IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 August 2023, accepted 31 August 2023, date of publication 5 September 2023, date of current version 8 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3312216

RESEARCH ARTICLE

High-Frequency Cybersickness Prediction Using Deep Learning Techniques With Eye-Related Indices

SHOGO SHIMADA[1], PEERAWAT PANNATTEE[1], YASUSHI IKEI[2], (Member, IEEE), NOBUYUKI NISHIUCHI[1], (Member, IEEE), AND VIBOL YEM[3], (Member, IEEE)

[1]Graduate School of Systems Design, Tokyo Metropolitan University, Tokyo 191-0065, Japan
[2]Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8656, Japan
[3]Graduate School of Science and Technology, Degree Program in Systems and Information Engineering, University of Tsukuba, Ibaraki 305-8573, Japan

Corresponding author: Nobuyuki Nishiuchi (nnishiuc@tmu.ac.jp)

This work was partially supported by Japan Society for Promotion of Science Grant-in-Aid for Scientific Research (JSPS KAKENHI) JP20K12511, JP18H04118, and Local-5G Project of Tokyo Metropolitan University.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Research Ethics Committee of Tokyo Metropolitan University under Approval No. H22-029.

ABSTRACT Cybersickness is a growing concern in the field of virtual reality (VR). It is characterized by symptoms, such as headache, sweating, disorientation, and nausea. These symptoms can considerably hinder the users' immersive experience in VR environments, leading to a pressing need for effective solutions to combat cybersickness. In this study, we aim to tackle cybersickness by presenting a novel high-frequency approach for detecting the timing at which users experience cybersickness. Our approach uses 1-, 5-, or 10-s time-series eye-related indices processed by deep learning algorithms to predict cybersickness severity. In five-fold cross-validation, we achieved 71.09% accuracy in classifying four classes of cybersickness severity when individuals were not distinguished. Furthermore, with individualized cross-validation, we achieved an accuracy of up to approximately 80%. Our approach outperforms other cybersickness prediction studies as it provides the highest frequency in predicting cybersickness. It is anticipated that our approach will be valuable not only for immediate evaluation by researchers investigating cybersickness mitigation but also for early detection and notification of users experiencing cybersickness symptoms. By predicting cybersickness, our approach has the potential to promote the future advancement of VR technology.

INDEX TERMS Cybersickness, deep learning, eye-related indices, high-frequency prediction, virtual reality.

I. INTRODUCTION

As we enter the digital age, virtual reality (VR) technology, once perceived only in science fiction, is gradually permeating our daily lives. However, many VR experiences can cause discomforting symptoms such as eye strain, headache, sweating, disorientation, and nausea, which resemble those of motion sickness [1]. These side effects are generally referred to as ''cybersickness (CS)'' or ''VR sickness,'' and they persist in the general use of VR devices such as head-mounted displays (HMDs), hindering immersion [2]. Once a user experiences CS, it becomes a psychological barrier for he/she to wear an HMD again. This is a serious problem. Studies have shown that more than half of the users have experienced CS in VR environments [3], [4]. Thus, CS is a major obstacle in the advancement of VR technology and deserves further study on methods for suppressing or preventing it during VR immersion.

It is crucial to develop a method for detecting CS in advance to confirm the effectiveness of countermeasures

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

against it. High-frequency CS detection techniques can contribute to the development of automatic CS suppression technology during VR experiences [5] and enable immediate evaluation of the experience. For instance, the VR content can be programmatically changed from bodily active motion to a relaxed experience when CS is detected in the user. This can prevent an increase in CS severity and overcome the psychological barrier of users. Therefore, a model that can predict the CS severity of users with high frequency and high accuracy is required.

In this study, we considered a method of predicting the occurrence of CS with high frequency using deep learning (DL) techniques by regarding time-series eye-related indices as features. Eye movements are the manifestations of the human visual system and exhibit temporal variations. Therefore, we posit that capturing the significant attributes of these temporal changes can be effectively achieved by treating them, as time-series data. Currently, eye-related indices data can be simply collected from sensors built into HMDs, which efficiently contribute to the development of VR technologies.

Our study offers several valuable contributions to the field of VR technology, primarily aiming at the capability to predict CS severity at a higher frequency than previous studies. We described the potential of employing time-series eye-related indices and DL models to achieve this objective. Furthermore, we demonstrated the feasibility of training DL models for each individual to achieve individual-specific CS prediction, which is crucial in addressing individual differences and enhancing users' overall VR experience. These findings suggest that our approach is promising in addressing CS challenges in the development of VR technology.

II. RELATED WORK

This section discusses related studies and is divided into three subsections. Each subsection provides a comprehensive discussion of the current state and challenges in CS prediction, focusing on the relationship between physiological and eye-related indices and CS, the application of machine learning (ML) and DL for CS prediction, and the aspect of high-frequency prediction.

A. PHYSIOLOGICAL INDICES AND CYBERSICKNESS PREDICTION

Numerous preceding studies have unveiled noteworthy associations between physiological signals, including gastric tachyarrhythmia [4], electroencephalogram (EEG) signals [4], [6], [7], [8], [9], heart rate (HR) [4], [7], [10], [11], breathing rate (BR) [7], [12], galvanic skin response (GSR) [7], [13], and CS, as well as simulator sickness and motion sickness. These observations underscore that alterations in the operations of the central and autonomic nervous systems accompany CS.

Furthermore, there are some studies that have been conducted using ML/DL to predict the occurrence and severity of CS based on objective data from multiple physiological indices [6], [7], [14], [15], [16], [17], [18]. Kim et al. [6] used EEG data obtained from over 200 subjects as a feature and worked on the classification tasks of 5-level CS severity obtained as subjective evaluations at the end of each content. The total data was divided into 80% train data, 10% validation data, and 10% test data, and learning was performed using convolutional neural networks (CNNs) and long short-term memory (LSTM). As a result, a maximum test accuracy of 89.16% (standard deviation (SD) = 1.87) was achieved. Islam et al. [7] collected HR, BR, and GSR data as features from 31 participants and predicted the current and 2-min future severity of CS using a support vector machine (SVM), CNNs, and LSTM. The fast motion sickness scale (FMS) [19] was used for CS evaluation and was divided into three severity classes based on the distribution of evaluations. As a result, their proposed CNN-LSTM classifier model achieved an accuracy of 97.44% for predicting current CS severity and 87.38% for predicting future CS severity. Garcia-Agundez et al. [14] collected electrocardiogram (ECG), EEG, respiratory data, skin conductivity data, and relevant game parameters, such as avatar linear and angular speed, acceleration, head movements, and on-screen collisions, from 66 participants, and predicted the severity of CS using SVM, K-nearest neighbors, and artificial neural networks. The simulator sickness questionnaire (SSQ) [20] score obtained before and after the experiment was used for CS evaluation. A maximum classification accuracy of 82% was achieved for binary classification and 56% for ternary classification.

Although the aforementioned studies demonstrated the effectiveness of using physiological indices to evaluate CS, the use of external sensors may lead to problems, such as restricting HMD users movement and interaction in a VR environment.

B. EYE-RELATED INDICES AND CYBERSICKNESS PREDICTION

Considering the problems outlined above arising from the use of external sensors, several recent studies have investigated the relationship between eye-related indices and CS [21], [22], [23].

Lopes et al. [21] demonstrated that the pupil position and eye blink pattern were substantially different between the sickness and non-sickness groups. Participants with the sick condition had a higher blink rate and count, and the data of the sickness group were considerably smaller in terms of the spread of the distribution than those of the non-sickness group. Chang et al. [22] described that the fixation time and the distance between the eye gaze and the object-position sequence are highly correlated with CS. When viewing a roller coaster video in a VR space, participants who gazed further away from the track tended to demonstrate a lower level of CS. According to Nam et al. [23], the varying pattern of CS was reflected in the center gaze ratio and scan-path length. Besides, other studies suggest that pupil diameter [24] and optokinetic-after-nystagmus [25] are correlated with CS.

Some studies have predicted the severity of CS using ML/DL models with eye-related indices as objective data. Islam et al. [26] used eye and head tracking and stereo-image data from 30 participants to classify four levels of CS severity. The evaluation of CS severity was based on self-reported evaluations based on FMS collected every 30 s during VR gameplay. As a result, their proposed deep fusion approach achieved an accuracy of 87.7% for predicting CS. When only the eye-related indices (pupil diameter, gaze direction, and convergence distance) were used as features, an accuracy of 80.7% was achieved. Chang et al. [22] developed a regression model to predict the severity of CS using eye-related indices obtained from 26 participants. The SSQ score obtained before and after the experiment was used for CS evaluation. As a result, their model could explain 34.8% of the total variance of CS, indicating a substantial improvement over the study performed by Wibirama et al. [27], which could explain only 4.2% of the total variance.

These studies have demonstrated the relationship between eye-related indices and CS utilizing diverse methods of employing ML/DL models to predict CS and yield promising results. Inspired by these findings, we also attempted to predict CS using eye-related indices and DL techniques.

C. HIGH-FREQUENCY CYBERSICKNESS PREDICTION

The above previous studies have demonstrated the effectiveness of methods for predicting CS using ML/DL. However, most of the measures used for CS severity in these studies were post-hoc evaluations using the SSQ or FMS. There are very few studies that consider high-frequency CS prediction during VR immersion. There are studies (Islam et al. [26]) that evaluate CS in VR immersion; however, their approach requires at least 30 s of historical data. Based on this background, our study considers approaches for predicting the occurrence and severity of CS employing eye-related indices with higher frequency for early CS detection. By considering eye-related indices along with exploring high-frequency prediction methods, our study aims to advance the CS prediction field in a VR environment.

III. EXPERIMENTAL DESIGN

A. PARTICIPANTS

Thirty participants (26 males and 4 females) aged between 21 and 39 years (mean age = 23.57; SD = 4.26) were recruited for the study. One participant who experienced severe CS symptoms could not complete all experimental tests. None of the participants suffered from vertigo, epilepsy, or any other condition that could be aggravated by wearing an immersive HMD. All participants had a normal naked-eye or corrected-to-normal vision with contact lenses. Written informed consent was obtained from all participants prior to the experiment. The participants could terminate the experiment at any time. The study was approved by the Research Ethics Committee of Tokyo Metropolitan University.

B. MATERIALS

The VR headset used in this study was the HTC Vive Pro Eye with Tobii® eye-tracking technology. The maximum sampling frequency of eye-related indices is 120 Hz. However, a flicker problem had occurred when obtaining eye-related indices at this frequency; thus, the eye-tracking frequency was reduced to 50 Hz. The HMD screen is a dual organic light-emitting diode screen with a resolution of 1440 × 1600 pixels per eye. The field of viewing angle is 110° with a refresh rate of 90 Hz. Audio can be played through the integrated Vive headphones. The personal computer (PC) used was equipped with an Intel Core i7-11800H CPU running at 2.30 GHz, 16 GB RAM, and an NVIDIA GeForce RTX 3070 GPU.

As CS-inducing scenes, we selected a car video and a roller coaster video. The car video included movements such as sudden deceleration, sudden reversals, sudden acceleration, swerving, and meandering driving. A VR camera was installed in the front passenger seat of a left-hand-drive car to provide a clear view of the surrounding scenery (Fig. 1(a)). The roller coaster video included rotations and diagonal movements of the cart. A VR camera was placed on the first cart of the train, giving a user a front-line experience (Fig. 1(b)). In the other video (Fig. 1(c)), hereafter called the non-sickness video, a user was ''seated'' beside a quiet lake, which was intended to induce a calm, relaxed sensation. No sickness-reducing technology was applied to these VR videos. All experiments were conducted through Steam VR and Unity 3D, and eye-related indices were collected using the VIVE Eye and Facial Tracking SDK (SRanipal SDK) provided by HTC.

C. PROCEDURE

In this experiment, we presented two conditions. Each condition involved only one of the sickness videos (either car or roller coaster video) to induce CS in the participants. We deemed it unethical to have participants view both sickness videos in a single condition. For transparency and consistency, all participants experienced both conditions, which consisted of the following sequences:

- Condition 1: Non-sickness video → Car video → Non-sickness video
- Condition 2: Non-sickness video → Roller coaster video → Non-sickness video

As illustrated in Fig. 2, each sickness video was preceded and followed by a non-sickness video. The sickness videos were approximately 5-min long, while the non-sickness video was 1-min long. Participants sat on a chair and viewed a combination of the prepared videos. After viewing one set of scenes, they were allowed to rest until their CS symptoms subsided before experiencing the other condition (Fig. 2). To reduce the order effect, the two conditions were presented in a counterbalanced way across the participants.

In all videos, the participants could freely change their field-of-view. Eye tracking was also calibrated before the start of the experiment. During the video experience, the

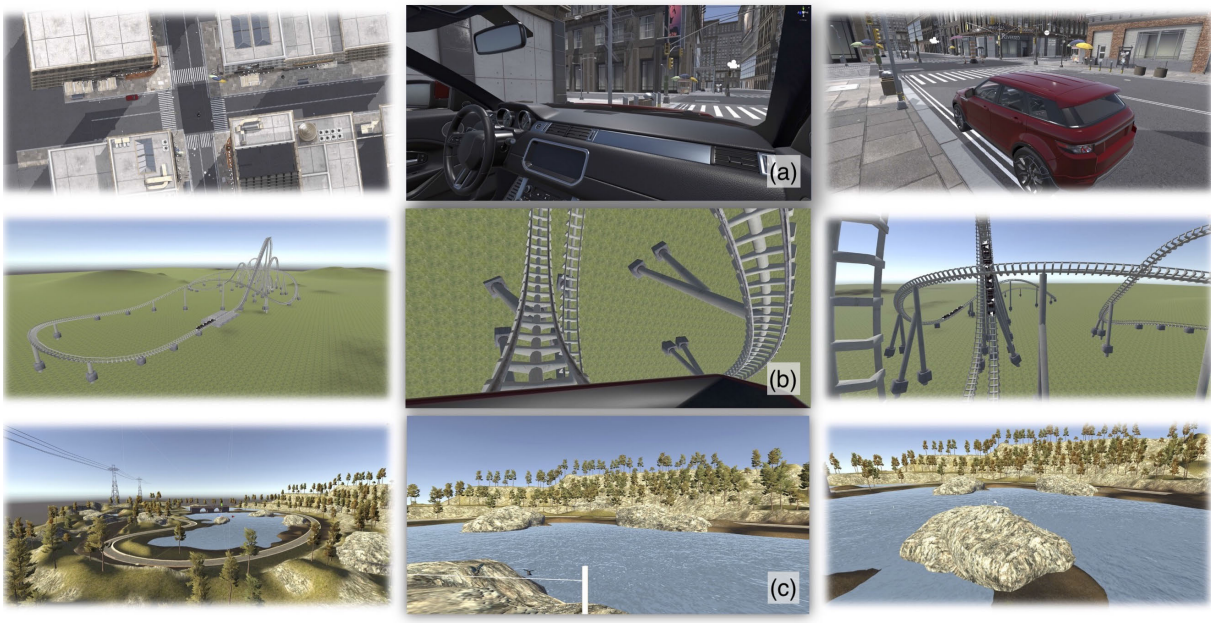


FIGURE 1. Screenshots of three experimental scenes in virtual environment: (a) car video, (b) roller coaster video, and (c) non-sickness video. The left image shows the appearance of the experimental environment, the middle image shows the scene from the perspective of the subject, and the right image is a single shot that captures the environment around the subject.

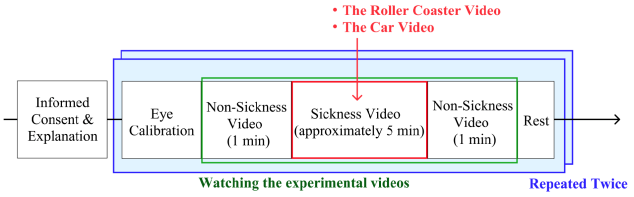


FIGURE 2. Procedure of the user data measurement.

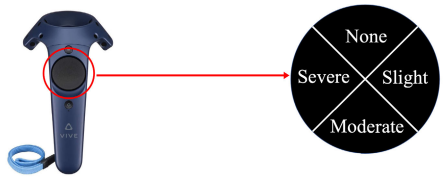


FIGURE 3. Illustration of the coordinate-segmentation method using the HTC Vive controller trackpad for 4-level CS evaluation.

participants subjectively indicated one of the four levels of CS severity by placing their thumb at the appropriate position on the trackpad of the HTC Vive controller (2018), as depicted in Fig. 3. The trackpad is equipped with a sensor that allows continuous recording by simply placing a finger on it (no forceful pushing or clicking was required). Analogous to the SSQ crafted as a subjective assessment tool for simulator sickness, we assessed four degrees of CS severity: None, Slight, Moderate, and Severe. Data on CS severity were recorded continuously at a sampling rate of 50 Hz. Each participant completed the entire experiment in approximately 1 h.

D. EYE-RELATED INDICES

As part of this study, we examined various eye-related indices to assess their potential as indicators for gauging the severity of CS. Among these indices, pupil diameter emerged as a notable candidate. Fluctuations in pupil diameter are widely recognized to be influenced by a range of factors, including emotional states [28], [29], and fatigue [30]–conditions frequently associated with CS experiences [1], [31], [32]. Therefore, variations in pupil diameter provide significant insights in this context. Nonetheless, we have taken into consideration that pupil diameter constitutes just among several contributing factors. Our approach comprehensively integrates an array of eye-related indices to aptly predict the severity of CS. We collected a total of 11 features from five types of eye-related indices as follows:

- Normalized pupil positions of both eyes (four features of x and y axes for both eyes)
- Pupil diameters of both eyes [mm] (two features)
- Gaze deviation from the center of the screen (one feature)
- Angular velocity of eye gaze at 0.02-s intervals [deg/s] (one feature)
- Gaze origin position [mm] (three features of x, y, and z axes)

As shown in Fig. 4 (right), the gaze deviation from the center of screen d_{xy} is defined in the HTC Vive Pro Eye coordinate system utilizing the x-direction gaze vector x_b, the y-direction gaze vector y_b, and the center of screen coordinates $(x_o, y_o) = (0,0)$:

$$d_{xy} = \sqrt{(x_b - x_o)^2 + (y_b - y_o)^2} \tag{1}$$

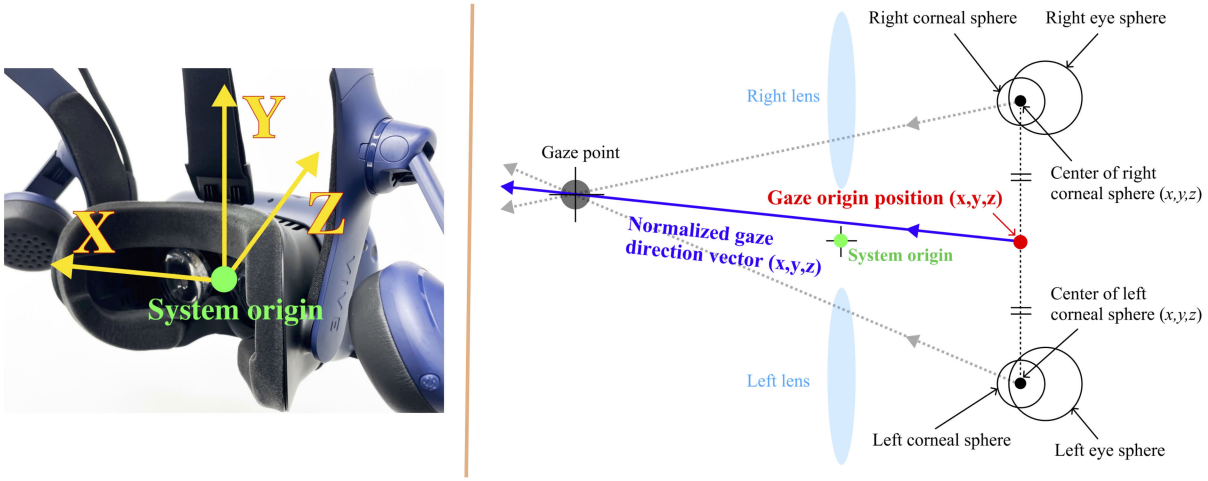


FIGURE 4. (Left) Coordinate system of HTC Vive Pro Eye, and (right) a diagram showing gaze origin position and normalized gaze direction vector.

The center of screen coordinates refers to the system origin situated at the center of the two lenses of the HMD (Fig. 4), while the gaze origin position represents the central coordinate of the straight line connecting the central coordinates of both corneas (Fig. 4 (right)). The pupil positions of both eyes and the gaze direction vector are automatically provided as normalized values according to the HTC Vive Pro Eye specifications. All data were recorded in the right-handed coordinate system (Fig. 4 (left)).

These eye-related indices were then individually normalized using the following equation for subsequent DL applications:

$$x_i^{norm} = \frac{x_i^{raw} - \mu}{\sigma} \tag{2}$$

Here, x represents time-series data, μ represents the mean of x, σ represents the standard deviation of x, x_i^{raw} represents the i_{th} data point of x, and x_i^{norm} represents the data after normalization. We adopted the methodology of individual normalization and integrated learning processes that accommodate temporal characteristics. This approach was used to alleviate the influence of static factors such as gender and age on eye-related indices [33]. The process of normalization has enabled us to mitigate these effects and regard them as inherent noise within the measured data. A noteworthy and pragmatic advantage of this methodology is its ability to obviate the necessity for distinguishing among VR participants, thereby simplifying the implementation process.

Furthermore, missing intervals in eye-related indices due to factors such as blinking were interpolated using a linear interpolation method based on data with and without missing values. Overall, we obtained approximately 42,000 data points for each participant who completed the experiment, comprising eye-related indices and CS severity data based on a sampling rate of 50 Hz for the full VR viewing time of 14 min per participant.

IV. DEEP LEARNING MODEL

In this study, we aimed to achieve highly accurate prediction of CS occurrence and severity utilizing the attention-based long short-term memory fully convolutional network (ALSTM-FCN)-based DL model proposed by Karim et al. [34]. The ALSTM-FCN model is a variation of the LSTM-FCN model that incorporates an attention mechanism. Besides, classification methods based on FCN and LSTM-FCN models are more accurate than conventional methods [34], [35].

A. TEMPORAL CONVOLUTIONAL NETWORKS (TCN)

We extracted eye-related indices as features using temporal convolutional networks (TCNs) in FCNs. A TCN is a CNN variant for sequence modeling that uses time-series data as input. As stated in Lea et al. [36], let $X_t \in \mathbb{R}^{F_0}$ be the input feature vector of length F_0 in time step t for $1 \leq t \leq T$. Each sequence may have a specific time T, and the number of time steps in each layer l is denoted as T_l. The true label for each frame is given by $y_t \in \{1, ..., C\}$, where C represents the number of classes.

For each convolutional layer, we applied a set of one-dimensional filters that capture changes in input signals. The filters for each layer l are parameterized by tensor $W^{(l)} \in \mathbb{R}^{F_l \times d \times F_{l-1}}$ and biases $b^l \in \mathbb{R}^{F_l}$, where d represents the filter duration. In the same layer, the i_{th} component of the unnormalized activation $\widehat{E}_t^{(l)} \in \mathbb{R}^{F_l}$ is a function of the incoming normalized activation matrix $E^{(l-1)} \in \mathbb{R}^{F_{l-1} \times T_{l-1}}$ from the previous layer [36]

$$\widehat{E}_{i,t}^{(l)} = f\left(b_i^{(l)} + \sum_{t'=1}^{d} \left\langle W_{i,t',.}^{(l)}, E_{.t+d-t'}^{(l-1)} \right\rangle\right) \tag{3}$$

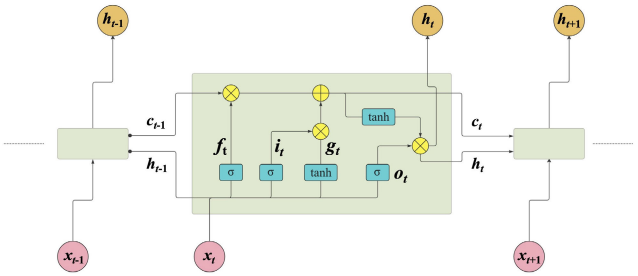


FIGURE 5. Schematic of the internal structure of the LSTM unit.

for each time t, where $f(\cdot)$ denotes the rectified linear unit (ReLU) function.

A basic convolution block consists of a convolution layer, followed by a normalization process and the ReLU activation function. Overall, the FCN model is composed of three convolutional blocks and a global average pooling (GAP) layer [37] applied after the final convolutional block.

B. LONG SHORT-TERM MEMORY (LSTM)

LSTM is a type of recurrent neural network (RNN) architecture designed to handle the problem of vanishing gradients in traditional RNNs [38]. LSTM has an internal memory cell that can retain information for longer periods, allowing the network to better handle time-dependent sequences. Fig. 5 shows a schematic of the LSTM unit. The LSTM unit is composed of three gates and a memory cell to store long-term information: which input gate i_t, forget gate f_t, output gate o_t, and memory cell c_t, respectively. The calculation process of LSTM is outlined below. Here, x represents the input to the unit, h represents the output of the unit, b represents the biases, W represents the weights, and $\odot$ denotes the element-wise product. In addition, the subscripts of weights (W) and biases (b) indicate which weight or bias is by the first and second letters.

The input gate i_t and the input data g_t are calculated using two activation functions (σ: sigmoid function; *tanh*: hyperbolic tangent function), as expressed in Eqs. (4) and (5).

$$i_t = \sigma(W_{xi}x_t + b_{xi} + W_{hi}h_{t-1} + b_{hi}) \tag{4}$$
$$g_t = tanh(W_{xg}x_t + W_{hg}h_{t-1} + b_g) \tag{5}$$

The forget gate f_t decides which information to preserve and which to discard, and it is controlled by the sigmoid function, as expressed in Eq. (6).

$$f_t = \sigma(W_{xf}x_t + b_{xf} + W_{hf}h_{t-1} + b_{hf}) \tag{6}$$

The state of the memory cell c_t at time step t is given by Eq. (7), and it is the sum of the input gate i_t multiplied by the input data g_t and the forget gate f_t multiplied by the state of the memory cell c_{t-1} from the previous time step.

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{7}$$

The output gate o_t and the output h_t are calculated using two activation functions, as expressed in Eqs. (8) and (9).

$$o_t = \sigma(W_{xo}x_t + b_{xo} + W_{ho}h_{t-1} + b_{ho}) \tag{8}$$
$$h_t = o_t \odot tanh(c_t) \tag{9}$$

Note that the input gate i_t and input data g_t of LSTM are different and that the input gate i_t determines the addition of new information to the memory cell. The input gate i_t is a combination of input data x_t and previous output h_{t-1}.

C. ATTENTION MECHANISM

The attention mechanism is a DL technique for extracting important information from multiple input sequences and combining them to generate an output sequence. The attention concept was introduced by Bahdanau et al. [39]. The attention mechanism incorporates contextual information from input sequences. The output element a_i is determined by a sequence of annotations $(h_1, h_2, \ldots, h_n)$, where n represents the maximum length of an input sequence. Each annotation h_i contains information about the entire input sequence, with a strong focus on parts surrounding the i_{th} element of the input sequence. a_i can be calculated using Eq. (10).

$$a_i = \sum_{j=1}^{n} W_{ij}h_j \tag{10}$$

Here, W_{ij} represents the weights of each annotation h_j. The attention weights are calculated using a dot product between the annotations, and the softmax function is used to normalize the results. This can be mathematically represented as follows:

$$W_{ij} = \frac{\exp(h_i^T h_j)}{\sum_{k=1}^{n} \exp(h_i^T h_k)} \tag{11}$$

where h_i and h_j denote the i_{th} and j_{th} annotations, respectively, and T represents the transpose of the vector.

D. ALSTM-FCN MODEL

The architecture of the ALSTM-FCN model consists of two parts: an LSTM network with an attention mechanism and an FCN, as discussed in previous sub-subsections. The ALSTM-FCN model used in this study is depicted in Fig. 6.

The model has two LSTM layers stacked on top of each other. The first LSTM layer has 256 units and is followed by a"Layer Normalization (LN)" layer [40], which normalizes the output of the LSTM layer. The output of this layer is passed to a Dropout layer with a dropout rate of 0.2. This means that during training, 20% of the activations in the output of the first LSTM layer will be set to zero. This is done to prevent overfitting by randomly dropping out some activations. The model is forced to learn multiple independent representations of the same input, enhancing the robustness of the model to unseen data. The second LSTM layer has 128 units, followed by a LN layer, an Attention layer, and a Dropout layer with a dropout rate of 0.2. The Attention layer

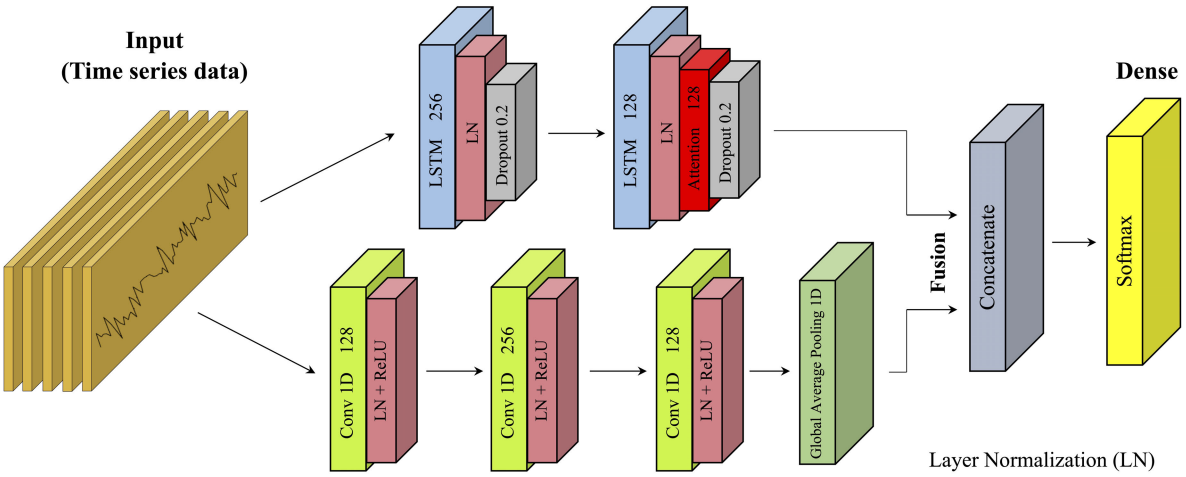


FIGURE 6. Configuration diagram of ALSTM-FCN model used in the experiment.

is used to weigh the importance of each timestep in the LSTM output and generate a weighted sum of the LSTM output.

Similarly, FCN layers are also defined. The FCN layers are defined using a series of one-dimensional convolutional (Conv1D) layers. The first Conv1D layer has 128 filters of size 8, the second Conv1D layer has 256 filters of size 5, and the third Conv1D layer has 128 filters of size 3. Each Conv1D layer is followed by a LN layer and an activation layer with the ReLU activation function. This function is used to introduce nonlinearity into the model and allow the model to learn more complex representations of the data. The output of the last Conv1D layer is passed through a one-dimensional GAP layer. This layer takes the average of the values of the last Conv1D layer along the temporal axis. This is used to reduce the number of parameters in the model and to make the model more robust to variations in input sequences.

The output of the LSTM and FCN layers are concatenated and passed through a Dense layer with the softmax activation. The Dense layer has the same number of units as the number of classes in the output. The final output is a probability distribution over the classes.

V. CYBERSICKNESS PREDICTION USING DEEP LEARNING

A. EVALUATION METHODS

In this analysis, we conducted two evaluations to verify the feasibility of high-frequency CS prediction:

- Five-fold cross-validation using all data obtained from all subjects.
- Five-fold cross-validation on individual data, repeated for all subjects.

The k-fold cross-validation is a method for evaluating the performance of ML models. It involves dividing a dataset into k equally sized folds, training the model on k-1 of the folds, and evaluating it on the remaining one. This process is repeated k times, with each fold serving as the test set once. This method helps reduce the variance in model performance estimates and provides a better understanding of how the model will perform on unseen data [41]. The cross-validation using all data was conducted to verify the feasibility of high-frequency CS prediction with a high generalization ability suitable for all data. The cross-validation experiment on individual data was conducted to develop individual learning models and verify the feasibility of high-frequency CS prediction suitable for each individual.

The ALSTM-FCN model was used as the training model. For comparison purposes, a simple LSTM (sLSTM) model (with 128 units), the FCN model (only the FCN part of ALSTM-FCN), and the LSTM-FCN model (the same model as ALSTM-FCN without the Attention layer) were also used. We used n-second ($n = 1$, 5, and 10) time-series eye-related indices as features for each learning. To the best of our knowledge, no study has verified the possibility of predicting CS faster than the prediction interval of 30 s by Islam et al. using eye-related indices [26]. The ground truth label was based on the evaluation of CS recorded 0.5 s after the last time-series eye-related indices for each n [s]. This time value considers the delay in evaluating the CS of a subject, as demonstrated by Nalivaiko et al. [11].

In this analysis, we performed a 4-level severity classification task for CS prediction, with the categories being 0 (none), 1 (slight), 2 (moderate), and 3 (severe), as well as a binary classification task for CS occurrence, dividing into the non-sickness group (0 and 1) and sickness group (2 and 3). The accuracy, precision, and recall metrics were used to evaluate the obtained DL models calculated from the following equations and confusion matrix (Fig. 7):

$$accuracy = \frac{T_P + F_N}{T_P + T_N + F_P + F_N} \tag{12}$$

$$precision = \frac{T_P}{T_P + F_P} \tag{13}$$

$$recall = \frac{T_P}{T_P + F_N} \tag{14}$$

TABLE 1. Results from a 4-level classification of CS severity through a five-fold cross-validation approach, using data from all participants. The "*n*-value" denotes the duration in seconds of the time-series data used as features.

n-value	Model	Accuracy	Precision				Recall			
			0 (None)	1 (Slight)	2 (medium)	3 (Severe)	0 (None)	1 (Slight)	2 (medium)	3 (Severe)
1 [s]	ALSTM-FCN	0.7109	0.8175	0.5634	0.5920	0.5764	0.8821	0.5243	0.5372	0.5019
	LSTM-FCN	0.6850	0.8063	0.5368	0.5340	0.4714	0.8767	0.5062	0.4628	0.4068
	sLSTM	0.6806	0.8139	0.5057	0.5512	0.5391	0.8429	0.5047	0.5169	0.4733
	FCN	0.5899	0.7973	0.4629	0.3584	0.3094	0.7478	0.3616	0.5624	0.2624
5 [s]	ALSTM-FCN	0.6873	0.8008	0.5364	0.5282	0.5731	0.8845	0.4327	0.5803	0.3698
	LSTM-FCN	0.6662	0.7801	0.5052	0.5433	0.4729	0.8608	0.4635	0.4859	0.3623
	sLSTM	0.6720	0.8108	0.4841	0.5543	0.5043	0.8332	0.4821	0.5344	0.4462
	FCN	0.5648	0.6188	0.3862	0.4169	0.3827	0.8903	0.1111	0.3388	0.2385
10 [s]	ALSTM-FCN	0.6754	0.7978	0.5367	0.4921	0.5349	0.8843	0.3654	0.5690	0.5208
	LSTM-FCN	0.6704	0.7626	0.5346	0.5922	0.4967	0.8525	0.4935	0.4843	0.4280
	sLSTM	0.6769	0.8109	0.5084	0.5220	0.6210	0.8519	0.5510	0.4251	0.4753
	FCN	0.5648	0.6459	0.4110	0.3797	0.4918	0.8442	0.3059	0.2685	0.1852

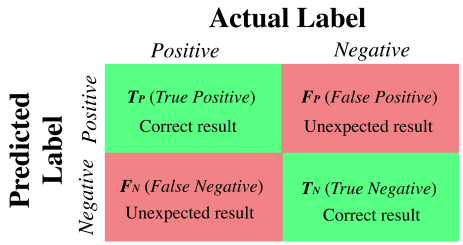


FIGURE 7. Confusion matrix of predicted labels and actual labels.

B. HYPERPARAMETER AND LOSS FUNCTION

The number of units for the two LSTM layers was determined through hyperparameter tuning using data from all subjects, whereas the parameters for the Conv1D layers were configured in accordance with the model proposed by Karim et al. [34]. To counteract overfitting, early stopping was implemented with a patience value of 10 during model training. The model was compiled using the Adam optimizer [42] with 300 epochs and a batch size of 128. The Adam optimizer iteratively adjusts the parameters of the model during training to minimize the loss function, which is computed iteratively, thereby enhancing model performance. Additionally, distinct loss functions were used for the multiclass and binary classification tasks. For the multiclass classification task, categorical cross entropy was used as the loss function:

$$L_{categorical} = -\sum_{i=1}^{C} t_i * log(y_i) \tag{15}$$

In this context, C denotes the overall count of classes; t_i represents the true label for class i, presented as a one-hot encoded vector; and y_i is the predicted probability for class i, calculated by the model. For the binary classification task, we used binary cross entropy as the loss function:

$$L_{binary} = -(t * log(y) + (1 - t) * log(1 - y)) \tag{16}$$

where t represents the true label, and y denotes the predicted probability for the positive class.

C. EVALUATION OF MODELS FOR ALL SUBJECTS

In this section, the results of five-fold cross-validation using all data obtained from 27 of the 30 participants in the experiment are presented. The data from three participants were excluded: the data of two participants could not be obtained due to PC issues; one participant retired from the experiment and could not complete it. We conducted two classification tasks for CS: 4-level classification for the severity of CS, and binary classification for the occurrence of CS.

For the 4-level classification task, the accuracy, precision, and recall values of the four models (ALSTM-FCN, LSTM-FCN, sLSTM, and FCN) obtained from five-fold cross-validation for *n*-second time-series eye-related indices are summarized in Table 1. The confusion matrix between predicted and actual labels using the ALSTM-FCN model is shown in Fig. 8. The ALSTM-FCN model achieved an accuracy of 71.09% when using 1-second time-series eye-related indices as features, with high precision (81.75%) and recall (88.21%) values for CS severity of 0 (none). However, this model struggled to distinguish between severity levels 1 (slight), 2 (moderate), and 3 (severe).

For the binary classification task, the accuracy, precision, and recall values of the four models obtained from five-fold cross-validation for *n*-second time-series eye-related indices are summarized in Table 2. The confusion matrix between predicted and actual labels using the ALSTM-FCN model is depicted in Fig. 9. In the case of the ALSTM-FCN and sLSTM model, an accuracy of approximately 82% was achieved, regardless of the value of n used. However, the precision and recall values for the sickness group are lower than those for the non-sickness group, with a difference of up to 58.2% in recall values observed when using the ALSTM-FCN model with an *n*-value of 10.

D. EVALUATION OF MODELS FOR EACH SUBJECT

In this analysis, we conducted five-fold cross-validation on individual data to develop individual DL models and investigate whether we can perform high-frequency CS

TABLE 2. Results from a binary classification of CS occurrence through a five-fold cross-validation approach, using data from all participants. The "*n*-value" denotes the duration in seconds of the time-series data used as features.

n-value	Model	Accuracy	Precision		Recall	
			Non-Sickness	Sickness	Non-Sickness	Sickness
1 [s]	ALSTM-FCN	0.8253	0.8828	0.6247	0.8915	0.6041
	LSTM-FCN	0.8287	0.8733	0.6483	0.9094	0.5588
	sLSTM	0.8282	0.8772	0.6407	0.9032	0.5773
	FCN	0.7929	0.8408	0.5693	0.9012	0.4334
5 [s]	ALSTM-FCN	0.8357	0.8754	0.6793	0.9150	0.5801
	LSTM-FCN	0.8125	0.8260	0.7058	0.9568	0.3395
	sLSTM	0.8194	0.8723	0.6286	0.8936	0.5801
	FCN	0.8019	0.8253	0.6408	0.9405	0.3477
10 [s]	ALSTM-FCN	0.8212	0.8337	0.7298	0.9575	0.3755
	LSTM-FCN	0.7949	0.8125	0.6518	0.9501	0.2988
	sLSTM	0.8260	0.8686	0.6527	0.9105	0.5498
	FCN	0.7828	0.7986	0.6154	0.9566	0.2233

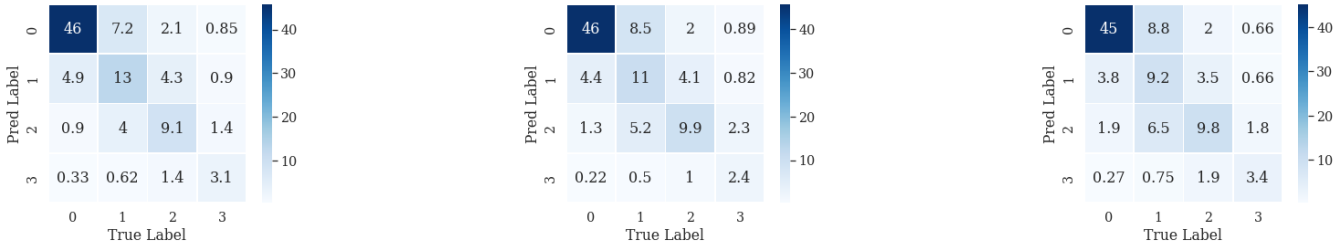


FIGURE 8. Illustration of the confusion matrix of the 4-level classification of CS severity, established through a five-fold cross-validation approach using the ALSTM-FCN model. The analysis is conducted using data from all participants (%). Subcaptions denoting the duration in seconds of the time-series data used as features are as follows: (a) 1 [s], (b) 5 [s], (c) 10 [s].

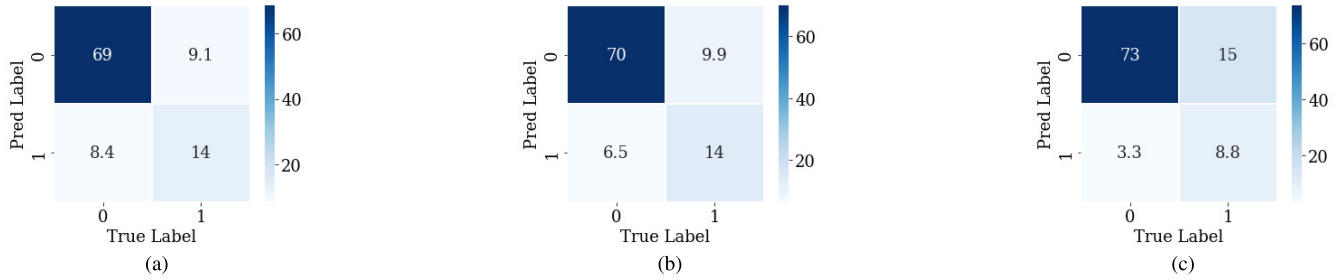


FIGURE 9. Illustration of the confusion matrix of the binary classification of CS occurrence, established through a five-fold cross-validation approach using the ALSTM-FCN model. The analysis is conducted using data from all participants (%). Subcaptions denoting the duration in seconds of the time-series data used as features are as follows: (a) 1 [s], (b) 5 [s], (c) 10 [s].

TABLE 3. Results from a 4-level classification analysis of CS severity, established through a five-fold cross-validation methodology for each individual participant. The "*n*-value" denotes the duration in seconds of the time-series data used as features.

n-value	Model	Accuracy	Precision				Recall			
			0 (None)	1 (Slight)	2 (medium)	3 (Severe)	0 (None)	1 (Slight)	2 (medium)	3 (Severe)
1 [s]	ALSTM-FCN	0.7934	0.8684	0.7145	0.7208	0.7230	0.9134	0.6741	0.6931	0.6979
	LSTM-FCN	0.7225	0.8356	0.6160	0.6042	0.6119	0.8675	0.5758	0.5983	0.6212
	sLSTM	0.7084	0.8353	0.5965	0.5770	0.5940	0.8465	0.5816	0.5762	0.5985
	FCN	0.6358	0.7451	0.5329	0.4938	0.5057	0.8169	0.4818	0.4391	0.5076
5 [s]	ALSTM-FCN	0.8116	0.8779	0.7413	0.7507	0.7543	0.9227	0.6989	0.7255	0.7302
	LSTM-FCN	0.7228	0.8372	0.6142	0.6103	0.6260	0.8611	0.5878	0.6120	0.6046
	sLSTM	0.7369	0.8509	0.6339	0.6232	0.6354	0.8682	0.6031	0.6232	0.6692
	FCN	0.6191	0.7326	0.4880	0.5150	0.4861	0.8018	0.4666	0.4580	0.3992
10 [s]	ALSTM-FCN	0.7587	0.8459	0.6619	0.6729	0.7220	0.8961	0.6183	0.6597	0.6616
	LSTM-FCN	0.7426	0.8514	0.6322	0.6633	0.6306	0.8735	0.6183	0.6373	0.6426
	sLSTM	0.7334	0.8511	0.6196	0.6252	0.6808	0.8554	0.6107	0.6331	0.6730
	FCN	0.6283	0.7452	0.5092	0.5231	0.4649	0.8075	0.4752	0.4916	0.4030

TABLE 4. Results from a binary classification analysis of CS occurrence, established through a five-fold cross-validation methodology for each individual participant. The "*n*-value" denotes the duration in seconds of the time-series data used as features.

n-value	Model	Accuracy	Precision		Recall	
			Non-Sickness	Sickness	Non-Sickness	Sickness
1 [s]	ALSTM-FCN	0.9123	0.9359	0.8389	0.9474	0.8085
	LSTM-FCN	0.8801	0.9134	0.7760	0.9272	0.7414
	sLSTM	0.8751	0.9150	0.7571	0.9175	0.7510
	FCN	0.8134	0.8548	0.6590	0.9034	0.5487
5 [s]	ALSTM-FCN	0.9177	0.9408	0.8478	0.9493	0.8253
	LSTM-FCN	0.8926	0.9223	0.8006	0.9348	0.7687
	sLSTM	0.8809	0.9202	0.7656	0.9202	0.7656
	FCN	0.8079	0.8469	0.6534	0.9062	0.5189
10 [s]	ALSTM-FCN	0.8967	0.9218	0.8183	0.9408	0.7697
	LSTM-FCN	0.8895	0.9242	0.7902	0.9266	0.7845
	sLSTM	0.8867	0.9207	0.7858	0.9273	0.7697
	FCN	0.8093	0.8467	0.6632	0.9074	0.5261

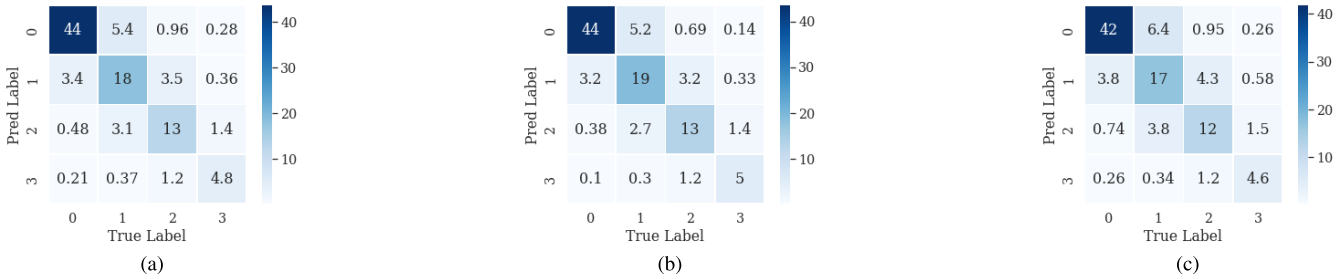


FIGURE 10. Illustration of the confusion matrix of the 4-level classification of CS severity, established through a five-fold cross-validation approach using the ALSTM-FCN model for each individual participant (%). Subcaptions denoting the duration in seconds of the time-series data used as features are as follows: (a) 1 [s], (b) 5 [s], (c) 10 [s].

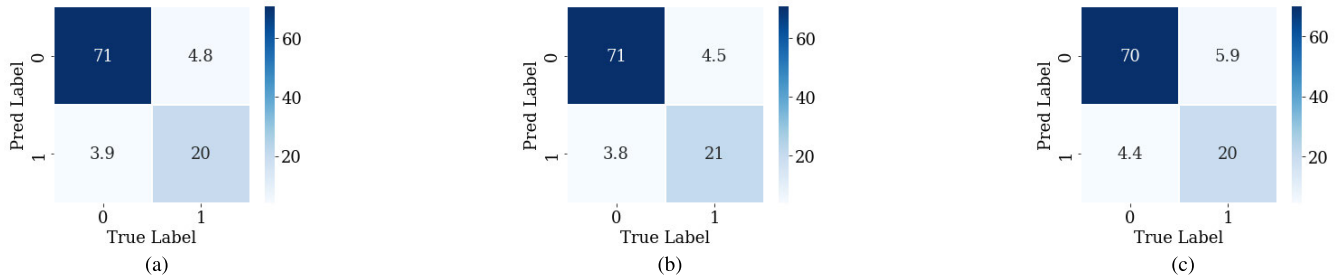


FIGURE 11. Illustration of the confusion matrix of the binary classification of CS occurrence, established through a five-fold cross-validation approach using the ALSTM-FCN model for each individual participant (%). Subcaptions denoting the duration in seconds of the time-series data used as features are as follows: (a) 1 [s], (b) 5 [s], (c) 10 [s].

prediction suitable for each individual. The data analyzed are from 25 of the 30 participants obtained from the user study. The data from five participants were excluded: the data of two participants could not be obtained because of PC issues, one participant retired from the experiment and could not complete it, and the remaining two participants did not reach a severity level of 2 or 3 for CS throughout the experiment. This cross-validation was performed on the data of all 25 participants. We conducted two classification tasks for CS: 4-level classification for the severity of CS, and binary classification for the occurrence of CS, as described in the previous section.

For the 4-level classification task, the accuracy, precision, and recall values of the four models (ALSTM-FCN, LSTM-FCN, sLSTM, and FCN) obtained from five-fold cross-validation for *n*-second time-series eye-related indices are summarized in Table 3. The confusion matrix between predicted and actual labels using the ALSTM-FCN model is depicted in Fig. 10. The ALSTM-FCN model achieved an accuracy of 81.16% when using 5-second time-series eye-related indices as features.

For the binary classification task, the accuracy, precision, and recall values of the four models obtained from five-fold cross-validation for *n*-second time-series eye-related indices are summarized in Table 4. The confusion matrix between predicted and actual labels using the ALSTM-FCN model is depicted in Fig. 11. The maximum accuracy of the ALSTM-FCN model was above 90%. Similarly, the

non-sickness group obtained precision and recall values of over 90%. This reveals the high classification capabilities of this model.

Overall, in both the 4-value classification and binary classification tasks, the ALSTM-FCN model outperformed the other models in terms of accuracy, precision, and recall. Moreover, the best results were achieved when 5-second time-series eye-related indices were used as features.

VI. DISCUSSION

In this study, our endeavor was to achieve a high-frequency prediction of the onset and intensity of CS through the usage of time-series eye-related indices and DL techniques. The ensuing content is organized into two distinct subsections. The first focuses on the discussion of prediction outcomes concerning CS, whereas the second entails a comparison of various DL models used for the purpose of CS prediction.

A. PREDICTION RESULTS OF CYBERSICKNESS

According to the CS prediction results of the ALSTM-FCN model and data obtained from all subjects, without distinguishing individuals, the accuracy values of 71.09% and 83.57% were obtained for the 4-level classification of CS severity and the binary classification of CS occurrence or absence, respectively. Islam's et al. [26] deep fusion approach achieved an accuracy of 80.7% using time-series eye-related indices in the 4-level classification of CS severity every 30 s. Although our approach is inferior, in terms of accuracy, to their results, we have demonstrated that it is possible to predict CS severity with higher frequency using time-series eye-related indices every 1 s.

Our greatest contribution is high-frequency CS prediction. Conversely, in terms of precision, and recall, the CS severity detection performance was biased in our approach, similar to Islam et al. [26]. In particular, our approach struggled to classify people with higher CS severity compared with those with lower CS severity. According to the confusion matrix, as the severity of CS increases, the proportion of the actual labels decreases in distribution. This shows that the biased distribution of the actual labels may have affected the results. Moreover, individual differences in the evaluations of the participants could have influenced the results. These results may highlight the limitations of relying on subjective evaluations of CS.

It is not possible to directly compare our results with previous studies. This is because the physiological indices used as features, the evaluation methods of CS (during or after immersion in VR), the length of the time-series data used as features, and the number of classes in the classification are different among previous studies. However, we compared our results with those of previous studies from the perspective of the use of sensors integrated into HMDs and external sensors (Table 5). Kim et al. [6] used EEG data as features and predicted the 5-level severity of CS with 89.16% accuracy. Garcia-Agundez et al. [14] used ECG, EEG, respiratory data, skin conductivity data, and relevant game parameters, such as avatar linear, and angular speed, acceleration, head movements, and on-screen collisions, as features and obtained 82% accuracy in binary classification and 56% accuracy in ternary classification. Chang et al. [22] used multiple eye-related indices, such as the fixation time and the distance between the eye gaze and object-position sequence as features. Their model could explain 34.8% of the total variance of CS. Islam et al. [7] used HR, BR, and GSR in the preceding 2-min as features and predicted the ternary severity of CS with 97.44% accuracy. Islam et al. [26] used multiple eye-related indices, such as pupil diameter, gaze direction, and convergence distance, as features. Their approach predicted the 4-level severity of CS with 80.7% accuracy. Our approach is superior to these previous studies in terms of high-frequency CS prediction.

In addition, we investigated the feasibility of training individual DL models for each participant and performing high-frequency prediction of CS occurrence and severity that is tailored to each participant. As a result, we achieved an accuracy of approximately 80% in the 4-level classification of CS severity and approximately 90% in the binary classification of CS occurrence. These findings indicate that the use of time-series eye-related indices and DL for high-frequency CS prediction is effective. This could be a valuable approach for addressing the issue of CS in the future advancements of VR technology. In addition, we demonstrated that it is possible to develop CS prediction models for each individual, which is expected to be an important concept for considering individual differences when using VR technology.

B. COMPARISON OF DEEP LEARNING MODELS FOR CYBERSICKNESS PREDICTION

We used four distinct DL models, namely, ALSTM-FCN, LSTM-FCN, sLSTM, and FCN, for the prediction of CS. Notably, the model exclusively relying on FCN exhibited the least accurate outcomes across all analyses. The LSTM and TCN are harnessed to integrate temporal sequencing within the learning process. The LSTM incorporates a gating mechanism that governs the retention or forgetting of information, allowing it to sequentially process data points [38]. On the other hand, TCN functions as a feature extraction module in an FCN branch. As a convolutional network, TCN effectively extracts and processes localized features within specific time windows [36]. Consequently, LSTM is adept at capturing intricate temporal features and long-term dependencies, whereas TCN excels at discerning broader patterns. Given the superior accuracy achieved by models incorporating LSTM (LSTM-FCN and sLSTM) compared to those solely using FCN coupled with these distinctive algorithmic attributes, it is posited that intricate severity-dependent patterns of change exist within eye-related indices. The capacity of LSTM to effectively learn and comprehend these intricate patterns is believed to underlie the observed results.

Additionally, the ALSTM-FCN model, incorporating an attention mechanism, generally outperformed the

TABLE 5. Comparative analysis of CS prediction outcomes between previous studies and our proposed approach.

Reference	Physiological indices used as features	Evaluation of CS	Learning model	Results
Kim et al. [6] (2019)	EEG	Original subjective evaluation (after experiment, not in immersive)	CNN, LSTM	Accuracy of 89.16% in 5-level classification
Agundez et al. [14] (2019)	ECG, EED, BR, Skin conductivity, Game parameters in VR	SSQ (after experiment, not in immersive)	SVM, KNN, Multilayer perceptron	Accuracy of 82% in binary classification. Accuracy of 56% in ternary classification.
Islam et al. [7] (2020)	HR, BR, GSR	Shortened FMS collected every 2 min. (in immersive)	CNN-LSTM	Accuracy of 97.44% in ternary classification.
Islam et al. [26] (2021)	Eye-related indices	Shortened FMS collected every 30 s. (in immersive)	CNN-LSTM	Accuracy of 80.7% in 4-level classification.
Chang et al. [22] (2021)	Eye-related indices	SSQ (after experiment, not in immersive)	Multiple regression	34.8% of the total variance of cybersickness
Our approach	Eye-related indices	Subjective evaluation based on SSQ every 1, 5, or 10 s. (in immersive)	ALSTM-FCN, LSTM-FCN, sLSTM, FCN	Accuracy of 71.09% in 4-level classification. (1-second time-series) Accuracy of 83.57% in binary classification. (5-second time-series)

LSTM-FCN model in terms of accuracy. Attention is a mechanism that enables models to prioritize significant segments of time-series data by evaluating the importance of each time point within the data and subsequently emphasizing the most crucial time points [39]. This mechanism is believed to have facilitated the model's ability to concentrate on pivotal aspects of the time-series eye-related indices, resulting in the highest performance among the tested models.

In conclusion, our analysis indicates that models relying solely on FCN are inadequate for predicting CS, especially in task settings such as the one in this study. Conversely, the algorithmic attributes of LSTM and the incorporation of the attention mechanism prove more effective in utilizing eye-related indices as features for predicting CS, as evidenced by the observed results.

VII. LIMITATIONS

Our proposed approach improved the performance of high-frequency prediction; however, some limitations need to be mentioned.

A. IMBALANCED SAMPLE SIZE

A key limitation of our study pertains to the uneven sample distribution, in gender and age. Among the 30 participants, 26 were male, whereas only 4 were female. Moreover, the age distribution leans toward a younger demographic, as evidenced by a mean age of 23.57 and a standard deviation of 4.26.

A previous study has suggested that there may be gender differences in the effects of HMDs on CS [3]. For instance, several large studies (sample sizes ranging from 160 to 837) reported that females experience greater CS than males [43], [44], [45]. Furthermore, a study focusing on eye-related indices found significant differences between males and females. Namely, the study found that females tend to exhibit more exploratory gaze behavior, as indicated by larger saccade amplitudes and longer scan paths and inspect images faster than males due to a shorter ratio of fixation durations to saccade durations [46]. In another study, as discussed by Cantoni et al. [33], it was highlighted that pupil diameter is recognized to vary depending on factors such as gender and age.

As a consequence, the imbalanced sample that favors male participants and a younger demographic might have introduced potential biases in the outcomes of our study. However, we effectively mitigated these effects through the individually normalized methodology adopted in this study. Nevertheless, ensuring a more equitable representation of both genders and various age groups among the participants would be prudent in subsequent investigations.

B. NON-INTERACTIVE STUDY DESIGN AND SUBJECTIVE EVALUATION

The second limitation of our study is that it focused solely on non-interactive cases in which users passively view VR video content. Previous studies on predicting CS have focused on various cases, such as cases in which users passively view VR videos without interaction [7], [22], cases in which users actively engage in interactions in the VR environment [21], [47], [48], and cases that encompass both types of scenarios [17], [26]. Additionally, a crucial factor that we did not account for in our study is the type of controller used. Prior research has indicated that the choice

of controller can impact the extent of CS experienced in interactive scenarios [49].

Therefore, our findings may not apply to a range of scenarios that involve user interaction with VR content, and further investigation is needed to determine the applicability of our approach in these cases. Additionally, our study relied on subjective self-reporting to measure CS severity, which may be influenced by individual differences. To address this, future studies should include objective measures, such as physiological data, to supplement self-reported data.

VIII. APPLICATION AND FUTURE WORK

Our study investigated the high-frequency prediction of CS and explored its potential via offline simulations using ML/DL techniques. For practical use and real applications, we propose the development of a system capable of real-time generation of predictions covering CS severity for new participants as they immerse themselves in VR environments. This system would rely on recorded physiological data, such as eye-related indices, to make predictions, thereby allowing for the assessment of CS severity and the adjustment of VR environments to mitigate the effects of CS.

We believe that combining our high-frequency CS prediction approach with methods to mitigate CS would be effective to provide high quality of VR experience. For example, the use of a visual field-of-view restrictor has been proposed to mitigate CS severity [50], [51], [52], [53]. It has also been suggested that image manipulation strategies, such as image blurring techniques, have a significant effect on mitigating CS [54], [55], [56]. Implementing a system that combines our high-frequency CS prediction method and CS mitigation methods can have considerable implications for the VR industry, enhancing user experience and ensuring a safer and more comfortable environment for users.

It is crucial to note that our study is limited to offline simulations for high-frequency prediction and does not include the construction of a real-time, high-frequency prediction system. Therefore, future studies in the field of CS should focus on the development and implementation of such a prediction system, which will not only contribute to a better understanding of CS but also lead to practical solutions that can be employed across various VR applications, improving the overall user experience in VR environments.

IX. CONCLUSION

In this study, we presented an approach for predicting the occurrence and severity of CS with higher frequency than in previous studies. Based on our approach, the ALSTM-FCN model achieved an accuracy of 71.09% for the 4-level classification of CS severity with high frequency and 83.57% for the binary classification of CS occurrence using data obtained from all participants without distinguishing individuals. We also examined whether we could develop DL models for each participant and perform high-frequency CS predictions suitable for each participant. The results indicate that a CS prediction model that corresponds to individual differences can be developed, with approximately 80% accuracy for the 4-level classification of CS severity and approximately 90% accuracy for the binary classification of CS occurrence.

We employed the time-series eye-related indices taken every 1-, 5-, and 10-s as features, and our approach is the fastest to predict CS to the best of our knowledge. Although it may be challenging due to limited processing power based on PC specifications, we believe that CS can be predicted with high frequency using the same approach, as many HMDs released in recent years have built-in eye-tracking sensors. For the future development of VR, we further highlight the importance of combining our high-frequency CS prediction approach with methods to mitigate CS, particularly when CS is detected.

Subsequently, we plan to conduct user studies to achieve more accurate high-frequency CS prediction by incorporating more factors related to video content, such as avatar movement in VR space, the distance between objects in VR space, and line of sight as features. In addition, we plan to develop real-time and high-frequency prediction systems that are not restricted to offline simulations, thereby further enhancing the practical application of our approach.

REFERENCES

[1] J. J. LaViola, "A discussion of cybersickness in virtual environments," *ACM SIGCHI Bull.*, vol. 32, no. 1, pp. 47–56, Jan. 2000.

[2] S. Davis, K. Nesbitt, and E. Nalivaiko, "A systematic review of cybersickness," in *Proc. Conf. Interact. Entertainment*, Dec. 2014, pp. 1–9.

[3] J. Munafo, M. Diedrick, and T. A. Stoffregen, "The virtual reality head-mounted display oculus rift induces motion sickness and is sexist in its effects," *Exp. Brain Res.*, vol. 235, no. 3, pp. 889–901, Mar. 2017.

[4] Y. Kim, H. Kim, E. Kim, H. Ko, and H.-T. Kim, "Characteristic changes in the physiological components of cybersickness," *Psychophysiology*, vol. 42, pp. 25–616, Oct. 2005.

[5] R. Islam, "A deep learning based framework for detecting and reducing onset of cybersickness," in *Proc. IEEE Conf. Virtual Reality 3D User Interface Abstr. Workshops (VRW)*, Mar. 2020, pp. 559–560.

[6] J. Kim, W. Kim, H. Oh, S. Lee, and S. Lee, "A deep cybersickness predictor based on brain signal analysis for virtual reality contents," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10579–10588.

[7] R. Islam, Y. Lee, M. Jaloli, I. Muhammad, D. Zhu, P. Rad, Y. Huang, and J. Quarles, "Automatic detection and prediction of cybersickness severity using deep neural networks from user's physiological signals," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Nov. 2020, pp. 400–411.

[8] H. Oh and W. Son, "Cybersickness and its severity arising from virtual reality content: A comprehensive study," *Sensors*, vol. 22, no. 4, p. 1314, 2022.

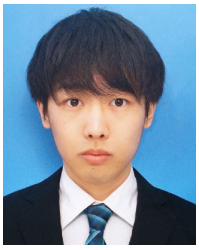
[9] S. A. A. Naqvi, N. Badruddin, M. A. Jatoi, A. S. Malik, W. Hazabbah, and B. Abdullah, "EEG based time and frequency dynamics analysis of visually induced motion sickness (VIMS)," *Australas. Phys. Eng. Sci. Med.*, vol. 38, no. 4, pp. 721–729, Dec. 2015.

[10] Y. Yokota, M. Aoki, K. Mizuta, Y. Ito, and N. Isu, "Motion sickness susceptibility associated with visually induced postural instability and cardiac autonomic responses in healthy subjects," *Acta Oto-Laryngolog.*, vol. 125, no. 3, pp. 280–285, Mar. 2005.

[11] E. Nalivaiko, S. L. Davis, K. L. Blackmore, A. Vakulin, and K. V. Nesbitt, "Cybersickness provoked by head-mounted display affects cutaneous vascular tone, heart rate and reaction time," *Physiol. Behav.*, vol. 151, pp. 583–590, Nov. 2015.

[12] S. Bruck and P. A. Watters, "The factor structure of cybersickness," *Displays*, vol. 32, no. 4, pp. 153–158, Oct. 2011.

[13] T. Irmak, D. Pool, and R. Happee, "Objective and subjective responses to motion sickness: The group and the individual," *Exp. Brain Res.*, vol. 239, pp. 1–17, Feb. 2021.
[14] A. Garcia-Agundez, C. Reuter, H. Becker, R. Konrad, P. Caserman, A. Miede, and S. Göbel, "Development of a classifier to determine factors causing cybersickness in virtual reality environments," *Games Health J.*, vol. 8, no. 6, pp. 439–444, Dec. 2019.
[15] R. K. Kundu, R. Islam, P. Calyam, and K. A. Hoque, "TruVR: Trustworthy cybersickness detection using explainable machine learning," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Oct. 2022, pp. 777–786.
[16] W. Jin, J. Fan, D. Gromala, and P. Pasquier, "Automatic prediction of cybersickness for virtual reality games," in *Proc. IEEE Games, Entertainment, Media Conf. (GEM)*, Aug. 2018, pp. 1–9.
[17] R. Islam, K. Desai, and J. Quarles, "Towards forecasting the onset of cybersickness by fusing physiological, head-tracking and eye-tracking with multimodal deep fusion network," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Oct. 2022, pp. 121–130.
[18] M. S. Anwar, J. Wang, W. Khan, A. Ullah, S. Ahmad, and Z. Fei, "Subjective QoE of 360-degree virtual reality videos and machine learning predictions," *IEEE Access*, vol. 8, pp. 148084–148099, 2020.
[19] B. Keshavarz and H. Hecht, "Validating an efficient method to quantify motion sickness," *Hum. Factors*, vol. 53, pp. 415–426, Aug. 2011.
[20] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness," *Int. J. Aviation Psychol.*, vol. 3, no. 3, pp. 203–220, Jul. 1993.
[21] P. Lopes, N. Tian, and R. Boulic, "Eye thought you were sick! Exploring eye behaviors for cybersickness detection in VR," in *Proc. Motion, Interact. Games*, Oct. 2020, pp. 1–10.
[22] E. Chang, H. T. Kim, and B. Yoo, "Predicting cybersickness based on user's gaze behaviors in HMD-based virtual reality," *J. Comput. Design Eng.*, vol. 8, no. 2, pp. 728–739, Apr. 2021.
[23] Y. Nam, U. Hong, H. Chung, and S. R. Noh, "Eye movement patterns reflecting cybersickness: Evidence from different experience modes of a virtual reality game," *Cyberpsychol., Behav., Social Netw.*, vol. 25, no. 2, pp. 135–139, Feb. 2022.
[24] H. Oyamada, A. Iijima, A. Tanaka, K. Ukai, H. Toda, N. Sugita, M. Yoshizawa, and T. Bando, "A pilot study on pupillary and cardiovascular changes induced by stereoscopic video movies," *J. NeuroEng. Rehabil.*, vol. 4, no. 1, p. 37, Dec. 2007.
[25] C. Guo, J. Ji, and R. So, "Could OKAN be an objective indicator of the susceptibility to visually induced motion sickness?" in *Proc. IEEE Virtual Reality Conf.*, Mar. 2011, pp. 87–90.
[26] R. Islam, K. Desai, and J. Quarles, "Cybersickness prediction from integrated HMD's sensors: A multimodal deep fusion approach using eye-tracking and head-tracking data," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Oct. 2021, pp. 31–40.
[27] S. Wibirama, P. I. Santosa, P. Widyarani, N. Brilianto, and W. Hafidh, "Physical discomfort and eye movements during arbitrary and optical flow-like motions in stereo 3D contents," *Virtual Reality*, vol. 24, no. 1, pp. 39–51, Mar. 2020.
[28] M. M. Bradley, L. Miccoli, M. A. Escrig, and P. J. Lang, "The pupil as a measure of emotional arousal and autonomic activation," *Psychophysiology*, vol. 45, no. 4, pp. 602–607, Jul. 2008.
[29] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, "Eye-tracking analysis for emotion recognition," *Comput. Intell. Neurosci.*, vol. 2020, pp. 1–13, Sep. 2020.
[30] Y. Wang, G. Naylor, S. E. Kramer, A. A. Zekveld, D. Wendt, B. Ohlenforst, and T. Lunner, "Relations between self-reported daily-life fatigue, hearing status, and pupil dilation during a speech perception in noise task," *Ear Hearing*, vol. 39, no. 3, pp. 573–582, May 2018.
[31] M. Kaufeld, J. Bourdeinik, L. M. Prinz, M. Mundt, and H. Hecht, "Emotions are associated with the genesis of visually induced motion sickness in virtual reality," *Exp. Brain Res.*, vol. 240, no. 10, pp. 2757–2771, Oct. 2022.
[32] H. Kim, D. J. Kim, W. H. Chung, K.-A. Park, J. D. K. Kim, D. Kim, K. Kim, and H. J. Jeon, "Clinical predictors of cybersickness in virtual reality (VR) among highly stressed people," *Sci. Rep.*, vol. 11, no. 1, Jun. 2021, Art. no. 12139.
[33] V. Cantoni, L. Cascone, M. Nappi, and M. Porta, "Demographic classification through pupil analysis," *Image Vis. Comput.*, vol. 102, Oct. 2020, Art. no. 103980.
[34] F. Karim, S. Majumdar, H. Darabi, and S. Chen, "LSTM fully convolutional networks for time series classification," *IEEE Access*, vol. 6, pp. 1662–1669, 2018.
[35] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 1578–1585.
[36] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *Proc. Eur. Conf. Comput. Vis.*, G. Hua and H. Jégou, Eds. Cham, Switzerland: Springer, 2016, pp. 47–54.
[37] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*.
[38] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[39] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.
[40] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
[41] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of K-fold cross-validation," *J. Mach. Learn. Res.*, vol. 5, pp. 1089–1105, Dec. 2004.
[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[43] D. M. Shafer, C. P. Carbonara, and M. F. Korpi, "Factors affecting enjoyment of virtual reality games: A comparison involving consumer-grade virtual reality technology," *Games Health J.*, vol. 8, no. 1, pp. 15–23, Feb. 2019.
[44] T. Luong, A. Pléchata, M. Möbus, M. Atchapero, R. Böhm, G. Makransky, and C. Holz, "Demographic and behavioral correlates of cybersickness: A large lab-in-the-field study of 837 participants," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Oct. 2022, pp. 307–316.
[45] T. A. Doty, J. W. Kelly, M. C. Dorneich, and S. B. Gilbert, "Does interpupillary distance (IPD) relate to immediate cybersickness?" in *Proc. IEEE Conf. Virtual Reality 3D User Interface Abstr. Workshops (VRW)*, Mar. 2023, pp. 661–662.
[46] B. A. Sargezeh, N. Tavakoli, and M. R. Daliri, "Gender-based eye movement differences in passive indoor picture viewing: An eye-tracking study," *Physiol. Behav.*, vol. 206, pp. 43–50, Jul. 2019.
[47] D. Monteiro, H.-N. Liang, X. Tang, and P. Irani, "Using trajectory compression rate to predict changes in cybersickness in virtual reality games," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Oct. 2021, pp. 138–146.
[48] J. Wang, H.-N. Liang, D. Monteiro, W. Xu, and J. Xiao, "Real-time prediction of simulator sickness in virtual reality games," *IEEE Trans. Games*, vol. 15, no. 2, pp. 252–261, Jun. 2023.
[49] D. Monteiro, H.-N. Liang, J. Wang, H. Chen, and N. Baghaei, "An in-depth exploration of the effect of 2D/3D views and controller types on first person shooter games in virtual reality," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Nov. 2020, pp. 713–724.
[50] A. S. Fernandes and S. K. Feiner, "Combating VR sickness through subtle dynamic field-of-view modification," in *Proc. IEEE Symp. 3D User Interface (3DUI)*, Mar. 2016, pp. 201–210.
[51] N. Kala, K. Lim, K. Won, J. Lee, T. Lee, S. Kim, and W. Choe, "P-218: An approach to reduce VR sickness by content based field of view processing," in *SID Symp. Dig. Tech. Papers*, vol. 48, no. 1, 2017, pp. 1645–1648.
[52] N.-G. Kim and B.-S. Kim, "The effect of retinal eccentricity on visually induced motion sickness and postural control," *Appl. Sci.*, vol. 9, no. 9, p. 1919, May 2019.
[53] M. Al Zayer, I. B. Adhanom, P. MacNeilage, and E. Folmer, "The effect of field-of-view restriction on sex bias in VR sickness and spatial navigation performance," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2019, pp. 1–12.
[54] P. Budhiraja, M. R. Miller, A. K. Modi, and D. A. Forsyth, "Rotation blurring: Use of artificial blurring to reduce cybersickness in virtual reality first person shooters," 2017, *arXiv:1710.02599*.
[55] W. Qionghua, W. Hui, and W. Qiang, "Some experimental results of relieving discomfort in virtual reality by disturbing feedback loop in human brain," Mar. 2019, *arXiv:1903.12617*.
[56] G.-Y. Nie, H. B. Duh, Y. Liu, and Y. Wang, "Analysis on mitigation of visually induced motion sickness by applying dynamical blurring on a user's retina," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 8, pp. 2535–2545, Aug. 2020.



SHOGO SHIMADA received the B.S. degree in computer science from Tokyo Metropolitan University, in 2022, where he is currently pursuing the master's degree with the Department of Computer Science, Graduate School of Systems Design. His research interests include virtual reality, machine learning, deep learning, and human–computer interaction. He is a member of ACM.

PEERAWAT PANNATTEE received the B.E. degree in electronics and telecommunication engineering and the M.E. degree in electrical engineering from the King Mongkut's University of Technology Thonburi (KMUTT), Bangkok, Thailand, in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree in computer science with Tokyo Metropolitan University, Tokyo, Japan. His research interests include image processing, computer vision, machine learning, deep learning, virtual reality, and human–computer interaction.

YASUSHI IKEI (Member, IEEE) received the Ph.D. degree in industrial mechanical engineering from The University of Tokyo, in 1988. He joined the Tokyo Metropolitan Institute of Technology, in 1992. After he worked for Tokyo Metropolitan University as a Professor, he moved to The University of Tokyo. He is currently a Professor with the Graduate School of Information Science and Technology and a Professor Emeritus with Tokyo Metropolitan University. His research interests are in the areas of virtual reality, ultra reality, telepresence, multisensory display, and cognitive engineering. He is a member of ACM, JSME, and a fellow. He was the former Vice President of the Virtual Reality Society of Japan (VRSJ).

NOBUYUKI NISHIUCHI (Member, IEEE) received the Ph.D. degree in engineering from Yokohama National University, Japan, in 2004. He is currently a Professor with the Faculty of Systems Design, Tokyo Metropolitan University. His main research fields are human interface, usability engineering, image processing, ergonomics, and biometrics. He served as a member of the International Program Committee for international conferences, such as ICBAKE and CISIM.

VIBOL YEM (Member, IEEE) received the Ph.D. degree in engineering from the University of Tsukuba, in 2015. He was an Assistant Professor with Tokyo Metropolitan University, from April 2018 to February 2023. He is currently an Associate Professor with the Faculty of Engineering, Information and Systems, University of Tsukuba. His research interests are human interface, tactile/haptic devices, VR/AR, and wearable, robotics. He is a member of ACM and the IEEE Computer Society.