0.78	0.81	0.79
	High	0.91	0.45	0.50

Table 13
Classification Accuracy for each Participant with data from all the sessions with 60% used for training and 40% for testing.

Participants	Accuracy (RF) %	Accuracy (LDA) %	Accuracy (kNN) %	Accuracy (XGBoost)%
1	0.99	0.75	1.0	0.99
2	0.99	0.60	1.0	0.99
3	0.99	0.70	1.0	0.99
4	0.99	0.72	1.0	0.99
5	1.0	0.90	1.0	1.0
6	1.0	0.97	1.0	1.0
7	1.0	0.72	1.0	1.0
8	1.0	0.70	1.0	1.0
9	1.0	0.72	1.0	1.0
10	1.0	0.87	1.0	1.0
11	1.0	0.85	1.0	1.0
12	1.0	0.85	1.0	1.0
13	0.99	0.70	0.99	0.99
14	0.99	0.72	1.0	0.99
15	0.99	0.70	1.0	0.99
16	0.99	0.75	1.0	0.99
17	0.99	0.65	0.99	1.0
18	1.0	0.85	1.0	1.0
19	1.0	0.90	1.0	1.0
20	1.0	0.87	1.0	1.0

requires a supervised approach for data labeling. Furthermore, we found that the deep learning models did not perform well when tested on new data from a different session not used for training (see Table 9). Thus, these models are not generalizable based on the feature set used in this experiment. We found that the computed distance feature (see 4.4), 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 4.7). Thus, we decided to conduct a follow up experiment where we changed the distance calculations and added some more features (see Section 4.4). In this experiment, we compared our best model from the first experiment (CNN-LSTM) with two other machine learning models (Random forest and XGBoost).

Our followup experiment, experiment 2, revealed that with the new distance value, the gaze angles (3 angles in total, two for eye gaze and one for head gaze) and gaze origin features significantly improved the accuracy of all the three models tested (CNN-LSTM, RF, and XGBoost). The RF model performed the best with an accuracy of 98.88% and an accuracy of 97.62% when gaze origin features were not used. Thus, we can conclude that using features which are relative to important objects (e.g., target/highlighted object, the teacher avatar, etc.) in the scene (such as gaze angles) and features independent of the VR environment (such as gaze origin) makes the model more accurate (see Table 11). In terms of generalizability, the random forest model performed better (see Fig. 15) for all sessions than the other models tested. The CNN-LSTM model performed better than XGBoost model when session 3 and session 4 was used for testing. The XGBoost model performed better than CNN-LSTM when session 1 and session 2 was used for testing. Overall RF model performed consistently better for all four test cases. Furthermore, our results revealed that personalized machine learning models could significantly improve the classification accuracy using RF, XGBoost and kNN models (see Table 13). Therefore, in a real VR-based classroom, one should consider this approach since the same set of students will attend classes for a given semester. Potentially, the model could be trained at the beginning of the semester and could then be used for students for the rest of the semester.

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 [66]) 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 min). 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 s 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 in experiment 1 and 20 in experiment 2). Our choice of features could also have an affect on the generalizability of the classification models tested. For the future, we could consider features characterizing fixations and saccades from eye tracking data [67]. 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 [68], 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 machine learning system to automatically detect the distraction level of students in a VR classroom. We tested several classification models and found that the Random Forest model had a better accuracy (98.88%) in classifying the data into three distraction classes (low, mid and high). We found that unsupervised learning (using k-means) does not work for classifying distractions based on gaze data since it leads to a low accuracy. Furthermore, we found that personalized machine learning approach with Random Forest, kNN or XGBoost model could significantly improve the classification accuracy.

In this experiment, 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. It would also be interesting to see the performance of these techniques for detecting distracted students for a networked VR class with multiple students.

CRediT authorship contribution statement

Sarker Monojit Asish: Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing – original draft, Writing – review & editing. **Arun K. Kulshreshth:** Conception and design of study, Writing – original draft, Writing – review & editing. **Christoph W. Borst:** Conception and design of study, Writing – original draft.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Christoph W. Borst reports financial support was provided by National Science Foundation. Arun K. Kulshreshth reports financial support was provided by Louisiana Board of Regents.

Data availability

Data will be made available on request.

Acknowledgments

This material is based upon work supported by the National Science Foundation, USA under Grant No. 1815976 and by the Louisiana Board of Regents, USA under contract No. LEQSF(2022–25)-RD-A-24. We would like to thank Yee Tan for his contribution to the educational environment, and Tomas Parker and Jesse Marin for helping us with pilot studies. Special thanks to Ekram Hossain for helping us to implement the data labeling algorithm. All authors approved version of the manuscript to be published.

References

[1] Youngblut C. Educational uses of virtual reality technology. Tech. rep., Institute for defense analysis, Alexendria VA; 1998.
[2] Blascovich J, Bailenson J. Infinite reality: Avatars, eternal life, new worlds, and the dawn of the virtual revolution. William Morrow & Co; 2011.
[3] Yoshimura A, Borst CW. A study of class meetings in VR: Student experiences of attending lectures and of giving a project presentation. Front Virtual Real 2021;2:34. http://dx.doi.org/10.3389/frvir.2021.648619, URL https://www.frontiersin.org/article/10.3389/frvir.2021.648619.
[4] Rahman Y, Asish SM, Fisher NP, Bruce EC, Kulshreshth AK, Borst CW. Exploring eye gaze visualization techniques for identifying distracted students in educational VR. In: 2020 IEEE conference on virtual reality and 3D user interfaces. IEEE; 2020, p. 868–77.
[5] Broussard DM, Rahman Y, Kulshreshth AK, Borst CW. 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. 2021, p. 284–90. http://dx.doi.org/10.1109/VRW52623.2021.00058.
[6] D'Mello S, Olney A, Williams C, Hays P. Gaze tutor: A gaze-reactive intelligent tutoring system. Int J Hum-Comput Stud 2012;70(5):377–98.
[7] Nakano YI, 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, p. 139–48.
[8] 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. IEEE; 2018, p. 1588–93.
[9] Sims SD, Conati C. A neural architecture for detecting user confusion in eye-tracking data. In: Proceedings of the 2020 international conference on multimodal interaction. 2020, p. 15–23.
[10] 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, p. 163–73.
[11] 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, p. 1–13.
[12] Antonenko P, Paas F, Grabner R, Van Gog T. Using electroencephalography to measure cognitive load. Educ Psychol Rev 2010;22(4):425–38.
[13] Ayres P. Using subjective measures to detect variations of intrinsic cognitive load within problems. Learn Instr 2006;16(5):389–400.
[14] Ij H. Statistics versus machine learning. Nature Methods 2018;15(4):233.
[15] Asish SM, Hossain E, Kulshreshth AK, Borst CW. Deep learning on eye gaze data to classify student distraction level in an educational VR environment. In: Orlosky J, Reiners D, Weyers B, editors. ICAT-EGVE 2021 - International conference on artificial reality and telexistence and eurographics symposium on virtual environments. The Eurographics Association; 2021, http://dx.doi.org/10.2312/egve.20211326.
[16] Gallagher AG, Cates CU. Virtual reality training for the operating room and cardiac catheterisation laboratory. Lancet 2004;364(9444):1538–40.
[17] Oberhauser M, Dreyer D. A virtual reality flight simulator for human factors engineering. Cogn, Technol Work 2017;19(2–3):263–77.
[18] Buttussi F, Chittaro L. Effects of different types of virtual reality display on presence and learning in a safety training scenario. IEEE Trans Vis Comput Graphics 2017;24(2):1063–76.
[19] Bertram J, Moskaliuk J, Cress U. Virtual training: Making reality work? Comput Hum Behav 2015;43:284–92.
[20] Won M, Mocerino M, Tang KS, Treagust DF, Tasker R. Interactive immersive virtual reality to enhance students' visualisation of complex molecules. In: Research and practice in chemistry education. Springer; 2019, p. 51–64.
[21] Mikropoulos TA, Natsis A. Educational virtual environments: A ten-year review of empirical research (1999–2009). Comput Educ 2011;56(3):769–80.
[22] Duchowski AT. A breadth-first survey of eye-tracking applications. Behav Res Methods Instrum Comput 2002;34(4):455–70.

[23] Mrochen M, Eldine MS, Kaemmerer M, Seiler T, Hütz W. Improvement in photorefractive corneal laser surgery results using an active eye-tracking system. J Cataract Refract Surg 2001;27(7):1000–6.
[24] Kim M, Lee MK, 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, p. 1253–8.
[25] Lengyel G, Carlberg K, Samad M, Jonker TR. Predicting visual attention using the hidden structure in eye-gaze dynamics. In: CHI2021 Eye movements as an interface to cognitive state (EMICS) workshop proceedings. ACM. 2021.
[26] Hu Z, Bulling A, Li S, Wang G. Ehtask: Recognizing user tasks from eye and head movements in immersive virtual reality. IEEE Trans Vis Comput Graphics 2021.
[27] Skjermo J, Roche-Cerasi I, Moe D, Opland R. Evaluation of road safety education program with virtual reality eye tracking. SN Comput Sci 2022;3(2):1–11.
[28] Bixler R, D'Mello S. Automatic gaze-based user-independent detection of mind wandering during computerized reading. User Model User-Adapt Interact 2016;26(1):33–68.
[29] Mills C, Gregg J, Bixler R, D'Mello SK. Eye-mind reader: An intelligent reading interface that promotes long-term comprehension by detecting and responding to mind wandering. Hum–Comput Interact 2021;36(4):306–32.
[30] Mendoza JS, Pody BC, Lee S, Kim M, McDonough IM. The effect of cellphones on attention and learning: The influences of time, distraction, and nomophobia. Comput Hum Behav 2018;86:52–60.
[31] Dumoulin S, Bouchard S, Loranger C, Quintana P, Gougeon V, Lavoie KL. Are cognitive load and focus of attention differentially involved in pain management: An experimental study using a cold pressor test and virtual reality. J. Pain Res 2020;13:2213.
[32] 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, p. 11–20.
[33] 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, p. 121–5.
[34] Gardony AL, Brunyé TT, Mahoney CR, Taylor HA. How navigational aids impair spatial memory: Evidence for divided attention. Spatial Cogn Comput 2013;13(4):319–50.
[35] Healy BC. Machine and deep learning in MS research are just powerful statistics–no. Multiple Scler J 2021;27(5):663–4.
[36] Butcher KR. The multimedia principle. In: The Cambridge Handbook of Multimedia Learning, vol. 2, Cambridge University Press New York, NY; 2014, p. 174–205.
[37] Han Y, Miao Y, Lu J, Guo M, Xiao Y. Exploring intervention strategies for distracted students in VR classrooms. In: CHI conference on human factors in computing systems extended abstracts. 2022, p. 1–7.
[38] Makransky G, Terkildsen TS, Mayer RE. Adding immersive virtual reality to a science lab simulation causes more presence but less learning. Learn Instr 2019;60:225–36. http://dx.doi.org/10.1016/j.learninstruc.2017.12.007.
[39] Meyer OA, Omdahl MK, Makransky G. Investigating the effect of pre-training when learning through immersive virtual reality and video: A media and methods experiment. Comput Educ 2019;140:103603. http://dx.doi.org/10.1016/j.compedu.2019.103603.
[40] Rucinski CL, Brown JL, Downer JT. Teacher–child relationships, classroom climate, and children's social-emotional and academic development. J Educ Psychol 2018;110(7):992.
[41] Borst CW, Ritter KA, Chambers TL. Virtual energy center for teaching alternative energy technologies. In: 2016 IEEE virtual reality. IEEE; 2016, p. 157–8.
[42] Chung L-Y. 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, p. 29–33. http://dx.doi.org/10.1109/NASNIT.2011.6111116.
[43] Radianti J, Majchrzak TA, Fromm J, Wohlgenannt I. A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. Comput Educ 2020;147:103778.
[44] David P, Kim JH, Brickman JS, Ran W, Curtis CM. Mobile phone distraction while studying. New Media Soc 2015;17(10):1661–79.
[45] Agrawal P, Sahana H, De' R. Digital distraction. In: Proceedings of the 10th international conference on theory and practice of electronic governance. 2017, p. 191–4.
[46] Kosmyna N, Maes P. Attentivu: An EEG-based closed-loop biofeedback system for real-time monitoring and improvement of engagement for personalized learning. Sensors 2019;19(23):5200.
[47] Herbig N, Düwel T, Helali M, Eckhart L, Schuck P, Choudhury S, et al. 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, p. 88–97.
[48] 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. IEEE; 2020, p. 387–99.
[49] Islam R, Lee Y, Jaloli M, Muhammad I, Zhu D, Rad P, et al. 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. IEEE; 2020, p. 400–11.
[50] Wallace JC, Kass SJ, Stanny CJ. The cognitive failures questionnaire revisited: Dimensions and correlates. J Gen Psychol 2002;129(3):238–56.
[51] Jennett C, Cox AL, Cairns P, Dhoparee S, Epps A, Tijs T, et al. Measuring and defining the experience of immersion in games. Int J Hum-Comput Stud 2008;66,9:641–61.
[52] Hussein MK, Salman TM, Miry AH, Subhi MA. Driver drowsiness detection techniques: A survey. In: 2021 1st Babylon international conference on information technology and science. IEEE; 2021, p. 45–51.
[53] Rahman A, Sirshar M, Khan A. Real time drowsiness detection using eye blink monitoring. In: 2015 National software engineering conference. IEEE; 2015, p. 1–7.
[54] Mandal B, Li L, Wang GS, Lin J. Towards detection of bus driver fatigue based on robust visual analysis of eye state. IEEE Trans Intell Transp Syst 2016;18(3):545–57.
[55] Dubey R, Zhou J, Wang Y, Thompson PM, Ye J, Initiative ADN, et al. Analysis of sampling techniques for imbalanced data: An n=648 ADNI study. NeuroImage 2014;87:220–41.
[56] Patil SS, Sonavane SP. Improved classification of large imbalanced data sets using rationalized technique: Updated class purity maximization over_Sampling technique (UCPMOT). J Big Data 2017;4(1):1–32.
[57] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artificial Intelligence Res 2002;16:321–57.
[58] Thaseen IS, Kumar CA, Ahmad A. Integrated intrusion detection model using chi-square feature selection and ensemble of classifiers. Arab J Sci Eng 2019;44(4):3357–68.
[59] Kasongo SM, Sun Y. A deep learning method with wrapper based feature extraction for wireless intrusion detection system. Comput Secur 2020;92:101752.
[60] Zhao B, Lu H, Chen S, Liu J, Wu D. Convolutional neural networks for time series classification. J Syst Eng Electron 2017;28(1):162–9.
[61] Agrawal A, Mittal N. Using CNN for facial expression recognition: A study of the effects of kernel size and number of filters on accuracy. Vis Comput 2020;36(2):405–12.
[62] Sainath TN, 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. IEEE; 2015, p. 4580–4.
[63] Kingma DP, Ba J. Adam: A method for stochastic optimization. 2014, arXiv preprint arXiv:1412.6980.
[64] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016, p. 785–94.
[65] Asish SM, Kulshreshth AK, Borst CW. Supervised vs unsupervised learning on gaze data to classify student distraction level in an educational VR environment. In: Symposium on spatial user interaction. 2021, p. 1–2.
[66] Yoshimura A, Khokhar A, Borst CW. Eye-gaze-triggered visual cues to restore attention in educational VR. In: 2019 IEEE conference on virtual reality and 3D user interfaces. IEEE; 2019, http://dx.doi.org/10.1109/vr.2019.8798327.
[67] George A, Routray A. A score level fusion method for eye movement biometrics. Pattern Recognit Lett 2016;82:207–15.
[68] Liebling DJ, 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. ACM Press; 2014, http://dx.doi.org/10.1145/2638728.2641688.