International Journal of Artificial Intelligence in Education
https://doi.org/10.1007/s40593-022-00316-z

ARTICLE

Exploring Gender Differences in Computational Thinking Learning in a VR Classroom: Developing Machine Learning Models Using Eye-Tracking Data and Explaining the Models

Hong Gao[1] · **Lisa Hasenbein**[2] · **Efe Bozkir**[1] · **Richard Göllner**[2] · **Enkelejda Kasneci**[3]

Accepted: 10 October 2022

Abstract

Understanding existing gender differences in the development of computational thinking skills is increasingly important for gaining valuable insights into bridging the gender gap. However, there are few studies to date that have examined gender differences based on the learning process in a realistic classroom context. In this work, we aim to investigate gender classification using students' eye movements that reflect temporal human behavior during a computational thinking lesson in an immersive VR classroom. We trained several machine learning classifiers and showed that students' eye movements provide discriminative information for gender classification. In addition, we employed a Shapley additive explanation (SHAP) approach for feature selection and further model interpretation. The classification model trained with the selected (best) eye movement feature set using SHAP achieved improved performance with an average accuracy of over 70%. The SHAP values further explained the classification model by identifying important features and their impacts on the model output, namely gender. Our findings provide insights into the use of eye movements for in-depth investigations of gender differences in learning activities in VR classroom setups that are ecologically valid and may provide clues for providing personalized learning support and tutoring in such educational systems or optimizing system design.

Keywords Gender classification · Machine learning · Explainable AI · Computational thinking · Eye movements · Virtual reality

✉ Hong Gao
hong.gao@informatik.uni-tuebingen.de

Extended author information available on the last page of the article

Published online: 04 November 2022

Introduction

Computational thinking (CT), which refers to the thought processes involved in expressing solutions as computational steps or algorithms that can be carried out by a computer (Wing, 2011), is considered as an essential skill that will provide people with a strong competitive edge in the digital future (García-Peñalvo and Mendes, 2018). With the growing popularity of user-friendly and open-source programming languages such as Python[1], CT has already been incorporated into K-12 education in many countries such as UK (Sentance and Csizmadia, 2015), Singapore (Seow et al., 2019), and New Zealand (Bell et al., 2014) to equip students with this 21st century skills. However, studies show that although the gender gap has been narrowing down in recent years, gender differences in students' interests and attitudes toward CT and in their CT skills are still observed (Kong et al., 2018; Sullivan and Bers, 2016). In these works, gender differences are typically examined through the analysis of commonly used self-reports, i.e., measures that do not tap into the learning process, including students' individual characteristics that have been shown to be associated with gender differences, acquired CT skills, or similar learning outcomes. However, the gender differences that might emerge during the process of CT development and that are reflected in real-time human behavior have not yet been investigated due to the lack of advanced methods to measure these differences.

In recent years, virtual reality (VR) has become increasingly popular and prevalent in education, offering benefits such as supporting distance learning and teaching (Cryer et al., 2019; Hernández-de Menéndez et al., 2019; Grodotzki et al., 2018) with increased immersion (Casu et al., 2015). With the growing availability of modern head-mounted displays (HMDs), immersive VR learning experiences can be provided to students in the near future at a reasonable cost and with relatively reduced effort. For instance, immersive VR classrooms in particular, which emulate traditional classrooms, have the potential to provide more flexible and engaging learning contexts for students to develop their CT skills. Furthermore, integrated eye trackers in such setups open up additional opportunities to investigate students' gazing behavior in standardized and yet realistic environments, ultimately offering an in-depth understanding of individual differences in learning, e.g., CT skill development. Indeed, eye-tracking technology has already been used in variety of educational applications for studying learning processes, training, and assessment, such as in mathematics education (Strohmaier et al., 2020), medical education (Ashraf et al., 2018), multimedia learning (Molina et al., 2018).

Compared to commonly used questionnaires and surveys, eye tracking offers the opportunity to obtain objective measurements from subjects in a non-intrusive manner, and these measurements could also be used in real-time for various purposes. Previous studies in this context have mainly focused on investigating various aspects of human behavior using eye movements, such as stress (Hirt et al., 2020), visual attention (Bozkir et al., 2019; Gao et al., 2021), and problem solving (Eivazi and

[1] https://www.python.org/

Bednarik, 2011), either in VR or conventional setups. In addition, eye movements have been found to provide discriminative information for various psychological behavior related predictions using machine learning methods, such as cognitive load (Appel et al., 2018; Yoshida et al., 2014), personality traits (Hoppe et al., 2018; Berkovsky et al., 2019), and IQ test performance (Kasneci et al., 2022). Closer to our work, several studies in the field of human-computer interaction have shown that gender differences can be inferred from eye movements by analyzing subjects' visual viewing and search behavior with 2D stimuli (Sammaknejad et al., 2017; Hwang and Lee, 2018; Mercer Moss et al., 2012). Gender differences were also found to be predictive using classification models developed based on eye-tracking data in reading (Al Zaidawi et al., 2020) and indoor picture viewing tasks (Abdi Sargezeh et al., 2019). However, eye movement information used in these studies was limited to fixation- and saccade-related statistics, and tasks were performed in relatively simple contexts, i.e., screen-based tasks with 2D stimuli. Furthermore, the relationship between eye movements and gender has not yet been fully investigated using explainable machine learning approaches. Therefore, it is an open question whether it is possible to detect gender differences by using eye movement information in learning activities that require more effort in more complex contexts (e.g., in VR-based learning), based on machine learning and explainability approaches.

From an educational perspective, predicting gender differences based on machine learning and explainability approaches to analyzing eye movements information provides great potential: It is widely acknowledged that boys and girls differ in their achievement and interest, especially in STEM subjects, but respective educational research is predominantly based on questionnaires relying on self-reports of students. Hence, little is known about gender differences in the actual learning process. Respective insights into systematic differences of how boys and girls objectively differ in their learning behaviors, not only allow a deeper understanding of gender differences but make it possible to adapt learning environments (especially in VR settings) to the different needs of girls compared to boys to equally foster the skill development of both genders.

Therefore, in this work, we investigated to what extent eye movement data provide discriminative information for gender classification in CT learning in an immersive VR classroom. To this end, we examined a large set of eye movement features that characterize students' real-time visual attention and cognitive behaviors in a VR lesson. Several machine learning models were developed for gender classification, including Support Vector Machine (SVM), Logistic Regression, three ensemble machine learning models, i.e., Random Forest, eXtreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM). To improve the performance of the model, we performed a feature selection procedure by applying the Shapley additive explanations (SHAP) approach (Shapley, 1953). Furthermore, we interpreted the classification model by SHAP approach as well.

In summary, the contributions of this work are four-fold. **(i)** We extracted a large number of eye movement features using different time windows similar to the work of Bulling et al. (2011) from a lesson on computational thinking in an immersive VR classroom. **(ii)** We developed five machine learning models for gender classification using all extracted eye movement features. We performed feature selection using

SHAP and improved the performance of the LightGBM(best) model by training the model with the selected (best) eye movement feature set. Furthermore, **(iii)** we investigated the importance of the eye movement features and explored the effects of high contribution features on classification output, i.e., gender, using the SHAP approach. **(iv)** We provided a fundamental methodology for future studies aimed at investigating gender differences using eye movement information not only in CT but also in other learning activities in immersive VR environments for educational purposes. This could help to gain further insights to optimize the design of educational systems and thus offer personalized tutoring in such educational systems.

Related Work

As the integration of CT into K-12 STEM education has increased, so has the need for effective teaching and learning methods for CT instruction (Chalmers, 2018; Hsu et al., 2018), and an understanding of gender differences can shed light on this. In the literature to date, there are various findings on gender differences in CT development. For instance, it has been found that girls tend to show less interest and self-efficacy in STEM subjects (McGuire et al., 2020; Wang and Degol, 2017). Kong et al. (2018) conducted a study with 287 senior primary school students to investigate their interest in CT and collaboration attitude. It was found that boys showed more interest in programming than girls and thus found programming more meaningful and exhibited higher creative self-efficacy. Similarly, Baser (2013) observed that in an introductory computer programming course, males had more positive attitude towards programming than females and this attitude was positively correlated with their performance in programming. In addition, Nourbakhsh et al. (2004) examined gender differences in a robotics course designed to develop CT skills in high school students. Girls were found to have less confidence in their CT skills than boys at the beginning of the course and, according to weekly surveys, girls were more likely to report that they struggled with programming. Notably, these studies measured gender differences using students' individual characteristics in terms of their self-reported interests and attitudes toward CT. In addition, gender differences are often tied to students' achievement test results. Boys are typically found to score higher in CT tests than girls (Polat et al., 2021). Girls in turn have been found to require more training time than boys to achieve the same skill level (Atmatzidou et al., 2016). Angeli and Valanides (2020) demonstrated that boys and girls benefit from different scaffolding and learning activities when working on CT-related tasks.

The aforementioned works have highlighted the growing need for research on gender differences in CT-related education, as knowledge of how different gender groups exhibit different attitudes and learning outcomes can inform educators to better support both girls and boys who typically differ in their prerequisites and requirements for CT skill acquisition (Nourbakhsh et al., 2004; Angeli and Valanides, 2020). However, most studies on gender differences in STEM (and CT) learning do not take into account the learning process and the respective differences of boys and girls in how they acquire knowledge, especially the differences reflected in their real-time visual attention and cognitive behavior during learning; however,

they could offer valuable insights for providing tailored support or developing tailored tutoring systems aimed at reducing gender disparities in STEM subjects (CT skill development) for different gender groups, e.g., tailored VR classrooms that can be easily realized through software. Eye tracking holds such potential.

Recently, several works suggest that eye movement data provide a more intuitive interface for studying conscious and unconscious human behaviors in various tasks, such as visual search pattern recognition (Raptis et al., 2017), web search (Dumais et al., 2010), the n-back task (Appel et al., 2018), decision making measure for intelligent user interfaces (Zhou et al., 2015), and learning in the VR context (Gao et al., 2021; Bozkir et al., 2021b). Moreover, eye movements were found to complement self-reports in providing discriminative information for the prediction of subjects' psychological behavior. In terms of eye movements, Hoppe et al. (2018) utilized a machine learning approach to predict personality traits and perceptual curiosity during an everyday task solely from eye movements. In addition to personality traits, eye movement data have also been used for detecting other individual psychological behaviors. Prediction of cognitive load based on eye movements was investigated by Appel et al. (2019) and it was found that the models trained based on eye movement data were discriminative in predicting cognitive load in a simulated emergency game. Furthermore, Zhou et al. (2021) combined eye movements with demographic information and self-reports to predict situational awareness in a takeover task during conditionally automated driving. The results show that the LightGBM model, which was developed based on eye movements alone, performed better than other models in predicting situational awareness with a mean absolute error of 0.096. In addition, Kasneci et al. (2022) utilized Gradient Boosted Decision Trees (GBDT) model to examine individual differences in IQ tests by using a set of eye movement variables in combination with socio-demographic variables as features. Specifically, eye movements alone were found to be discriminative in predicting participants' intelligence performance on the Cultural-Fair IQ Test 20-R (CFT 20-R). Notably, the eye movement and socio-demographic features together were observed to provide complementary information, indicating that eye movement information is a reliable and effective behavioral measure of learning and performance processes.

Furthermore, eye movements were also found to be connected to gender. Sammaknejad et al. (2017) used eye movement data to determine gender differences in a face viewing task. It was found that male and female participants exhibited significantly different eye movement transition patterns, indicated by saccades, when viewing facial photographs of male and female subjects that were unknown to them. Similarly, gender differences in eye movement patterns were found in an indoor picture viewing task (Abdi Sargezeh et al., 2019), where females showed more exploratory gaze behavior, as indicated by longer scanpaths and larger saccade amplitudes. In their study, a support vector machine classifier was utilized to predict gender by using ten eye movement features (i.e., features related to fixations, saccades, and scanpaths), achieving an overall accuracy of about 70%. In the work of Hwang and Lee (2018), gender differences in online shopping were examined using area-of-interest information based on eye-tracking and it was concluded that females paid more attention to shopping content than males. Closer to the CT learning task in this work, different eye movement behaviors were observed in two gender groups during

an algorithmic problem solving task (Obaidellah and Haek, 2018), with females fixating more on the indicative verbs, while males fixated more on the operational statements. However, in contrast to the present study, the participants were undergraduates and the task was performed in a conventional context with screen-based stimuli. Al Zaidawi et al. (2020) performed gender classification of children aged 9-10 years, whose ages were comparable to the participants in this work, based on eye movements with reading stimuli using machine learning. Several classifiers were developed based on a group of extracted fixation- and saccade-related features, and accuracies of 63.8% and 60.7% were achieved for the non-dyslexic and dyslexic participants groups, respectively. Furthermore, gender differences in eye movement behavior were found not only in these 2D-based stimulus tasks but also in 2D-based stimulus tasks. In a map direction pointing task (Liao and Dong, 2017), males fixated on landmarks significantly longer than females in the 3D map, and a reverse difference between males and females was observed in the 2D map. In addition to these 2D screen-based experimental setups, gender was found to be predictable in a VR-based reading task using only eye movement features and support vector machines, with accuracies near 70% (Steil et al., 2019); however, the reading stimuli in VR were still 2D.

As previous works have examined gender differences in relatively conventional contexts (i.e., with screen-based stimuli) and in VR contexts (i.e., with 2D stimuli in VR) with limited spatial and temporal characteristics of eye movements, first, it is an open question whether such findings apply to immersive virtual reality environments (e.g., learning environments) for developing computational thinking. It should be mentioned that, to our knowledge, there is no similar research on gender classification based on eye movements with 3D stimuli rendered in VR learning environments. Second, more complex models with multi-modal data and model explanation approach can reveal relationships between gender and the most contributed features, ordered by feature importance, to support computational thinking training rather than analyzing differences separately based on summary statistics (e.g., mean fixation or saccade duration). Gender information is in fact considered protected and should be hidden in the data (Steil et al., 2019; Bozkir et al., 2021a), especially when using the commercial application. However, gender recognition is critical for education domain and the development of commercial and noncommercial human-computer interaction applications, and has been studied in depth by a number of researchers (Lin et al., 2016). Particularly in subjects where gender differences typically exist (Reilly et al., 2017), gender prediction can help in providing personalized support during learning and further expand implications for the design of VR and intelligent tutoring systems. Therefore, we investigate gender differences as a proof-of-concept by using eye movements that are obtained in a learning space in an immersive VR environment.

Dataset

In this section, we give an overview of data acquisition and preprocessing.

(a) Overall view of the VR classroom (b) Participant's view

Fig. 1 Computational thinking (CT) learning in an immersive virtual reality classroom

Data Acquisition

Participants

Data were collected from 381 sixth-grade volunteer students (179 female, 202 male; average age = 11.5, SD = 0.5), including eye-tracking data and questionnaire data (i.e., demographic data, self-reports of participants' learning background, and VR learning experience). The study was IRB-approved and all participants and their legal guardians provided informed consent in advance.

Apparatus

We used the HTC Vive Pro Eye with a refresh rate of 90 Hz and a field of view of 110° in our study. Eye-tracking data were recorded using the integrated Tobii eye tracker with a sampling frequency of 120 Hz after a 5-point calibration routine. Virtual environment was rendered using Unreal Game Engine v4.23.1.

Experimental Design and Procedure

An immersive VR classroom as depicted in Fig. 1, similar to conventional classrooms, was used for data collection. Virtual avatars, including the teacher and peer learners, were rendered in one of two visualization styles (i.e., cartoon and realistic). Participants sat in the classroom (i.e., front or back row) and listened to an approximately 15-minute virtual lesson about basic CT principles delivered by the virtual teacher. The lesson *"Understanding how computers think"* consists of four sessions. In the first session, the virtual teacher gives an introduction to CT and asks five simple questions to prompt interaction with learners; in the second session, the teacher explains the terms "loop" and "sequence" to the learners, with four questions following each explanation; in the third session, after the knowledge input by the teacher, the learners are given two exercises to apply the learned content, after a

short reflection period, the teacher gives the answers to each question; in the fourth session, the teacher ends the lesson with a summary. During the VR lesson, all relevant learning content, including the CT terms, questions, answers, and exercises, are displayed on the screen on the front wall of the VR classroom. In addition, to mimic a real classroom and increase immersion, a fixed percentage (20%, 35%, 65%, 80%) of virtual peer learners interact with the teacher by raising their hands after each question and by turning around from time to time throughout the lesson.

Each experimental session took about 45 minutes in total, including a paper-based pre-test, the VR lesson, and a post-test. Since the different experimental conditions (i.e., the aforementioned visualization styles of the virtual avatar, the seating positions of the participants in the VR classroom, and the percentages of virtual peer learners who were preprogrammed with hand-raising behavior) were not the focus of the present study (see detailed investigation in our previous work, Gao et al. (2021), wetrained our classification models based on all eye-tracking data.

Data Preprocessing

We collected raw sensor data including participants' head-poses, gaze vectors, and pupil diameters. Data from participants who experienced sensor-related issues, such as low tracking ratios (less than 90% of eye-tracking signal was recorded), incomplete VR lesson experiences, were excluded. Given that the summary session of the VR lesson ($\approx$ 1.5 minutes) does not include learning activities, we excluded the data from this session. Consequently, data from 280 participants (140 female, 140 male) were used with an average of 13 minutes of head-pose and eye-tracking data. To ensure the quality of the data, we then performed preprocessing of the data for further feature engineering as follows.

Since pupillometry data are affected by noisy sensor readings and blinks, we smoothed and normalized pupil diameter using Savitzky-Golay filter (Savitzky and Golay, 1964) and the divisive baseline correction method (Mathôt et al., 2018) with a baseline duration of $\approx$ 1 seconds, respectively. We used a $7°/s$ threshold to detect stationary ($< 7°/s$) and moving ($> 7°/s$) head activities similar to the work of Agtzidis et al. (2019). In addition, we performed a linear interpolation for the missing gaze vectors. Eye movement events, including fixations and saccades, were detected based on a modified Velocity-Threshold Identification (I-VT) method suitable for the VR setting which takes into account head movements (Agtzidis et al., 2019). In the absence of prior knowledge on how to determine gaze velocity and duration thresholds for fixation and saccade detection in the VR learning context, we set these thresholds based on previous literature (Salvucci and Goldberg, 2000; Holmqvist et al., 2011; Agtzidis et al., 2019), but make some adjustments to fit our study. Fixations were detected within stationary head activities using a maximum gaze velocity threshold of $30°/s$, with additional thresholds for a minimum duration of $100ms$ and maximum duration of $500ms$. Saccades were detected by a minimum gaze velocity threshold of $60°/s$ with additional thresholds for a minimum duration of $30ms$ and maximum duration of $80ms$.

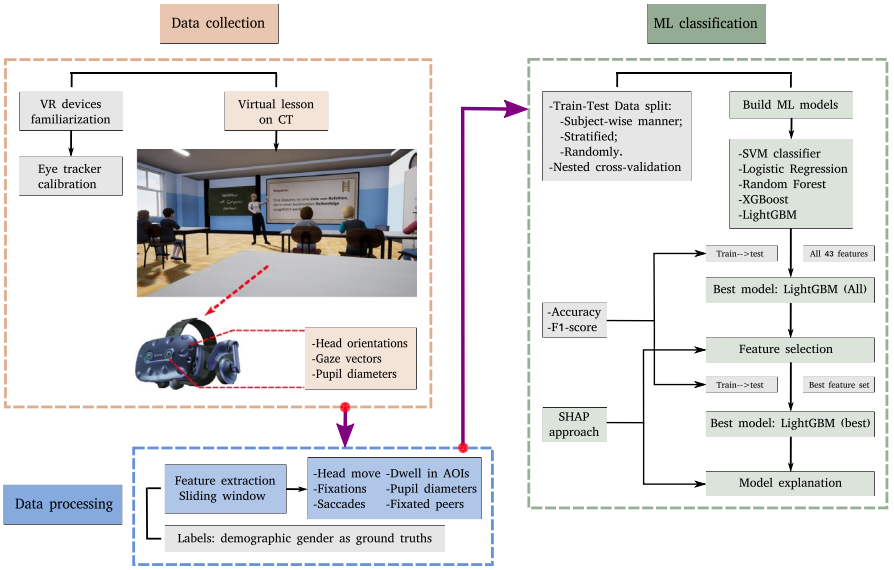

Fig. 2 The ML approach for gender prediction in CT development in the immersive VR classroom

Methods

In this section, we discuss the feature extraction pipeline, machine learning model development for gender classification, model evaluation, and the SHAP explanation approach for feature selection and model interpretation. The procedure of our ML approach for gender prediction is shown in Fig. 2.

Feature Extraction

To extract the temporal features from the sensory data, we adopted a sliding-window approach similar to previous studies (Bulling et al., 2011; Hoppe et al., 2018). Since there is no gold standard for the selection of window size in VR learning scenarios, considering different preprogrammed activities (see Section 3.1) occur during the virtual lesson, we initially used a set of window sizes ranging from 10*s* to 100*s* with a step of 10*s*. For each window, a vector of 43 features was extracted, most of which related to eye movement information, while one feature was HMD-related; the details of the extracted features and the description of the features are given in Table 1. For simplicity, we refer to these 43 features collectively as the eye movement features in the following.

- HMD-related features: In a previous study, head movements were analyzed and found to be indicative of shifts in social attention during participation in a virtual classroom (Seo et al., 2019). Therefore, we used similar measurement in this

Table 1 Eye movement features extracted for gender classification model

Feature	Description
1. HMD moving rate	Number of moving head activities per second;
2. Fixation rate	Number of fixations per second;
3-7. Fixation duration	Mean, min, max, sum, SD of fixation duration;
8-10. Number of fixation on object of interest	Number of fixations on peer learners, teacher, and screen;
11-22. Duration of fixation on object of interest	Mean, min, max, SD of fixation duration on peer learners, teacher, and screen;
23-25. Dwell time on object of interest	Dwell times on peer learners, teacher, and screen;
26. Saccade rate	Number of saccades per second;
27-31. Saccade duration	Mean, min, max, sum, SD of saccade duration;
32-36. Saccade amplitude	Mean, min, max, sum, SD of saccade amplitude;
37-40. Saccade peak velocity	Mean, min, max, SD of saccade peak velocity;
41-42. Pupil diameter	Mean, SD of pupil diameter;
43. Fixated peer learners	Number of peer learners fixated by the participant.

Min, max, and SD stand for the minimum, maximum, and standard deviation of the relevant features

study. The number of moving head activities per second, i.e., **hmdMoveRate**, is used as a feature in the classification models (Feature 1 in Table 1).

- Fixation-related features: Fixations are periods of time when the visual gaze is maintained in a single location. Fixations have been used to understand the learning processes and are considered indicators of attention and cognitive processing activity (Negi and Mitra, 2020; Chien et al., 2015). We used features related to fixations, including **fixationRate** and **fixationDuration** (Features 2-7 in Table 1). Moreover, in our previous study (Bozkir et al., 2021b), we found that participants' attention in the VR classroom mainly switches between three virtual objects, also called object-of-interest (OOI), including the virtual peer learners, the virtual teacher, and the screen displaying the instructional content. Therefore, we extracted the number of fixations on these OOIs, i.e., **fixationNumberOnPeer/Teacher/Screen**, as well as their duration, i.e., **peer/teacher/screenFixationDuration** (Features 8-22 in Table 1). Since dwell time quantifies the time spent looking within an OOI, which includes all fixations and saccades within the OOI as well as revisits (Holmqvist et al., 2011), we additionally extracted **dwellOnPeer/Teacher/Screen** (Features 23-25 in Table 1). In addition, we extracted the number of peer learners fixated by the participant during the virtual lesson, i.e., **fixatedPeerNumber** as a feature (Feature 43 in Table 1).
- Saccade-related features: Saccades indicate the rapid shift of the eye from one fixation to another and are also informative eye movements that are highly correlated with visual search behavior (Holmqvist et al., 2011). In a fixation-like manner, we have the number of saccades per second, i.e., **saccadeRate**, and their durations, i.e., **saccadeDuration**. Additionally, **saccadeAmplitude** and **saccadePeakVelocity** are employed as features in our models (Features 26-40 in Table 1).

- Pupil-related features: It is known from previous studies that pupil diameter reflects cognitive load in various human cognitive processes, e.g., visual attention during scene perception (Gao et al., 2021), visual search (Castner et al., 2020), and sustained attention (Appel et al., 2018). Therefore, we extracted features related to pupil diameter (Features 41-42 in Table 1).

Classification Models

We used the reported gender demographic data as the ground truth and used discrete variables to represent each gender, i.e., for the purpose of machine learning research only, we set 0 for the female class and 1 for the male class. To be clear, these two numbers have no specific meaning. In this work, we developed five supervised machine learning models to detect participants' gender in a VR lesson on CT learning using eye movement features. Specifically, we used Support Vector Machine (SVM), Logistic Regression, and three ensemble machine learning models, namely Random Forest, eXtreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM). Thus, a binary classification task was performed on our dataset, which consists of features and the target: (x_{pw}, y_{pw}), $x_{pw} = [x_{pw1}, x_{pw2}, ..., x_{pwN}]$, $x_{pw} \in R^d$, where $1 < p < K$, $K = 280$, p represents the sequence number of the participant, $1 < w < M$, $M = 780s/window\_size$, w represents the sequence number of time windows of each participant, and N represents the total number of the eye movement features used, $N = 43$. In short, x_{pw} is the $p * w$-th input vector of all features used for model training; $y_{pw} = \{0, 1\}$ is the target variable (gender), where 0 is class-0 (class-female) and 1 is class-1 (class-male). Before model training, predictor variables were normalized using the maximum-absolute scaling normalization technique.

For training the model based on eye movement features extracted with a specific time window, we performed the nested cross-validation approach to optimize open parameters, i.e., the hyperparameters of the models and window size. A stratified 5-fold cross-validation strategy was applied. For each iteration, we divided the data into a training set, a validation set, and a test set. Particularly, in each iteration, we selected 20% of participants as the test set, 20% of the remaining participants as the validation set, and the rest of the participants as the training set. Thus, for example, with a window size of 60s, there are more than 3600 data samples, including 2900 data samples for training, and 700 data samples for testing. After 5-fold cross-validation, all participants appeared in the training set and the test set. Note that to avoid overfitting and to generalize our models to unseen data, we performed all data splits in a participant-dependent manner, meaning that all data samples from the same participant should remain in one data set (i.e., either the training, validation, or test set). In addition, participants were randomly assigned without regard to identity. Furthermore, we performed 5-fold cross-validation 10 times, each time selecting different groups of participants as the test set, which further eliminated the participant-group effect on the model. Thus, our models were trained for 50 iterations, and in each iteration, five models were trained on the training set and evaluated on the validation set. We used the F1-score to select the most optimized model hyperparameters. The

best hyperparameters were selected based on the validation results. For the final performance evaluation, we trained our models on the unified training and validation set and tested them on the test set to generalize our models to unseen data. Since our dataset is nearly balanced with respect to gender, the chance level is about 50%.

Model Evaluation

Since we have a balanced dataset for binary classification, we evaluated the performance of the models in terms of accuracy and F1-score. TP, TN, FP, and FN stand for True Positive, True Negative, False Positive, and False Negative, respectively. $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$, is the ratio of correctly predicted observations to total observations. $Precision = \frac{TP}{TP+FP}$, is the ratio of correctly predicted observations to the total predicted positive observations. $Recall = \frac{TP}{TP+FN}$, measures the percentage of true positives that are identified correctly. We measured the F1-score, i.e., $F_1 = 2 \times \frac{Precision*Recall}{Precision+Recall}$, as it accounts for both false positives and false negatives.

SHAP Explanation Approach

Explainability becomes significant in the field of machine learning as it provides insights into how a model can be improved. SHAP (Shapley additive explanation), a game-theoretic approach, is one of the proposed methods to support the interpretation of the prediction results and analyzing the importance of individual features, where the individual feature values are assumed to be in a cooperative game whose payout is the prediction (Shapley, 1953). Given that the Shapley value of a feature is its contribution to the payout, weighted and summed across all possible feature value combinations, the Shapley value for a model with a prediction function of $f(x)$ is given as follows. Given $F = \{x_1, x_2, \ldots, x_N\}$ including all the features,

$$\phi_j(f_x) = \sum_{S \subseteq F \setminus \{x_j\}} \frac{|S|!(N - |S| - 1)!}{N!} \left(f_{p=S\cup\{j\}}\left(x_p\right) - f_S(x_S)\right) \tag{1}$$

where S is a subset of features and N is the number of features. $\phi_j(f_x) \in R$ stands for the Shapley value of the feature vector x_j (Lundberg and Lee, 2017). In our study, we used the local feature attribution method for tree models, TreeExplainer, introduced by Lundberg et al. (2020). TreeExplainer bridges theory and practice by building on previous model-agnostic research based on classic game-theoretic Shapley values (Shapley, 1953; Štrumbelj and Kononenko, 2014; Datta et al., 2016; Lundberg and Lee, 2017; Sundararajan and Najmi, 2020). More details on TreeExplainer can be found in the study by Lundberg et al. (2020).

SHAP feature importance is measured as mean absolute Shapley values. We used the SHAP approach not only to explain the machine learning models at the feature level (explaining the effects of each predictor variable on the model output, i.e., gender), but also to perform feature selection according to the calculated SHAP feature importance to improve the performance of the classification model. See details in Section 5.

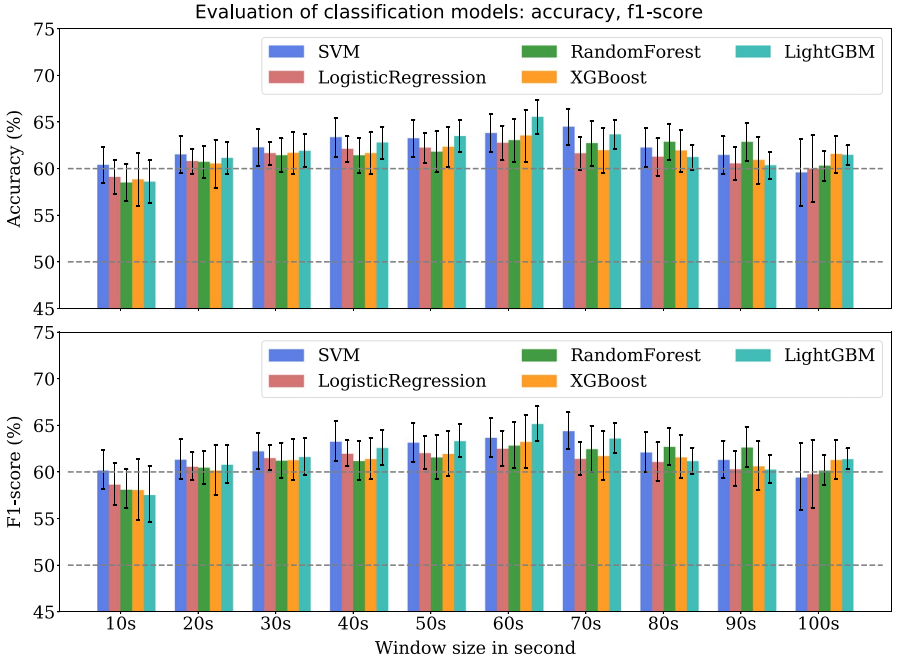

Fig. 3 Performance of all five classification models trained with 43 eye movement features extracted using 10 different time windows; mean values and standard deviations calculated from 50 training iterations

Results

We present gender classification, feature selection by SHAP, and SHAP explainability results as follows.

Classification Results

To find the optimal window size for eye movement feature extraction in gender classification, we extracted ten sets of eye movement features in ten different time windows. Then we trained five classification models separately with ten extracted feature sets. The hyperparameters of the classifiers were tuned during the training process. Figure 3 shows the performance results of five models trained with all eye movement features extracted with different window sizes. As shown, the performance of five models is higher than the chance level (50%) in all time windows. Particularly, we found that a window size of 60*s* provides an optimal trade-off for gender classification compared to time windows of other lengths, as shown by the evaluation results of all five models (see Fig. 3). Therefore, it can be stated that 60*s* is the optimal window size for eye movement feature extraction in gender classification in this study. In the following analysis, we report the performance of gender

Table 2 Performance of all five classification models trained with all (43) eye movement features and with the selected (best, *feature_number*) eye movement features

Classification model	Accuracy	F1-score
SVM (all)	63.8(±2.1)	63.7(±2.1)
Logistic Regression (all)	62.7(±1.9)	62.5(±1.9)
Random Forest (all)	63.1(±2.3)	62.9(±2.5)
XGBoost (all)	63.5(±2.8)	63.3(±2.7)
LightGBM (all)	**65.5(±1.8)**	**65.2(±1.8)**
SVM (best, 26)	66.7 (±1.9)	66.5 (±1.9)
Logistic Regression (best, 21)	67.1 (±2.2)	66.9 (±2.2)
Random Forest (best, 20)	67.2(±2.1)	67.4 (±2.2)
XGBoost (best, 26)	67.9 (±2.4)	67.8 (±2.3)
LightGBM (best, 24)	**70.8(±1.7)**	**70.6(±2.5)**

'all' and 'best (*number*)' mean that the model is trained with all (43) eye movement features and the best eye movement feature set selected by the SHAP approach, respectively. The bold font represents the best performance of the model trained with different feature sets

classification models trained with eye movement features extracted with a window size of 60*s*.

The comparative performance of all five classification models based on all 43 eye movement features extracted a window size of 60*s* is shown in the upper part of Table 2. In particular, the LightGBM classifier performs best with an average accuracy of 65.5% ($SD = 1.8\%$), followed by SVM ($M = 63.8\%$, $SD = 2.1\%$) and XGBoost classifier ($M = 63.5\%$, $SD = 2.8\%$). The best hyperparameters used for each model trained with all features are listed in the upper part of Table 3.

Furthermore, we performed feature selection according to the feature importance, represented by Shapley values, to further improve the performance of the machine learning models in gender classification. Specifically, we trained all five models separately with a series of selected feature sets: There are 43 eye movement features in the current feature set, we dropped the least important feature (feature with the lowest feature importance according to the SHAP approach) and trained the model with the remaining features (i.e., the selected feature set); we used the selected feature set as the current feature set for the next loop; in each loop, we dropped the least important feature from the current feature set, we continue the loop until there is only one feature left in the selected feature set. In this way, 43 feature sets with length from 43 to 1 were obtained and used for training.

The comparative performance of all five classification models based on selected best eye movement features extracted with a window size of 60*s* is shown in the lower part of Table 2. As can be seen, all models achieved better performance after feature selection by SHAP than the models trained with all 43 features. In particular, the improvement of the LightGBM model is the largest, over 5% improvement in accuracy from 65.5% ($SD = 1.8\%$) to 70.8% ($SD = 1.7\%$) trained with the top 24 features. In contrast, the SVM classifier shows the least improvement in accuracy, about 3%, trained with the top 26 features. Nevertheless, LightGBM still achieved

Table 3 Best hyperparameters of all five classification models trained with all (43) eye movement features and with the selected (best) eye movement features

Classification model	Best hyperparameters
SVM (all)	'C': 1, 'kernel': 'RBF', 'gamma': 0.01;
Logistic Regression (all)	'penalty': 'l2', 'C': 0.1, 'solver': 'newton-cg', 'max_iter': 10000;
Random Forest (all)	'n_estimators': 100, 'max_depth': 10, 'min_samples_split': 15, 'min_samples_leaf': 20, 'max_features': 'log2';
XGBoost (all)	'gamma': 0.1, 'learning_rate': 0.01, 'max_depth': 3, 'min_child_weight': 12, 'subsample": 0.7, 'colsample_bytree': 0.6, 'reg_lambda': 0.5, 'reg_alpha': 0.8;
LightGBM (all)	'n_estimators': 600, 'num_leaves': 150, 'learning_rate': 0.01, 'max_depth': 8, 'min_child_weight': 0.01, 'min_child_samples': 180, 'subsample": 0.7;
SVM (best)	'C': 10, 'kernel': 'RBF', 'gamma': 0.1;
Logistic Regression (best)	'penalty': 'l1', 'C': 0.01, 'solver': 'newton-cg', 'max_iter': 10000;
Random Forest (best)	'n_estimators': 150, 'max_depth': 8, 'min_samples_split': 15, 'min_samples_leaf': 26, 'max_features': 'log2';
XGBoost (best)	'gamma': 0.1, 'learning_rate': 0.01, 'max_depth': 5, 'min_child_weight': 12, 'subsample": 0.7, 'colsample_bytree': 0.5, 'reg_lambda': 0.5, 'reg_alpha': 0.6;
LightGBM (best)	'n_estimators': 600, 'num_leaves': 200, 'learning_rate': 0.01, 'max_depth': 8, 'min_child_weight': 0.01, 'min_child_samples': 200, 'subsample": 0.8.

'all' and 'best' mean that the model is trained with all (43) eye movement features and the best eye movement feature set selected by the SHAP approach, respectively

the best performance in gender prediction among all five models after feature selection. Here, only the feature selection results of the best performed LightGBM model are given, as shown in Fig. 4. The best hyperparameters used for each model trained with the selected best features are listed in the lower part of Table 3.

SHAP Explanation

To understand the contribution of each eye movement feature to the output of the LightGBM(best) model trained with the top 24 features, we calculated the average SHAP values for each feature in the test data used in the model. Since LightGBM is tree-based model, the TreeExplainer[2] SHAP was used. A global feature importance and a local explanation summary plot of SHAP values that combines feature importance with feature effects are shown in Figs. 5 and 6, respectively. In the global feature importance plot, a standard bar-chart based on the average magnitude of the SHAP values is illustrated, where the x-axis represents the average impact of the features on the model output. In addition, local explanations are plotted in a beeswarm-style SHAP summary plot to examine both the prevalence and magnitude of features' effect. Each point in the summary plot represents the SHAP value for a

[2] https://shap-lrjball.readthedocs.io/en/latest/index.html

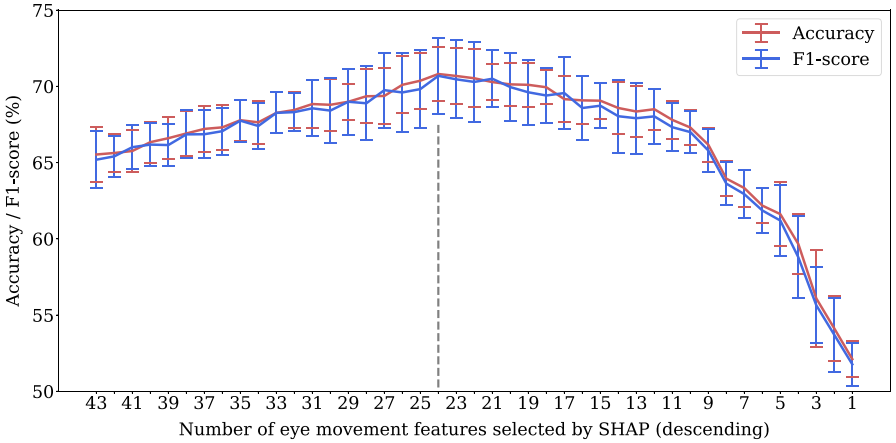

Fig. 4 Performance of the LightGBM model trained with a series of eye movement feature sets obtained by sequentially dropping the least important feature; mean values and standard deviations calculated from 50 training iterations

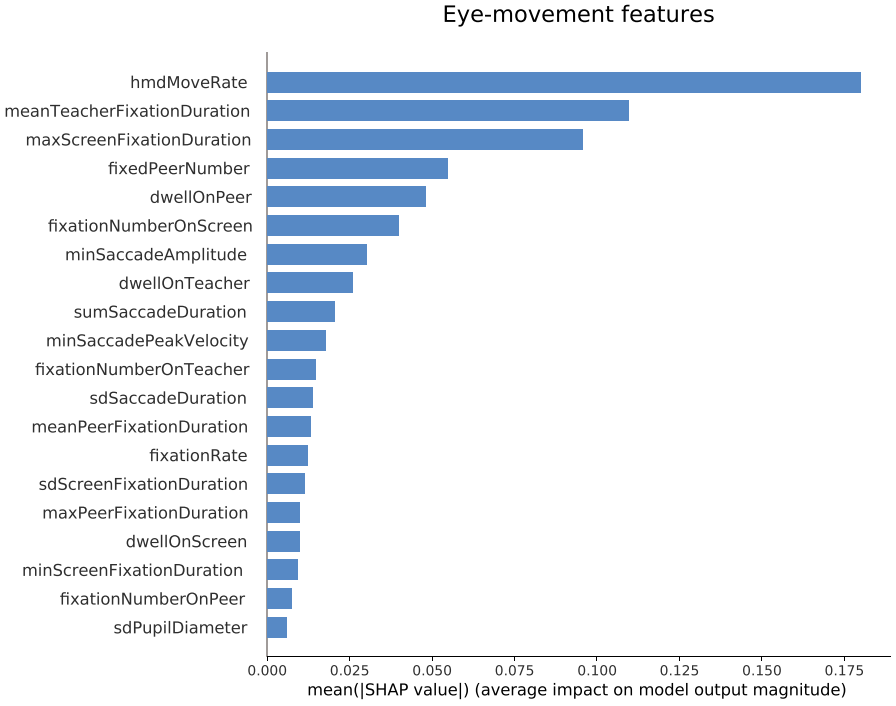

Fig. 5 SHAP global feature importance plot: the top 20 features of the LightGBM(best) model trained with the best eye movement feature set. Bar plot of mean absolute SHAP values of individual features

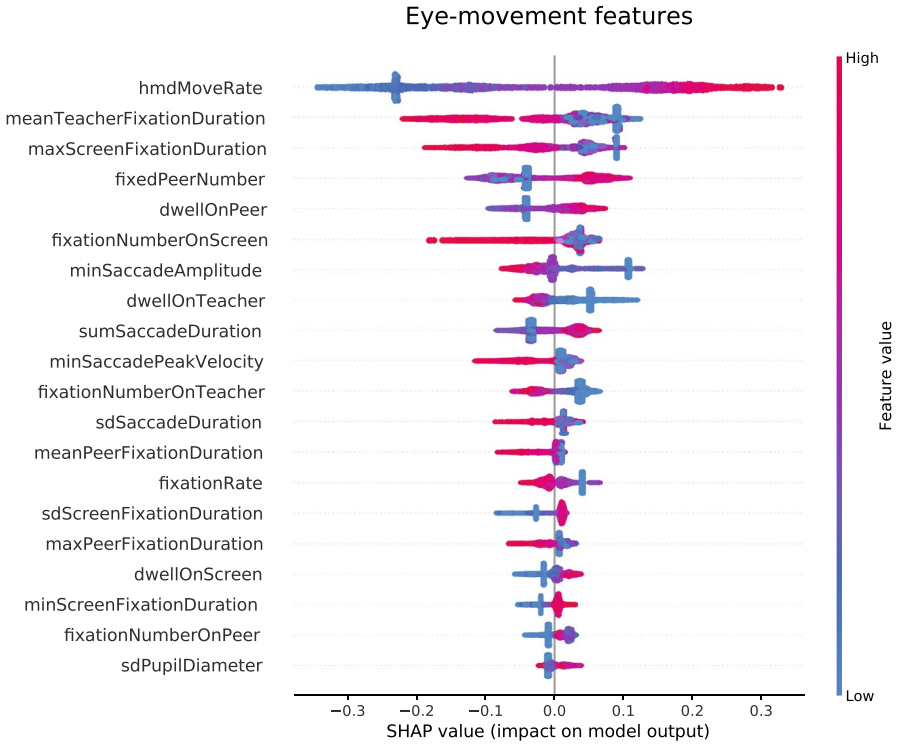

Fig. 6 SHAP local explanation summary plot: the top 20 features of the LightGBM(best) model trained with the best eye movement feature set. The color change in the summary plot (from left to right) of each feature from blue to red indicates a positive influence on classification into class-1; conversely, from red to blue indicates a negative (positive) influence on classification into class-1 (class-0)

feature and an observation. The position of each point is determined by the feature on the *y*-axis and by the SHAP value on the *x*-axis. For each feature, blue and red colors indicate low and high feature value, respectively. A change (from left to right) in the color from blue to red along the *x*-axis indicates that the feature has a positive impact on the prediction of the class-1 (class-male) and, on the contrary, a change (from left to right) from red to blue indicates a negative impact of the feature on the prediction of the class-1 (class-male) and thus a positive impact on the prediction of the class-0 (class-female). The greater the impact of a feature on the model output, the more spread out it is on the *x*-axis. In both plots, the features are sorted according to their importance from top to bottom in the *y*-axis.

As shown in Figs. 5 and 6, features **hmdMoveRate**, **meanTeacherFixationDuration**, and **maxScreenFixationDuration** followed by **fixedPeerNumber** and **dwellOnPeer** provide the maximum information for the LightGBM(best) model in gender classification. Interestingly, we found that pupil- and saccade-related features contribute less to the classification model than HMD- and fixation-related features, and many informative features are features related to objects of interest (i.e., the

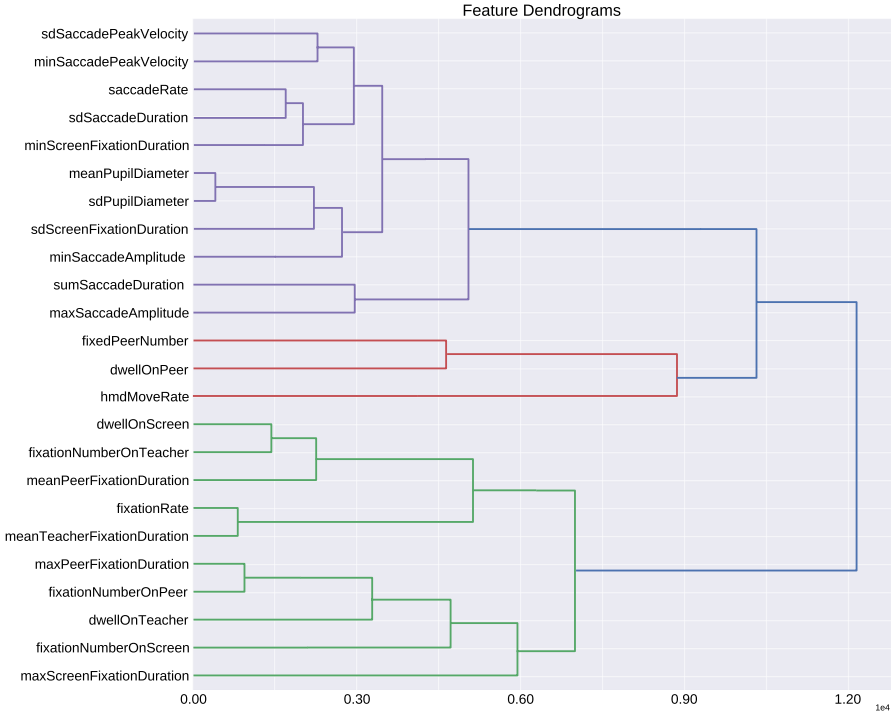

Fig. 7 Hierarchical clustering dendrogram for the features used in the best LightGBM

virtual teacher, virtual peer learners, and the screen) in VR. Furthermore, as can be seen in Fig. 6, which shows the local explanation of feature importance, features affect the model differently. For instance, the features **hmdMoveRate**, **fixedPeerNumber**, and **dwellOnPeer** have the highest positive influence on classification into class-1 (i.e., class-male); while the features **meanTeacherFixationDuration** and **maxScreenFixationDuration** have the highest negative influence on classification into class-1 (i.e., class-male), in other words, the highest positive influence on classification into class-0 (i.e., class-female).

Furthermore, before further discussion of the SHAP results in Section 6, we examine the hierarchical relationship between the features to check for redundancy. A dendrogram of the top 24 features used for the best LightGBM was created, as shown in Fig. 7.

Discussion

As most gender differences in CT development are assessed by commonly used subjective measures, i.e., questionnaires, acquired CT skills, or similar learning outcomes based on statistical analyses, gender differences assessed by using participants' temporal object behavior with machine learning techniques have not been addressed in previous work. In this work, we investigated the detection of gender

differences in CT development process in a VR context using eye movements, which provide a non-intrusive and real-time measure of participants' cognitive and visual behavior during the (VR) learning process. We developed five models using supervised machine learning techniques for gender classification and trained the models with all 43 eye movement features extracted from the recorded eye-tracking data using the selected optimal window size (60*s*, see Fig. 3). The results show that all classification models perform above chance level (50%), with LightGBM performing best among all models, followed by SVM and XGBoost (see Table 2). To improve the performance of the best classification model, i.e., LightGBM, feature selection was implemented using the SHAP approach, and the best eye movement feature set was identified (i.e., top 24 eye movement features according to SHAP feature importance, see Fig. 4). As a result, an improved average accuracy of 70.8% ($SD = 1.7\%$) for gender classification was achieved after feature selection (see Table 2). Therefore, it can be concluded that participants' eye movements provide discriminative information for the LightGBM(best) model in classifying gender. Our gender classification results are in line with previous literature (e.g., 64% in Al Zaidawi et al. (2020); 70% in Abdi Sargezeh et al. (2019)), although the stimulus and tasks in our immersive VR setup differ significantly from conventional 2D stimuli used in these works. This is further evidence that even though people's viewing behavior and attention models differ in virtual and real worlds, eye movements that are considered non-intrusive can nevertheless reveal gender differences in learning in VR environments where eye movements and their analysis are more complex. Another previous work (Steil et al., 2019) achieved about 70% accuracy in gender classification, which is similar to the accuracy of our work. As the reading task in their study was performed in a VR context, our results are more comparable; however, the reading task (2D reading stimulus in VR) is much less complex than the CT skills development in an immersive VR classroom in our study. These previous works have further implications for the success of our study regarding the use of eye movement features in predicting gender during more complex learning activities, i.e., CT development, in an immersive VR environment.

In addition to using the SHAP approach for feature selection, the SHAP was used to explain the model, namely to examine the contribution of features to the classification model and the relationships between predictor variables and the target. Specifically, we found that the HMD-related and fixation-related predictor variables influenced the classification model more than the saccade-related and pupillary predictor variables (see Figs. 5 and 6). Notably, the majority of the features in the best eye movement feature set are features related to virtual objects (i.e., the virtual teacher, virtual peer learners, and the screen) in VR, suggesting that participants' eye movement behavior toward these objects in VR provides discriminative information for differentiating gender differences. This finding suggests that the way participants attribute their visual attention to the different classroom content (i.e., the virtual teacher and the screen are instructional content, and virtual peer learners are social comparison information) reflects their gender information. From an educational perspective, this is crucial to know when aiming to design learning environments and instructional support tailored to the different needs of girls compared to boys in STEM subjects (in this case, the development of CT skills). The

results imply that, for instance, intelligent tutoring support by virtual peer learners might vary in its positive impact as it is likely to receive different levels of attention depending on the gender of the learners. Similarly, scaffolds and guidance in the instructional material are attended to differently by boys compared to girls, necessitating differentiated implementation (in line with the conclusions drawn by Angeli and Valanides (2020).

Notably, the frequency of participants' head movements was found to be the most informative feature for predicting gender. It was observed that head movement has the greatest influence on classification into class-male, implying that more head movements could be associated with male gender and, in particular, that it is highly likely that boys show more head movement than girls during a VR lesson on CT. This finding is consistent with previous research based on participants' self-reports (Kong et al., 2018; Baser, 2013) showing that boys typically exhibit higher levels of interest and self-efficacy in CT and consequently exhibit more exploratory behavior, which is particularly reflected in head movements and further exemplified by eye movements. With regards to the pedagogical design of CT learning environments, this result suggests that girls require more guidance when exploring the VR classroom; girls' interest and self-efficacy in the CT lesson need to be explicitly promoted whereas boys naturally tend to exhibit respective behaviors.

In addition, there are several fixation-related features that have a high impact on the gender classification model, including fixations and dwell on virtual objects, as well as non-specific fixations. This suggests that different gender groups may exhibit different attentional behaviors while participating the VR CT lesson. In particular, participants' attentional behavior toward the virtual teacher has high contribution to the classification model. The features mean fixation duration on the teacher, number of fixations on the teacher, and dwell time on the teacher have high negative impacts on classification into class-male, i.e., a lower feature value than the feature average drives classification into class-male, whereas a higher feature value than the feature average drives classification into class-female. This suggests that different gender groups might display different attentional behavior toward the virtual teacher: Girls were more likely to direct their visual attention to the virtual teacher than boys. Regarding the screen, another instructional object of interest in VR, we found similar finding as for the teacher. In particular, the features maximum fixation duration and number of fixations on the screen have high negative impacts on classification into class-male, implying that a lower feature value than the feature average drives classification into class-male. This may indicate that boys tend to pay less attention to the screen compared to girls. Taken together, our results suggest that the girls' group may pay more visual attention (as indicated by longer and more fixations) to the instructional content (i.e., the virtual teacher and the screen) than the boys' group, which is in line with previous work that girls pay more attention to learning material (Papavlasopoulou et al., 2020).

Moreover, participants' attention to virtual peer learners was also found to provide discriminative information for the model. The features number of peers fixated by participants and dwell time on peers are positively related to the gender classification output; a feature value higher than the feature average leads to classification into class-male. This may indicate that participants' visual attention to

virtual peer learners differs between genders, with boys more likely to direct their visual attention to the social comparison information (virtual peer learners) than girls. This could be further supported by the positive impact of feature sum of the saccade duration on classification into class-male, as the teacher, the screen, and especially the peer learners draw attention from different directions around the participants, resulting participants exhibiting longer visual search behavior.

Taken together, all these results further indicate that girls and boys might differ with regards to how they distribute their attention in the VR lesson. Girls, in particular, seem to focus more on the instructional and lesson content. Consistent with previous research (Atmatzidou et al., 2016; Angeli and Valanides, 2020), these results may indicate that girls need more time and more conversation-based instructional support to acquire CT skills compared to boys. Based on the results, specifically instructional guidance provided by the teacher is likely to support girls' CT learning, as they tend to focus more on the teacher compared to virtual peer learners (who could also serve as intelligent tutors in a VR classroom but are more at the focus of boys' visual attention). In addition, the results indicate that boys appear to distribute their attention more across the whole classroom and spend more time decoding the information provided by peer learners in addition to the teacher and the screen. This is consistent with the findings on the differences in head movements between girls and boys (see above) and suggests stronger exploratory behavior in boys, which in turn is likely to indicate greater interest in CT among boys as well as their overall more positive attitudes toward CT development (Baser, 2013; Polat et al., 2021; Kong et al., 2018).

Notably, the inferred learning behaviors of boys and girls based on their eye movements in this study reflect relatively high-level and aggregated behavior concerning the overall visual search and attention in the VR classroom rather than responses to specific aspects of the lesson and learning materials. Therefore, implications from an educational perspective concern primarily the overall learning environment design and general instructional support in the VR classroom. Asserting these overall gender differences in learning behaviors in the VR environment provides ground for more fine-grained analyses that further inform the design of tailored learning environments and instructional supports for girls and boys, respectively. SHAP approach provides a valuable way to interpret machine learning models at the feature level by examining the importance of different features that contribute differently to the LightGBM(best) model and by revealing the relationship between eye movement features and gender. Our SHAP results suggest that participants' visual search and visual attention behaviors, particularly attention to different sources of information in the VR environment, such as the virtual teacher (i.e., instruction), the screen (i.e., lesson content), and peer learners (i.e., social orientation), provide different amount of discriminative information for gender classification during a VR lesson on CT. Future studies can build on these findings to examine gender differences in the response to more specific aspects of the learning experience (e.g., critical time points in the lesson, certain presentations of the instructional material) and in other subjects, particularly in those that typically yield gender differences and take place in VR environments.

Conclusion

To our knowledge, our study is the first to examine gender differences in a VR lesson based on only eye movements using machine learning techniques. Five classification models developed based on all extracted eye movement features were found to be predictive of gender, with LightGBM outperforming the other models. The LightGBM(best) model, which was developed based on the best eye movement feature set selected by SHAP approach, showed an improved performance with an average accuracy of over 70%. In addition, the SHAP approach was used for model interpretation. Our study provides a systematic way to detect gender and explore the relationship between different eye movements (e.g., different attentional, exploratory, search, and cognitive behaviors during CT learning in VR) and gender.

Our findings provide an important foundation for future use of eye movement data to study gender differences in learning in educational contexts, particularly in VR scenarios. Future research building on the findings of this work may offer a promising avenue for improving teaching and learning processes aimed at narrowing gender gaps by gaining a comprehensive understanding of how eye movement features differentially contribute to gender classification: for example, optimizing environmental design in terms of instructional content (i.e., teacher's instructions and screen-based content) and social counterparts (i.e., peer learners) rendered in VR, or optimizing relevant factors for user interface design tailored to different gender groups. These insights offer important implications for the design of future adaptive tutoring systems for educational purposes, particularly in STEM subjects such as CT where gender differences remain pronounced.

Acknowledgements The authors thank Jens-Uwe Hahn, Stephan Soller, Sandra Hahn, and Sophie Fink from the Hochschule der Medien Stuttgart for their work and support related to the immersive virtual reality classroom used in this study.

Author Contributions Hong Gao: Methodology, Data pre-processing, Data analysis, Writing - Original Draft, Writing - Review & Editing, Visualization; Lisa Hasenbein: Data collection, Methodology, Writing - Original Draft, Writing - Review & Editing; Efe Bozkir: Methodology, Writing - Review & Editing; Richard Göllner: Methodology, Writing - Review & Editing, Supervision, Funding acquisition; Enkelejda Kasneci: Methodology, Writing - Review & Editing, Supervision.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was partly supported by a grant to Richard Göllner funded by the Ministry of Science, Research and the Arts of the state of Baden-Württemberg and the University of Tübingen as part of the Promotion Program of Junior Researchers. Lisa Hasenbein was a doctoral candidate and supported by the LEAD Graduate School and Research Network, which is funded by the Ministry of Science, Research and the Arts of the state of Baden-Württemberg within the framework of the sustainability funding for the projects of the Excellence Initiative II. Hong Gao is a doctoral candidate and is supported by the University of Tübingen. This research was also partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC number 2064/1 - Project number 390727645.

Declarations

Conflicts of interest The authors have no conflicts of interest to disclose.

References

Abdi Sargezeh, B., Tavakoli, N., & Daliri, M. R. (2019). Gender-based eye movement differences in passive indoor picture viewing: An eye-tracking study. *Physiology & Behavior, 206*, 43–50. https://doi.org/10.1016/j.physbeh.2019.03.023

Agtzidis, I., Startsev, M., & Dorr, M. (2019). 360-degree video gaze behaviour: A ground-truth data set and a classification algorithm for eye movements. In: *Proceedings of the 27th ACM International Conference on Multimedia* (pp. 1007–1015). ACM: New York, MM '19. https://doi.org/10.1145/3343031.3350947

Al Zaidawi, S.M.K., Prinzler, M.H., Schröder, C., et al. (2020). Gender classification of prepubescent children via eye movements with reading stimuli. In: *Companion Publication of the 2020 International Conference on Multimodal Interaction* (pp. 1–6). ACM: New York, ICMI '20 Companion https://doi.org/10.1145/3395035.3425261

Angeli, C., & Valanides, N. (2020). Developing young children's computational thinking with educational robotics: An interaction effect between gender and scaffolding strategy. *Computers in Human Behavior, 105*(105), 954. https://doi.org/10.1016/j.chb.2019.03.018

Appel, T., Scharinger, C., Gerjets, P., et al. (2018). Cross-subject workload classification using pupil-related measures. In: *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. ACM: New York, ETRA '18. https://doi.org/10.1145/3204493.3204531

Appel, T., Sevcenko, N., Wortha, F., et al. (2019). Predicting cognitive load in an emergency simulation based on behavioral and physiological measures. In: *2019 International Conference on Multimodal Interaction* (pp. 154–163). ACM: New York, ICMI '19. https://doi.org/10.1145/3340555.3353735

Ashraf, H., Sodergren, M. H., Merali, N., et al. (2018). Eye-tracking technology in medical education: A systematic review. *Medical Teacher, 40*(1), 62–69. https://doi.org/10.1080/0142159X.2017.1391373

Atmatzidou, S., & Demetriadis, S. (2016). Advancing students' computational thinking skills through educational robotics: A study on age and gender relevant differences. *Robotics and Autonomous Systems, 75*, 661–670. https://doi.org/10.1016/j.robot.2015.10.008

Baser, M. (2013). Attitude, gender and achievement in computer programming. *Middle East Journal of Scientific Research, 14*, 248–255.

Bell, T., Andreae, P., & Robins, A. (2014). A case study of the introduction of computer science in NZ schools. *ACM Trans Comput Educ, 14*(2). https://doi.org/10.1145/2602485

Berkovsky, S., Taib, R., Koprinska, I., et al. (2019). Detecting personality traits using eye-tracking data. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). ACM: New York, CHI '19. https://doi.org/10.1145/3290605.3300451

Bozkir, E., Geisler, D., & Kasneci, E. (2019). Assessment of driver attention during a safety critical situation in VR to generate VR-based training. In: *ACM Symposium on Applied Perception 2019*. ACM: New York. https://doi.org/10.1145/3343036.3343138

Bozkir, E., Günlü, O., Fuhl, W., et al. (2021). Differential privacy for eye tracking with temporal correlations. *Plos One, 16*(8), 1–22. https://doi.org/10.1371/journal.pone.0255979

Bozkir, E., Stark, P., Gao, H., et al. (2021b). Exploiting object-of-interest information to understand attention in VR classrooms. In: *2021 IEEE Virtual Reality and 3D User Interfaces (VR)* (pp. 597–605). IEEE: New York. https://doi.org/10.1109/VR50410.2021.00085

Bulling, A., Ward, J. A., Gellersen, H., et al. (2011). Eye movement analysis for activity recognition using electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(4), 741–753. https://doi.org/10.1109/TPAMI.2010.86

Castner, N., Appel, T., Eder, T., et al. (2020). Pupil diameter differentiates expertise in dental radiography visual search. *Plos One, 15*(5), 1–19. https://doi.org/10.1371/journal.pone.0223941

Casu, A., Spano, L.D., Sorrentino, F., et al. (2015). RiftArt: Bringing masterpieces in the classroom through immersive virtual reality. In: *Smart Tools and Apps for Graphics - Eurographics Italian Chapter Conference. The Eurographics Association*, Geneva, Switzerland, pp 77–84. https://doi.org/10.2312/stag.20151294

Chalmers, C. (2018). Robotics and computational thinking in primary school. *International Journal of Child-Computer Interaction, 17*, 93–100. https://doi.org/10.1016/j.ijcci.2018.06.005

Chien, K. P., Tsai, C. Y., Chen, H. L., et al. (2015). Learning differences and eye fixation patterns in virtual and physical science laboratories. *Computers & Education, 82*, 191–201. https://doi.org/10.1016/j.compedu.2014.11.023

Cryer, A., Kapellmann-Zafra, G., Abrego-Hernández, S., et al. (2019). Advantages of virtual reality in the teaching and training of radiation protection during interventions in harsh environments. In:*2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)* (pp. 784–789). IEEE, New York. https://doi.org/10.1109/ETFA.2019.8869433

Datta, A., Sen, S., & Zick, Y. (2016) Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In: *2016 IEEE Symposium on Security and Privacy (SP)* (pp. 598–617). https://doi.org/10.1109/SP.2016.42

Dumais, S.T., Buscher, G., & Cutrell, E. (2010) Individual differences in gaze patterns for web search. In: *Proceedings of the third symposium on Information interaction in context* (pp. 185–194). ACM, New York, IIiX '10. https://doi.org/10.1145/1840784.1840812

Eivazi S, & Bednarik R (2011) Predicting problem-solving behavior and performance levels from visual attention data. In: *Proceedings of 2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction at IUI* (pp. 9–16). ACM, New York

Gao H, Bozkir E, Hasenbein L, et al. (2021) Digital transformations of classrooms in virtual reality. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, CHI '21 https://doi.org/10.1145/3411764.3445596

García-Peñalvo, F. J., & Mendes, A. J. (2018). Exploring the computational thinking effects in pre-university education. *Computers in Human Behavior, 80*, 407–411. https://doi.org/10.1016/j.chb.2017.12.005

Grodotzki, J., Ortelt, T. R., & Tekkaya, A. E. (2018). Remote and virtual labs for engineering education 4.0: Achievements of the ELLI project at the TU dortmund university. *Procedia Manufacturing, 26*, 1349–1360. https://doi.org/10.1016/j.promfg.2018.07.126,46th SME North American Manufacturing Research Conference, NAMRC 46, Texas, USA

Hernández-de Menéndez, M., Guevara, A. V., & Morales-Menendez, R. (2019). Virtual reality laboratories: A review of experiences. *International Journal on Interactive Design and Manufacturing (IJIDeM), 13*(3), 947–966. https://doi.org/10.1007/s12008-019-00558-7

Hirt, C., Eckard, M., & Kunz, A. (2020). Stress generation and non-intrusive measurement in virtual environments using eye tracking. *Journal of Ambient Intelligence and Humanized Computing, 11*(1), 1–13. https://doi.org/10.1007/s12652-020-01845-y

Holmqvist, K., Nyström, M., Andersson, R., et al. (2011). Eye tracking: A comprehensive guide to methods and measures. OUP Oxford

Hoppe, S., Loetscher, T., Morey, S. A., et al. (2018). Eye movements during everyday behavior predict personality traits. *Frontiers in Human Neuroscience, 12*, 105. https://doi.org/10.3389/fnhum.2018.00105

Hsu, T. C., Chang, S. C., & Hung, Y. T. (2018). How to learn and how to teach computational thinking: Suggestions based on a review of the literature. *Computers & Education, 126*, 296–310. https://doi.org/10.1016/j.compedu.2018.07.004

Hwang, Y. M., & Lee, K. C. (2018). Using an eye-tracking approach to explore gender differences in visual attention and shopping attitudes in an online shopping environment. *International Journal of Human-Computer Interaction, 34*(1), 15–24. https://doi.org/10.1080/10447318.2017.1314611

Kasneci, E., Kasneci, G., Trautwein, U., et al. (2022). Do your eye movements reveal your performance on an iq test? A study linking eye movements and socio-demographic information to fluid intelligence. *Plos One, 17*(3), 1–23. https://doi.org/10.1371/journal.pone.0264316

Kong, S. C., Chiu, M. M., & Lai, M. (2018). A study of primary school students' interest, collaboration attitude, and programming empowerment in computational thinking education. *Computers & Education, 127*, 178–189. https://doi.org/10.1016/j.compedu.2018.08.026

Liao, H., & Dong, W. (2017). An exploratory study investigating gender effects on using 3D maps for spatial orientation in wayfinding. *ISPRS International Journal of Geo-Information, 6*(3), 60. https://doi.org/10.3390/ijgi6030060

Lin, F., Wu, Y., Zhuang, Y., et al. (2016). Human gender classification: A review. *International Journal of Biometrics, 8*(3–4), 275–300. https://doi.org/10.1504/IJBM.2016.082604

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 4768–4777). Curran Associates Inc., Red Hook, NIPS'17

Lundberg, S. M., Erion, G., Chen, H., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Mathôt, S., Fabius, J., Van Heusden, E., et al. (2018). Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior Research Methods, 50*(1), 94–106. https://doi.org/10.3758/s13428-017-1007-2

McGuire, L., Mulvey, K. L., Goff, E., et al. (2020). STEM gender stereotypes from early childhood through adolescence at informal science centers. *Journal of Applied Developmental Psychology, 67*(101), 109. https://doi.org/10.1016/j.appdev.2020.101109

Mercer Moss, F. J., Baddeley, R., & Canagarajah, N. (2012). Eye movements to natural images as a function of sex and personality. *Plos One, 7*(11), 1–9. https://doi.org/10.1371/journal.pone.0047870

Molina, A. I., Óscar, N., Ortega, M., et al. (2018). Evaluating multimedia learning materials in primary education using eye tracking. *Computer Standards & Interfaces, 59*, 45–60. https://doi.org/10.1016/j.csi.2018.02.004

Negi, S., & Mitra, R. (2020). Fixation duration and the learning process: An eye tracking study with subtitled videos. *Journal of Eye Movement Research, 13*(6)

Nourbakhsh, I., Hamner, E., Crowley, K., et al. (2004). Formal measures of learning in a secondary school mobile robotics course. In: *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004* (Vol 2. pp. 1831–1836) IEEE, New York https://doi.org/10.1109/ROBOT.2004.1308090

Obaidellah, U., & Haek, M. A. (2018). Evaluating gender difference on algorithmic problems using eye-tracker. In: *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. ACM, New York,ETRA '18 https://doi.org/10.1145/3204493.3204537

Papavlasopoulou, S., Sharma, K., & Giannakos, M. N. (2020). Coding activities for children: Coupling eye-tracking with qualitative data to investigate gender differences. *Computers in Human Behavior, 105*(105), 939. https://doi.org/10.1016/j.chb.2019.03.003

Polat E, Hopcan S, Kucuk S, et al. (2021) A comprehensive assessment of secondary school students' computational thinking skills. *British Journal of Educational Technology, 52*(5) https://doi.org/10.1111/bjet.13092

Raptis, G. E., Fidas, C. A., & Avouris, N. M. (2017). On implicit elicitation of cognitive strategies using gaze transition entropies in pattern recognition tasks. In: *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 1993–2000). ACM, New York, CHI EA '17. https://doi.org/10.1145/3027063.3053106

Reilly, D., Neumann, D. L., & Andrews, G. (2017). Gender differences in spatial ability: Implications for STEM education and approaches to reducing the gender gap for parents and educators (pp. 195–224). Springer, Berlin. https://doi.org/10.1007/978-3-319-44385-0_10

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In: *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications* (pp. 71–78). Association for Computing Machinery, New York, ETRA '00 https://doi.org/10.1145/355017.355028

Sammaknejad, N., Pouretemad, H., Eslahchi, C., et al. (2017). Gender classification based on eye movements: A processing effect during passive face viewing. *Advances in Cognitive Psychology, 13*(3), 232–240. https://doi.org/10.5709/acp-0223-1

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry, 36*, 1627–1639. https://doi.org/10.1021/ac60214a047

Sentance, S., & Csizmadia, A. (2015) Teachers' perspectives on successful strategies for teaching computing in school. In: *IFIP TC3 Working Conference 2015: A New Culture of Learning: Computing and Next Generations*. Vilnius, Lithuania

Seo, S. H., Kim, E., Mundy, P., et al. (2019). Joint attention virtual classroom: A preliminary study. *Psychiatry Investigation, 16*, 292–299. https://doi.org/10.30773/pi.2019.02.08

Seow, P., Looi, C. K., How, M. L., et al. (2019). Educational Policy and Implementation of Computational Thinking and Programming: Case Study of Singapore (pp. 345–361). Springer Singapore https://doi.org/10.1007/978-981-13-6528-7_19

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games, 2*, 307–317.

Steil, J., Hagestedt, I., Huang, M. X., et al. (2019). Privacy-aware eye tracking using differential privacy. In: *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. ACM, New York, ETRA '19. https://doi.org/10.1145/3314111.3319915

Strohmaier, A. R., MacKay, K. J., Obersteiner, A., et al. (2020). Eye-tracking methodology in mathematics education research: A systematic literature review. *Educational Studies in Mathematics, 104*(2), 147–200. https://doi.org/10.1007/s10649-020-09948-1

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems, 41*(3), 647–665. https://doi.org/10.1007/s10115-013-0679-x

Sullivan, A., & Bers, M. (2016). Girls, boys, and bots: Gender differences in young children's performance on robotics and programming tasks. *Journal of Information Technology Education : Innovations in Practice, 15*, 145–165. https://doi.org/10.28945/3547

Sundararajan, M., & Najmi, A. (2020). The many shapley values for model explanation. In: I.I.I. HD, A. Singh (eds) *Proceedings of the 37th International Conference on Machine Learning, Proceedings of Machine Learning Research* (Vol. 119. pp 9269–9278). PMLR

Wang, M. T., & Degol, J. L. (2017). Gender gap in science, technology, engineering, and mathematics (STEM): Current knowledge, implications for practice, policy, and future directions. *Educational Psychology Review, 29*(1), 119–140. https://doi.org/10.1007/s10648-015-9355-x

Wing, J. (2011). Research notebook: Computational thinking–what and why. *The Link Magazine, 6*, 20–23.

Yoshida, Y., Ohwada, H., Mizoguchi, F., et al. (2014). Classifying cognitive load and driving situation with machine learning. *International Journal of Machine Learning and Computing, 4*, 210–215. https://doi.org/10.7763/IJMLC.2014.V4.414

Zhou, F., Yang, X. J., & de Winter, J. C. F. (2021). Using eye-tracking data to predict situation awareness in real time during takeover transitions in conditionally automated driving. *IEEE Transactions on Intelligent Transportation Systems* (pp. 1–12) https://doi.org/10.1109/TITS.2021.3069776

Zhou, J., Sun, J., Chen, F., et al. (2015). Measurable decision making with GSR and pupillary analysis for intelligent user interface. *ACM Trans Comput-Hum Interact, 21*(6). https://doi.org/10.1145/2687924

Authors and Affiliations

Hong Gao[1] · Lisa Hasenbein[2] · Efe Bozkir[1] · Richard Göllner[2] · Enkelejda Kasneci[3]

Lisa Hasenbein
lisa.hasenbein@uni-tuebingen.de

Efe Bozkir
efe.bozkir@uni-tuebingen.de

Richard Göllner
richard.goellner@uni-tuebingen.de

Enkelejda Kasneci
enkelejda.kasneci@tum.de

[1] University of Tübingen, Tübingen, Germany

[2] Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Germany, Tübingen

[3] Technical University of Munich, Munich, Germany