LiteVR: Interpretable and Lightweight Cybersickness Detection using Explainable AI

Ripan Kumar Kundu *
University of Missouri

Rifatul Islam †
Kennesaw State University

John Quarles ‡
University of Texas at San Antonio

Khaza Anuarul Hoque §
University of Missouri

ABSTRACT

Cybersickness is a common ailment associated with virtual reality (VR) user experiences. Several automated methods exist based on machine learning (ML) and deep learning (DL) to detect cybersickness. However, most of these cybersickness detection methods are perceived as computationally intensive and black-box methods. Thus, those techniques are neither trustworthy nor practical for deploying on standalone energy-constrained VR head-mounted devices (HMDs). In this work, we present an explainable artificial intelligence (XAI)-based framework *LiteVR* for cybersickness detection, explaining the model's outcome, reducing the feature dimensions, and overall computational costs. First, we develop three cybersickness DL models based on long-term short-term memory (LSTM), gated recurrent unit (GRU), and multilayer perceptron (MLP). Then, we employed a post-hoc explanation, such as SHapley Additive Explanations (SHAP), to explain the results and extract the most dominant features of cybersickness. Finally, we retrain the DL models with the reduced number of features. Our results show that eye-tracking features are the most dominant for cybersickness detection. Furthermore, based on the XAI-based feature ranking and dimensionality reduction, we significantly reduce the model's size by up to 4.3×, training time by up to 5.6×, and its inference time by up to 3.8×, with higher cybersickness detection accuracy and low regression error (i.e., on Fast Motion Scale (FMS)). Our proposed lite LSTM model obtained an accuracy of 94% in classifying cybersickness and regressing (i.e., FMS 1–10) with a Root Mean Square Error (RMSE) of 0.30, which outperforms the state-of-the-art. Our proposed *LiteVR* framework can help researchers and practitioners analyze, detect, and deploy their DL-based cybersickness detection models in standalone VR HMDs.

Keywords: Virtual Reality, Cybersickness Detection, Explainable Artificial Intelligence, Deep Learning, Model Reduction

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Virtual reality; Human-centered computing—Human computer interaction (HCI)—HCI design and evaluation methods.

1 INTRODUCTION

The applications of virtual reality (VR) are very diverse and are rapidly increasing in areas such as health care [19,56], education [2], military training [82], surgical training [48], and disaster management/public safety [16,51]. However, VR users often experience virtual reality sickness, or cybersickness, which hinders their immersive experience. Thus cybersickness has emerged as an important obstacle [49] to the wider acceptability of VR. Cybersickness can be defined as a set of unpleasant symptoms such as eyestrain, headache,

---
*e-mail: rkcgc@umsystem.edu
†e-mail: rislam11@kennesaw.edu
‡e-mail: john.quarles@utsa.edu
§e-mail: hoquek@umsystem.edu

nausea, disorientation, or even vomiting [59]. One of the popular techniques to detect cybersickness is to use post-immersive subjective questionnaires such as Simulator Sickness Questionnaire (SSQ) and the VR Sickness Questionnaire (VRSQ) [66]. In contrast, the Fast Motion Sickness Scale (FMS) [32] can assess cybersickness severity during immersion. Note, FMS relies on user feedback, i.e., needs human intervention during the VR immersion [47]. To overcome these limitations, deep learning (DL) and machine learning (ML) have recently become popular for cybersickness detection [1, 21, 22, 24, 27, 30, 58].

State-of-the-art DL models can detect cybersickness with good accuracy. For instance, Islam et al. [24] proposed a cybersickness severity detection using a deep fusion network with an accuracy of 87.7% from users' eye tracking and head tracking data. Other researchers used electroencephalography (EEG) [58, 72, 83], stereoscopic video [24, 52] and bio-physiological signal [21, 58] data for detecting cybersickness with higher accuracy. Despite the great prospect of DL models in detecting cybersickness, these methods have several limitations:

- Most of the cybersickness DL-based detection models rely on black-box models; thus, they lack explainability. The explainability of the DL models can significantly improve the model's understanding and provide insight into why and how the DL model arrived at a specific decision. Identifying and understanding important features leading to cybersickness can help designers to develop more effective cybersickness detection models.

- The multimodal data generated by HMD's integrated sensors as well as external physiological sensors (EEG, GSR, HR) results in a complex dataset and power-hungry DL models for cybersickness detection. As a result, deploying these models on standalone energy-constrained HMDs (e.g., Meta Quest Pro) is often challenging and not feasible. By identifying the dominant features, it can be useful to design and develop lightweight DL models that can improve computational costs, inference time, and model complexity.

To demonstrate the need for the proposed XAI approach, consider that a VR game developer may want to develop a cybersickness prediction model for their game based on alpha/beta players who are using a Meta Quest Pro HMD. If the developer was using a black box cybersickness prediction model without explainability in the game, they could not easily determine which of the model's features (e.g., pupil diameter) contributed to cybersickness prediction. Hence, they would have use trial and error to reduce the model size such that it would minimize resource usage on the already resource constrained Meta Quest Pro.

Limited research has been conducted to address these research gaps. Recently, Kundu et al. [38] used inherently interpretable ML models for cybersickness detection and explanation using bio-physiological [25] and gameplay datasets [57]. However, this work considered binary classifiers, which can only detect the presence or absence of cybersickness. Such binary classification for cybersickness detection is less effective for realistic VR applications.

Additionally, inherently interpretable models are typically dependent on the data properties and thus suffer from the dimensionality problem [64]. For instance, decision tree-based inherently interpretable models can suffer from overfitting problems because of over-complex trees that do not generalize the data well, which may eventually lead to their poor performance in cybersickness classification [54, 58].

Several other works use feature selection or dimensionality reduction techniques, such as principal component analysis (PCA) [34, 36, 45, 63, 70] to address the high-dimensionality issue in DL models for cybersickness detection. For instance, Mawalid et al. in [46] used time-domain feature extraction methods to extract the EEG statistical features for classifying cybersickness. Similarly, Lin et al. [41] applied ML models with PCA to extract the cybersickness-related features to predict the cybersickness level. Furthermore, Kottaimalai et al. in [37] used PCA to reduce dimensions, complexity, and computational time for EEG signals to detect cybersickness. Note, PCA attempts to cover as much variance as possible among the feature spaces in a dataset. If the number of principal components in a dataset is not selected carefully, it is possible to miss some information compared to the original set of features [3]. Hence, applying PCA-based dimensionality reduction may result in losing important features essential for accurate cybersickness detection.

To address the above-mentioned challenges, we propose a novel methodology, LiteVR–*an XAI-based framework for cybersickness detection*, explanation, and feature size reduction (also known as dimensionality reduction). Specifically, we first develop three DL models for cybersickness detection based on long-term short-term memory (LSTM), gated recurrent unit (GRU), and multilayer perceptron (MLP). Then we employ post-hoc explanation techniques, namely SHapley Additive exPlanations (SHAP) [44], to provide global and local explanations for analyzing, identifying, and ranking the dominant features causing cybersickness. The identified dominant features are then used to retrain the models, (i.e., to train them with a reduced number of features instead of using all the features). This makes a cybersickness detection model that is easy to use and has a much smaller number of trainable parameters. This makes training and inference faster while maintaining baseline accuracy. For instance, our results show that an LSTM model with all features classifies the cybersickness severity into 4 classes (none, low, medium, and high) with an accuracy of 92% and regresses (FMS 1–10) the ongoing cybersickness with a Root Mean Square Error (RMSE) value of 0.36. However, after the XAI-based feature reduction (i.e., with only 1/3 of the features of the baseline model) of the same LSTM model, we can classify the cybersickness severity and predict the ongoing cybersickness with an accuracy of 94% and an RMSE of 0.30, respectively. Furthermore, our XAI-based model reduction approach results in up to a 4.3× reduction in the model size (for LSTM) and up to a 5.5× and 3.8× improvement in training and inference time, respectively. Therefore, we believe that the proposed method can aid future researchers in comprehending, analyzing, and designing cybersickness detection models that are suitable for real-time cybersickness detection for standalone HMDs. To the best of our knowledge, this is the first work applying XAI for explaining and feature size reduction of DL-based cybersickness detection models.

2 RELATED WORKS

The most popular theory to explain the reason behind cybersickness is the sensory conflict [39]. This theory states that cybersickness occurs when the visual sensory system perceives motion, but the vestibular system does not. However, other theories, such as poison theory and postural instability theory, even cybersecurity have also been identified as the causes of cybersickness [17, 39, 74, 75]. In addition, factors such as age, gender, and prior VR experience of users can also impact the degree of cybersickness [12, 14, 20]. The state-of-the-art works in cybersickness detection can be broadly divided into two categories that use either subjective and objective measurements. Recently, automated cybersickness detection using advanced ML and DL methods is getting more popular as it does not require any human interventions. In this section, we highlight the state-of-the-art works for cybersickness detection.

To measure cybersickness, researchers have proposed several subjective measurements such as the Simulator Sickness Questionnaire (SSQ) [6, 7, 9, 13, 50, 66, 78, 84, 85], the FMS [32], and the Motion Sickness Susceptibility Questionnaire (MSSQ) [31]. In contrast, several researchers have also proposed objective measurements (i.e., physiological signals) for cybersickness [21, 25, 57] detection. Previous research has shown that objective measurements (e.g., heart rate, galvanic skin response, and electroencephalogram (EEG) signals) vary significantly when cybersickness occurs [11, 25, 42, 42, 60, 62]. For example, they found that HR and EEG delta waves correlate positively with cybersickness, whereas EEG beta waves correlate negatively [42]. On the contrary, another study reported that GSR has a stronger positive correlation with cybersickness than other objective measurements that can detect cybersickness [25, 71].

To automatically detect cybersickness from a variety of objective measurements (physiological signals) data, subjective measurements (FMSQ, SSQ) data and integrated sensors measurements (eye-tracking, head-tracking, motion-flow, etc.) data in HMD, numerous ML and DL-based approaches have recently been proposed [1, 18, 21, 24–26, 33, 53, 58, 61, 62, 77, 80, 81]. For instance, in [1], a machine–deep–ensemble learning method is applied to classify the cybersickness. In contrast, Porcino et al. [57] used a symbolic ML-based approach to identify the levels of cybersickness. In recent years DL has gained more attention from the cybersickness researchers. For instance, Qu et al. [58] used an LSTM-based attention network for detecting cybersickness using bio-physiological signals. On the other hand, Jin et al. [30] utilized three DL/ML algorithms: Convolutional neural network (CNN), LSTM, and support vector machine (SVM) for cybersickness detection. Similarly, Padmanaban et al. [52] used depth and optical flow features from the VR video data to predict cybersickness. In contrast, the authors in [40] used a 3D-CNN and a multi-modal deep fusion approach with optical-flow, disparity, and saliency features and reported a better cybersickness detection accuracy compared to Padmanaban et al. in [52]. Consequently, in [25], the authors used an LSTM model to classify the cybersickness severity from users' physiological signals (e.g., HR, GSR, etc.). In addition, a deep fusion approach was presented in [24] for classifying cybersickness severity to forecast cybersickness from integrated sensors in HMD data. On the other hand, Jeong et al. [26] used attention-based DL models for detecting cybersickness from multimodal sensor data. Even though DL models have shown tremendous success in cybersickness detection, there is a significant research gap in applying XAI to explain cybersickness models. Indeed, understanding the feature importance in the cybersickness models and explaining why samples are being correctly vs. incorrectly labeled as cybersickness is an important step towards applying the proper mitigation technique. Furthermore, applying XAI in cybersickness models can significantly improve the model's *trustworthiness* and provide insight into why and how the DL model arrived at a specific decision. In our context, the term trustworthy refers to using XAI for cybersickness explanation through a set of mechanisms, such as global and local explanations and explainable layers, to make the model transparent, understandable, and therefore, trusted by users [73]. Very recently, Kundu et al. [38] used inherently interpretable ML-based models for explaining cybersickness. They used explainable boosting Machine (EBM), decision tree (DT), and logistic regression (LR) to detect and explain the cybersickness from a user's bio-physiological and subjective measurement signals. But unfortunately, their proposed approach cannot explain DL-based cybersickness detection models.

Figure 1: Overview of LiteVR for cybersickness (CS) detection, explanation, and dimensionality reduction.

In contrast, we use a post-hoc explanation method based on SHAP to explain DL cybersickness detection models in our work. Another limitation in [38] is that their ML models are limited to binary classification (cybersickness vs. no cybersickness). In contrast, our work considers a multi-class classification problem (none, low, medium, and high) for cybersickness detection.

There exist several works in dimensionality reduction of cybersickness models. Several researchers used PCA-based methods to extract important features and reduce the model size [34, 36, 45, 63, 68, 70] for this purpose. For instance, authors in [41] used PCA to extract the cybersickness-related features from EEG signals. Then they used 3 ML/DL models: linear regression (LR), SVM, and self-organizing neural fuzzy inference network (SOFIN) to predict the user's level of cybersickness. Similarly, Kottaimalai et al. in [37] also used PCA to find the patterns from the EEG signals and neural networks (NN) to classify the cognitive tasks using the Colorado University EEG signal dataset. In contrast, Singla et al. in [68] used PCA to reduce the set of questions from the SSQ simulation. On the other hand, Mawalid et al. in [46] used time domain feature extraction based on the statistical features (e.g., mean, variation, standard deviation, number of peaks) and power percentage band to understand the cybersickness features and then applied K-Nearest Neighbor and Naive Bayes classifiers to classify cybersickness. However, PCA-based dimension reduction is not always trustworthy. This is because PCA maps high-dimensional data to low-dimensional space through projections, which often cause the loss of information from the original data [29]. In contrast to that, we use a *trustworthy*, i.e., XAI-based approach for dimensionality reduction of cybersickness detection models.

3 PROPOSED LITEVR FRAMEWORK

An overview of the proposed LiteVR framework for VR cybersickness, detection, explanation, and model reduction is shown in Figure 1. First, the training data [24] is used to train the DL models (i.e., LSTM, GRU, and MLP) for both classification and regression tasks [1]. Next, the trained DL models detect cybersickness on the test dataset by classifying them into four classes– none, low, medium, and high. The trained DL models can also forecast the next value of the ongoing cybersickness using the FMS score range of $0-10$. Next, we use the state-of-the-art tool SHAP to explain the cybersickness features, which can provide both local and global explanations. Global explanation identifies features crucial for the overall prediction, and local explanation identifies features dominating an individual sample prediction. After explaining the cybersickness features, we use the global explanation to identify and rank the dominating features. Finally, in the model reduction phase, we retrain the DL models with only the most dominating features (in our case, only the top 30%) to reduce the cybersickness models. This results in a significantly smaller model with fewer trainable parameters, hence a faster training and inference time.

---

[1]Dataset [24]: https://tinyurl.com/2p92p45h

3.1 Cybersickness classification

The network architecture of the LSTM, GRU, and MLP-based cybersickness classification model is shown in Tables 5.3 , 2, and 3, respectively. The LSTM and GRU models have 6 layers, whereas the MLP model has 8 layers. We chose LSTM and GRU models as they can capture datasets' temporal and spatial features. For both the LSTM and GRU models, we set a timestep of 60 and have 43 features. The first LSTM layer had a recurrent dropout of 20%, followed by a same-dropout layer of 20% to deal with overfitting. In the next LSTM layer, we reduce the dropout rate by up to 5% because a high dropout rate can slow the model's convergence rate and often harm final performance [79]. On the other hand, for the GRU model, all of the GRU layers used a recurrent dropout of 30% to reduce overfitting. For both the LSTM and GRU models, we used 'ReLu' as the activation function for the fifth dense layer. Finally, we used 'softmax' as the activation function in the last dense layer, which contains four outputs for the four cybersickness classes. Since all the 8 layers in MLP are dense, we do not have the timesteps like LSTM and GRU models. We use 'ReLu' in each dense layer as an activation function with a recursive dropout of 25%. Like GRU and LSTM models, the last dense layer of MLP has the softmax activation function to identify the four cybersickness severity classes. We use categorical cross-entropy as the loss function in our DL classification models as shown in eqn.1.

$$\mathscr{L}(y,\hat{y}) = -\sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij}\log(\hat{y}_{ij})) \tag{1}$$

where $\mathscr{L}$ denoted the loss function, $\hat{y}_{ij}$ is the predicted cybersickness severity class label, and (y_{ij}) is the actual cybersickness severity class label. Furthermore, p is the number of cybersickness classes, and q denotes the classifier's total number of training samples.

3.2 Cybersickness regression

The cybersickness regression task can be defined as follows: Given a history of observed VR data (e.g., eye-tracking data, head-tracking data, etc.) and the FMS score at previous time steps $t-1$, predict the FMS score at the next time steps t. Let us denote the cybersickness FMS score at time t as CS_t. For instance, if we predict the cybersickness FMS score at time $t = 10$ seconds, then $CS_t \rightarrow [S_{t-9}, S_{t-8}, S_{t-7}, S_{t-6} \ldots, S_t]$, where S denotes the user's cognitive state. We use the same three DL models that we used for classification for the cybersickness regression task as well except the loss function (RMSE for regression), i.e., to forecast the cybersickness FMS score in the range of 0 to 10. For the regression models, we use root mean squared error (RMSE) as the loss function as defined in Equation 2.

$$RMSE = \sqrt{\frac{1}{|N|}\sum_{y\in S}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2} \tag{2}$$

where y_t is the input, e.g., verbally reported FMS score from the user at time t, and $\hat{y}_t$ is the predicted FMS score by the model at that time. Consequently, N and S represent the total number of samples and time steps.

Algorithm 1: SHAP-based cybersickness feature reduction

Inputs	: Cybersickness DL Model θ,
	Training dataset X_{Train},
	Testing dataset X_{Test}
Outputs	: $F_{rank} = \{\ \}$

1: **for each** feature $i \in X_{train}$ **do**
2:     $Ex_{val} = \text{ModelExplainer}(\theta, X_{Train})$
3:     $F_{imp} = \text{SHAPval}(Ex_{val}, X_{Test})$
4:     $F_{rank} = F_{rank} \bigcup F_{imp}$
5: **end for**
6: **return** F_{rank}

3.3 Explaining DL models with SHAP

In this work, we use the SHAP-based post-hoc explanation technique for explaining the cybersickness outcomes. SHAP is a feature importance explanation approach that works by assigning a feature significance value to each prediction. It is based on the mathematical foundation of Shapley values from cooperative game theory [44]. For a given set of input samples (e.g., eye-tracking data, head-tracking data, etc.) and DL models (e.g., LSTM, GRU, and MLP), the goal of SHAP is to explain the prediction of input samples by calculating the contribution of each feature to the prediction. In other words, SHAP values explain how a given feature value increases or decreases a model's prediction. For example, given a cybersickness prediction model that predicts whether a sample has cybersickness, the SHAP explanation allows us to know to what extent a feature drove a given prediction. Such explanations can be global or local. The overall feature importance ranking (global explanation) of cybersickness can be visualized in terms of bar graphs in SHAP. On the contrary, for the local explanation, each sample is randomly chosen from the test dataset, which contains all the features and determines which features increase or reduce the likelihood of cybersickness. We use the *Deep SHAP* method [67] for computing SHAP values from the DL-based cybersickness models. After that, these SHAP values are used to calculate the feature's importance during cybersickness severity prediction.

3.4 Feature reduction through SHAP-based explanation

Algorithm 1 shows the overall method of feature reduction using SHAP. The algorithm takes a cybersickness detection model, training and testing datasets as input, and returns the feature ranking. For each feature in the dataset, given the model θ, the algorithm uses the `ModelExplainer()` function to obtain the expected value Ex_{val} (Lines 1-2). The `ModelExplainer()` function approximates the conditional expectations of SHAP values using the X_{Train} samples [44]. Next, the algorithm calculates the feature importance score F_{imp} for each feature using the `SHAPvalue()` function (Line 3), which is then appended to the set F_{rank} (Line 4). The `SHAPvalue()` function calculates the F_{imp} value through calculating the mean average value for each feature. Finally, when this process is complete for all the features, the algorithm returns the set F_{rank} (Line 6). Based on the ranked features F_{rank}, we sort (in descending order, i.e., the top feature means the most important feature) them based on their importance and use only a portion of the important/top features for retraining the model. For instance, in our case, we only considered the top $1/3$ of the features in the ranked list to retrain the cybersickness DL models. This results in a lighter model with significantly fewer trainable parameters and, thus, a faster training and inference time.

4 DATASET & EXPERIMENTAL SETUP

This section explains the experimental setup and data we used to validate our proposed LiteVR framework. We used Scikit-Learn [55] and TensorFlow-2.4 [65] for training and evaluating our DL models. For explaining the DL models, we used the SHAP [44] library. The DL models were trained on an Intel Core i9 Processor and 32GB RAM option with NVIDIA GeForce RTX 3080 Ti GPU.

4.1 Hyper-Parameter

We used the Adam optimizer to optimize our DL models with epochs of 300 and a batch size of 256. For training the DL models, we used the learning rate of 0.001. To prevent the model from overfitting, we deployed an early-stopping strategy with a patience value of 30 while training the DL models. We used a 10 fold cross-validation technique to train and test the performance of the DL models similar to [4] in which the dataset is partitioned into k groups (i.e., in our case $k = 10$). Only one partition out of k is utilized for testing the model, while the remaining partitions are used for training. The method is repeated k times, each time picking a new test partition and the remaining $(k-1)$ partition as a training dataset to eliminate bias.

4.2 Dataset

We used the integrated sensor dataset to validate the effectiveness of the proposed DL models. The integrated sensor dataset [24] contains the eye tracking, head tracking, and physiological signals for 30 participants immersed in 5 different VR simulations: Beach City, Road Side, Roller Coaster, SeaVoyage, and Furniture Shop. Eye tracking, head tracking, and physiological data consist of different subcategories. For instance, in the eye tracking data, the subcategories are Pupil Diameter (left), Pupil Position (x, y, z), Gaze Direction (x, y, z), Convergence Distance, and % of Eye Openness, and for the head tracking data Quaternion Rotation of X, Y, Z, and W axis, respectively. Similarly, for the physiological signals, the subcategories are electrodermal activity (EDA) and HR measurements. This dataset has a total of 20104 samples recorded with a maximum of 7 minutes of VR simulation. In addition, as mentioned earlier, the dataset contains four different cybersickness severity classes: none, low, medium, and high, and the FMS score ranges from 0 to 10, which is used for regression analysis. We used the 70% samples from this dataset for training the DL models and their remaining 30% samples for testing.

4.3 Performance Metrics

The standard quality metrics such as accuracy, recall, precision, and F1-score are used to evaluate the performance of DL models (e.g., LSTM, GRU, and GRU) for cybersickness severity classification [15]. Similarly, standard loss functions like mean squared error (MSE), root-mean-squared error (RMSE), mean absolute error (MAE), and R^2 score are used to evaluate the cybersickness regression model's performance [10]. For example, if y_t and $\hat{y}_t$ denote the actual and predicted cybersickness of a candidate at time t, respectively, then the MAE can be defined as:

$$MAE = \frac{1}{|N|} \sum_{y \in S} \sum_{t=1}^{N} (y_t - \hat{y}_t)$$

where N and S represent the total number of samples and time steps of all trajectories. The smaller the MAE, the better the regression model. If $\sum(y_t - \hat{y}_t)^2$ and $\sum(y_t - \bar{y})^2$ represent the Sum Squared Regression (SSR) and the Total Sum of Squares (SST) then, the R^2 score can be written as:

$$R^2 = 1 - \frac{\sum(y_t - \hat{y}_t)^2}{\sum(y_t - \bar{y})^2}$$

Table 1: LSTM model architecture (with all features)

Layer	Type	Output shape	#Param	Dropout	Activation
1	LSTM	128	66560	0.2	-
2	Dropout	128	0	0.2	-
3	LSTM	64	49408	0.15	-
4	Dropout	64	0	0.15	-
5	Dense	16	3136	-	ReLU
6	Dense	4	68	-	Softmax
Total no. of parameter:			**119, 172**		

Table 2: GRU model architecture (with all features)

Layer	Type	Output shape	#Param	Dropout	Activation
1	GRU	32	3360	0.3	-
2	Dropout	32	0	0.3	-
3	GRU	64	18816	0.3	-
4	Dropout	64	0	0.3	-
5	Dense	16	74496	-	ReLU
6	Dense	4	516	-	Softmax
Total no. of param:			**97,188**		

Table 3: MLP model architecture (with all features)

Layer	Type	Output shape	#Param	Dropout	Activation
1	Dense	128	5504	0.25	ReLU
2	Dropout	128	0	0.25	
3	Dense	64	8256	0.25	ReLU
4	Dropout	64		0.25	-
5	Dense	32	2080	0.25	ReLU
6	Dropout	32	0	0.25	-
7	Dense	16	1056	-	ReLU
8	Dense	4	132	-	Softmax
Total no. of param:			**17,028**		

Table 4: Training and inference time of: non-reduced DL models (all features) vs. reduced DL models (reduced features)

Models	Before XAI		After XAI	
	Training(s)	Inference(s)	Training(s)	Inference(s)
LSTM	2076.21	4.05	376.37	1.07
GRU	3236.21	4.94	774.29	1.32
MLP	51.25	0.15	21.55	0.09

Consequently, a low R^2 value indicates that the regression model does not adequately capture the output variance.

5 RESULTS

This section presents the results of cybersickness detection, explanation, and reduction.

5.1 Cybersickness classification and regression performance with all features

Before presenting the results of our LiteVR approach for cybersickness explanation and model reduction, in this section, we first present the details of the model development with their important statistics, cybersickness classification and regression accuracy. This will help us compare our approach's effectiveness in the following subsection of the paper.

5.1.1 Cybersickness detection model development

Tables 5.3, 2, 3 show the network architecture and number of trainable parameters for the LSTM, GRU, and MLP based cybersickness detection models. We observe that the ML, LSTM, and GRU model has total $17,028$, $119,172$, and $97,188$ trainable parameters. The MLP model has 7 times and 5.7 times less number of trainable parameters when compared to the LSTM and GRU models. These models' training and inference times are reported in Table 4. Training the MLP, LSTM, and GRU models (with all the features) requires 51.25, 2076.21, and 3236.21 seconds, respectively. Indeed, training the MLP model is 232.33 and 63.14 times faster than the GRU and LSTM models. Similarly, the inference time required for the MLP, LSTM, and GRU models are 0.15, 4.05, and 4.94 seconds, respectively. Indeed the inference on MLP is 27 and 33 times faster than the LSTM and GRU models, respectively.

5.1.2 Cybersickness classification and regression with all features

Table 5 summarizes the accuracy, precision, recall, and F-1 scores of cybersickness severity classification using LSTM, GRU, and MLP models. The cybersickness classification using the LSTM and GRU model resulted in an accuracy of 92% and 90%, whereas the accuracy of MLP is only 79%. Indeed the overall performance of the LSTM model is slightly better than the GRU model in terms of precision, recall, and F1-score for cybersickness severity classification. Furthermore, Table 6 summarizes the MSE, RMSE, R^2, and MAE of cybersickness regression using LSTM, GRU and MLP models. Like cybersickness classification, regression using the LSTM outperforms the GRU and MLP models. Like classification, the MLP also performs poorly for regression task.

5.2 Feature space reduction using SHAP-based cybersickness explanation

In this section, we apply the SHAP-based post-hoc explanation method to explain the DL model's outcome. These explanations, specifically the global explanation, are then used to identify the dominating features to aid in reducing the cybersickness DL models.

5.2.1 Cybersickness Severity Global Explanation

The overall feature importance for cybersickness severity classification using LSTM, GRU, and MLP models with all features is visualized in Figure 5. A shapely value is used to calculate the ranking of the most important features contributing to the cybersickness severity classification, with important features at the top and the least important ones at the bottom. The features with large absolute Shapley values are classed as important as they have a higher average impact on the model output. From Figure 5a and b, we observe that features such as *NrmRightEyeOriginZ*, *NrmLeftEyeOriginY*, *NrmRightEyeOriginY*, *NrmLeftEyeOriginZ* , *GazeOriginWrldSpc_Y*, etc., are the most dominating features in cybersickness severity classification for LSTM and GRU models, which corresponds to the eye tracking features. Interestingly, for both LSTM and GRU models, the eye-tracking features have a much stronger influence on the cybersickness severity classification than head-tracking features. The reason is that eye tracking measurement data contains insightful information such as gaze behavior, the position of the pupil, and the type of blink of the user for tracking the user's activity [24]. Therefore, these eye-tracking features can influence the level of cybersickness. On the other hand, features such as *HeadQRotationY*, *GazeOriginWrldSpc˙Y*, *GazeOriginWrldSpc_Z*, *HeadEulX*, *EDA*, etc., appear to be the most predictive features in cybersickness classification for the MLP model (Figure 5c). The MLP model has many false positives and negatives regarding cybersickness severity classification. However, the cybersickness severity classification using the MLP model has quite a low accuracy compared to LSTM and GRU models, as discussed in Section 5.1.2. Therefore, there is an incorrect ranking of the features such as *HeadQRotationY* at the top of the features list in Figure 5c. Previous work [8, 24, 43, 76] has suggested that eye tracking features are the most influential features in cybersickness severity classification as compared to head tracking features. Such cybersickness explanation about misclassification gives great insights into classification findings and increases trust in the model's prediction.

5.2.2 Cybersickness Severity Local Explanation

The individual samples are taken from the model randomly to explain their outcome for cybersickness local explanation. The results

Table 5: Performance of 10-Fold Cross Validation on Cybersickness Severity Classification (non-reduced DL models)

Model	Accuracy%	Precision%				Recall %				F1-score%			
		None	Low	Medium	High	None	Low	Medium	High	None	Low	Medium	High
LSTM	92	96	86	87	90	92	95	86	93	95	92	83	90
GRU	90	95	89	81	83	94	90	79	87	95	90	81	86
MLP	79	83	79	80	75	81	85	72	78	82	77	70	75

(a)

(b)

Figure 2: The local explanation of LSTM-based cybersickness classification for the (**a**) explanation for high cybersickness severity, (**b**) explanation for none cybersickness severity.

(a)

(b)

Figure 3: The local explanation of GRU-based cybersickness classification for (**a**) explanation for high cybersickness severity, (**b**) explanation for none cybersickness severity.

Table 6: Cybersickness regression using non-reduced DL models (with all features)

Regression Models	MSE	RMSE	R^2	MAE
LSTM	0.24	0.36	90	0.21
GRU	0.29	0.41	86	0.24
MLP	0.42	0.57	73	0.40

of the local explanation utilizing SHAP for the LSTM model are shown in Figure 2. Figure 2a illustrates the high cybersickness severity classification. In that figure, the red-colored line shows cybersickness probabilities for that individual outcome, the *x*-axis represents the model's output as log odds (the probabilities of feature importance in prediction), and the *y*-axis lists the model's features. It is observed that the eye tracking feature *NrmLeftEyeOriginY* is the most influential feature for high cybersickness severity classification, which has the highest probabilities of nearly 1.0. Most of

the features except *HeadQRotationY* have the highest probabilities of outcome belonging to eye tracking features; thus, an appropriate decision is established for high cybersickness severity classification. For example, in Figure 2b, none cybersickness severity classification has low probabilities score for eye tracking features, in which *HeadQRotationW* is the most influential feature with probabilities of 0.4 and *NrmRightEyeOriginZ* is the least influential feature with probabilities of 0. As a result, the correct classification is made for none cybersickness severity class. Similarly, the local explanation of the classified cybersickness using the GRU model is shown in Figure 3. In Figure 3 a and Figure 3 b, we observe that most features contribute to a high cybersickness severity class corresponding to the eye-tracking features and for none cybersickness severity class the dominating features are mostly from head-tracking features i.e., *HeadEulZ*, *HeadEulY*, *CLeftPupilPosInSensorY*, *Convergence˙distance*, etc.,. Likewise, the local explanation of the classified cybersickness using the MLP model is shown in Figure 4. From Figure 4a, it is observed that the most influential features for the

Figure 4: The local explanation of MLP-based cybersickness classification for the (**a**) explanation for high cybersickness severity, (**b**) explanation for low cybersickness severity.

Figure 5: Overall feature importance using global explanation using SHAP for (**a**) LSTM model (**b**) GRU model (**c**) MLP model

high cybersickness severity class belong to eye-tracking features such as *GazeOriginWrldSpc_Y*, *NrmLeftEyeOriginX*, etc. However, the features, such as *EDA*, *HeadQRotationY*, *HeadEulX*, etc., have significant importance in high cybersickness severity classification. This observation is misguiding since we know eye-tracking features are more critical in cybersickness prediction. However, this explanation also shows why the accuracy of the cybersickness classification using the MLP model is lower than that of the LSTM and GRU models. Consequently, Figure 4b shows that most of the features that influence the positive outcome (none cybersickness severity class) belong to head tracking and physiological signal features again. The *HeadEulZ*, *HeadEulZ*, *HeadQRotationW*, *HR*, etc., are the most predictive features for none cybersickness severity class. Such feature identification provides insights into the classification results and builds trust in the model outcome to make appropriate decisions.

5.3 Model reduction for cybersickness detection

This section presents the details of feature selection and cybersickness model reduction. Using the (global) explanation results described in the previous section, we first identify the top 1/3 of the features based on the SHAP explanation in Algorithm 1. Once the features are selected we retrain the DL models based on the selected features. We chose 1/3 of the dominating features to develop a reduced model from the base model because we were able to achieve a similar detection accuracy using this ratio. Note for different datasets, this ratio can be different. It is also worth mentioning that we use the explanation results for GRU and LSTM models as they have excellent accuracies compared to the MLP model. Table 7 shows the selected features through SHAP explanation for retraining the LSTM, GRU, and MLP models. Table 8, Table 9, and Table 10

show the network architectures of the retrained LSTM, GRU, and MLP models. We observe that the explanation-based model reduction reduces their size significantly. For instance, after reduction, the LSTM model has a total of $19,108$ trainable parameters, which is 6.26X times less compared to its non-reduced version in . Similarly, the total number of trainable parameters for the GRU and MLP models are $24,388$ and $4,020$ after model reduction, which is 4X, and 4.25X less compared to their non-reduced version in Tables 2 and 3, respectively. The reduced models' training and inference times are reported in Table 4. We observe that the training time of the reduced MLP, LSTM, and GRU models are 21.25, 376.37, and 774.29 seconds, which is 2.4X, 5.6X, and 4.2X faster than their non-reduced versions. Similarly, their inference time is also significantly faster after reduction. For instance, the inference time for the reduced MLP, LSTM, and GRU models are 0.09, 1.07, 1.32 seconds, which is 1.66X, 3.78X, and 3.72X faster than their non-reduced versions. In the following subsections, we describe the training and inference performance of the reduced cybersickness DL models.

5.3.1 Cybersickness classification with reduced DL models

Table 11 summarizes the accuracy, precision, recall, and F-1 scores of cybersickness classification using the reduced order LSTM, GRU, and MLP models. For instance, cybersickness classification using the reduced LSTM exhibits 94% accuracy, which is also slightly higher than the accuracy of the non-reduced LSTM model (see Table 5). In addition, the cybersickness classification accuracy of the reduced MLP and GRU models also increased by 11.5% and 9.7% compared to their non-reduced versions. Furthermore, other performance metrics, such as precision, recall, and F1-score for the none, low, medium, and high cybersickness classes, slightly increased for the reduced LSTM, GRU, and MLP models compared to their

Table 7: List of selected features for retraining the LSTM, GRU and MLP models through SHAP explanation

Measurement Type	Selected Features from SHAP Explanation
Eye-Tracking Data	NrmRightEyeOriginZ; NrmLeftEyeOriginY; NrmRightEyeOriginX; NrmRightEyeOriginY; NrmLeftEyeOriginZ; NrmLeftEyeOriginX; NrmSRLeftEyeGazeDirX; NrmSRLeftEyeGazeDirY; NrmSRRightEyeGazeDirY; GazeOriginWrldSpc_Y; GazeOriginWrldSpc_Z; GazeDirectionWrldSpc_Z; RightPupilDiameter; GazeOriginLclSpc_Z
Head-Tracking Data	HeadQRotationW; HeadQRotationY; HeadEulX; HeadEulZ

Table 8: Reduced LSTM model architecture (with reduced features)

Layer	Type	Output shape	#Param	Dropout	Activation
1	LSTM	64	16896	0.15	-
2	Dropout	64	0	0.15	
5	Dense	32	2080	-	ReLU
6	Dense	4	132	-	Softmax
Total no. of param:			19,108		

Table 9: Reduced GRU model architecture (with reduced features)

Layer	Type	Output shape	#Param	Dropout	Activation
1	GRU	32	3360	0.2	-
2	Dropout	32	0	0.2	
3	GRU	64	18816	0.2	-
4	Dropout	64	0	0.2	-
5	Dense	32	2080	-	ReLU
6	Dense	4	132	-	Softmax
Total no. of param:			24,388		

Table 10: Reduced MLP model architecture (with reduced features)

Layer	Type	Output shape	#Param	Dropout	Activation
1	Dense	64	1344	0.25	ReLU
2	Dropout	64	0	0.25	
3	Dense	32	2080	0.25	ReLU
4	Dropout	32		0.25	-
5	Dense	16	528	-	ReLU
6	Dense	4	68	-	Softmax
Total no. of param:			4,020		

value of 0.30, which is 16.8% less than its non-reduced version. The LSTM and GRU models performed great in both cybersickness detection and regression, whereas MLP performed poorly. It is not surprising that the MLP performs poorly in classification and regression since MLP is incapable of remembering the past sequence compared to LSTM and GRU. The accuracy of our LSTM models outperforms several state-of-the-art works in DL-based cybersickness detection. For instance, Islam et al. [24] employed the deep fusion model to classify the severity of cybersickness with an accuracy of 87.7% and reported an RMSE value of 0.51 using eye and head tracking data. The same authors in [23] used a deep temporal convolutional network (DeepTCN) to forecast the cybersickness FMS score (on a scale from 0–10) with an RMSE value of 0.49, based on eye tracking, heart rate, and galvanic skin response data. Our LSTM model's classification and regression accuracy outperform these works. There also exist other works which are relevant to our work. For instance, Qu et al. [58], Islam et. al [25], Kim et al. [35], and Jeong et al. [28] reported cybersickness detection accuracy of 96.85%, 97.44%, 89.16%, and 94.02%, respectively, using physiological and EEG/ECG signals.

Even though there is tons of work in cybersickness detection methods, to date, only a few studies have been conducted on identifying the causes of cybersickness [21, 24, 35, 52]. However, to the best of our knowledge, to date the exists no prior work on applying XAI to explain cybersickness DL models and to reduce their dimensions. Indeed, XAI-based explanations can help researchers understand the reasons behind correct and incorrect cybersickness classification and can be further utilized to develop effective cybersickness reduction methods. On the other hand, using complex DL models for real-time cybersickness detection is computationally intensive and may be impractical for integration into energy-constrained standalone VR HMDs [5]. Therefore, we believe that the proposed XAI-based cybersickness model reduction can help deploy those models on standalone HMDs.

non-reduced versions. For instance, the precision score for the none, low, medium, and high cybersickness classes for the reduced MLP are 97%, 88%, 90%, and 92%, which is slightly better than the non-reduced LSTM model. Likewise, the recall score for the none, low, medium, and high cybersickness classes for the reduced MLP model are 93%, 91%, 79%, and 85%, which is approximately 1.1X, 1.06X, 1.09X, and 1.15X higher than the non-reduced MLP model. These results suggest that we significantly improve cybersickness classification with reduced models in all cases.

5.3.2 Cybersickness regression with reduced DL models

Table 12 shows the performance of cybersickness regression using the reduced order LSTM, GRU, and MLP models. To illustrate, the MSE, RMSE, MAE, and (R^2) values for the reduced LSTM and GRU models are 0.18, 0.30, 0.17 and 0.92 and 0.24, 0.36, 0.20 and 0.88, respectively. Regarding the reduced MLP, we observe a significant improvement in the MSE, RMSE, R^2, and MAE. This is because MLP works more effectively with a small feature set. For example, we observe a decrease of RMSE, MSE, and MAE by approximately 14%, 17%, and 18%, and an increase of R^2 by 7.6% for the reduced MLP compared to its non-reduced version. Such improvements are evident in all reduced-order DL models for cybersickness regression.

6 DISCUSSION

The SHAP based global explanation reveals that features such as *normal eye origin*, *gaze origin*, *pupil diameter*, etc., are the most influential features for causing cybersickness. On the contrary, the SHAP-based local explanations of specific predictions offered useful insight for each sample, which helps explain misclassification scenarios. Consequently, we ranked the features' importance from the global explanation using SHAP, and important features were used to retrain DL models. Our results suggest that the SHAP-guided reduced DL models result in significantly faster training and inference time without sacrificing accuracy. For instance, the reduced LSTM cybersickness model, which was trained with only 1/3 of the features compared to its non-reduced version, classified the cybersickness severity with an accuracy of 94%. Similarly, while regressing cybersickness, the reduced LSTM obtained an RMSE

7 LIMITATION AND FUTURE WORKS

Although the proposed XAI-based DL-based cybersickness detection and reduction, specifically LSTM, outperformed the previous state-of-the-art cybersickness detection models, our approach has a few limitations. For instance, even though we demonstrated the effectiveness of our proposed XAI-based model reduction method with a fast training and inference time, we did not deploy them on an actual VR headset. We plan to address this limitation in the future. Furthermore, cybersickness might affect different people based on their unique characteristics, VR environment, and gender [69]. Therefore, in the future, we plan to conduct further research with people from broader demographic backgrounds and of equal gender represen-

Table 11: Performance of 10-Fold Cross Validation on Cybersickness Severity Classification using reduced DL models (reduced features)

Models	Feature Count		Accuracy%	Precision%				Recall%				F1-score			
	Original	Reduced		None	Low	Medium	High	None	Low	Medium	High	None	Low	Medium	High
LSTM	43	18	94	97	88	90	92	94	95	88	96	94	93	87	91
GRU	43	18	93	96	91	85	89	96	92	84	90	95	93	83	88
MLP	43	18	89	91	84	82	79	93	91	79	85	92	88	80	84

Table 12: Cybersickness regression using reduced DL models (with reduced features)

Regression Models	Feature count		MSE	RMSE	R^2	MAE
LSTM	43	18	0.18	0.30	0.92	0.17
GRU	43	18	0.24	0.36	0.88	0.20
MLP	43	18	0.35	0.49	0.79	0.33

tation. Also, this work uses only eye-tracking, head-tracking, and physiological signals to detect cybersickness. In the future, we plan to investigate the effect of stereo images and stereoscopic video data to detect cybersickness with explainability. Finally, we acknowledge that we do not have access to datasets with more than 7 minutes of exposure time. We plan to develop a new dataset with a long exposure time in the future to validate our approach further.

8 CONCLUSION

This paper proposed the LiteVR framework, an XAI-based framework for cybersickness detection through lightweight DL models and explanations. Specifically, we developed three cybersickness DL models: MLP, LSTM, and GRU, for cybersickness detection and applied XAI to explain and reduce their size. We illustrated the effectiveness of our proposed method using the integrated sensor dataset with eye-tracking, head-tracking, and physiological signals. Our global explanation results revealed that eye-tracking features are the most influential for causing cybersickness, and using the local explanation, we identified misclassification instances. Furthermore, based on the XAI-based feature ranking, we significantly reduced the model's size (up to 4.3X), training time (up to 5.6X), and inference time (up o 3.8X). For instance, our reduced LSTM cybersickness model could predict cybersickness with an accuracy of 94% and RMSE of 0.30, which outperforms the state-of-art works. We believe this research will be helpful for future researchers working on cybersickness detection mitigation and deployment of these models on consumer-level VR headsets.

REFERENCES

[1] Machine–deep–ensemble learning model for classifying cybersickness caused by virtual reality immersion. *Cyberpsychology, Behavior, and Social Networking*, 24(11):729–736, 2021.

[2] M. Alcañiz, I. A. Chicchi-Giglioli, L. A. Carrasco-Ribelles, J. Marín-Morales, M. E. Minissi, G. Teruel-García, M. Sirera, and L. Abad. Eye gaze as a biomarker in the recognition of autism spectrum disorder using virtual reality and machine learning: A proof of concept for diagnosis. *Autism Research*, 15(1):131–145, 2022.

[3] M. H. Alkinani and M. R. El-Sakka. Patch-based models and algorithms for image denoising: a comparative review between patch-based images denoising methods for additive noise reduction. *EURASIP Journal on Image and Video Processing*, 2017(1):1–27, 2017.

[4] Y. Bengio and Y. Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Advances in Neural Information Processing Systems*, 16, 2003.

[5] L. Castaneda and M. Pacampara. Virtual reality in the classroom-an exploration of hardware, management, content and pedagogy. In *Society for information technology & teacher education international conference*, pp. 527–534. Association for the Advancement of Computing in Education (AACE), 2016.

[6] T. T. Chan, Y. Wang, R. H. Y. So, and J. Jia. Predicting subjective discomfort associated with lens distortion in vr headsets during vestibulo-ocular response to vr scenes. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2022. doi: 10.1109/TVCG.2022.3168190

[7] T. T. Chan, Y. Wang, R. H. Y. So, and J. Jia. Predicting subjective discomfort associated with lens distortion in vr headsets during vestibulo-ocular response to vr scenes. *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[8] E. Chang, H. T. Kim, and B. Yoo. Predicting cybersickness based on user's gaze behaviors in hmd-based virtual reality. *Journal of Computational Design and Engineering*, 8(2):728–739, 2021.

[9] U. A. Chattha, U. I. Janjua, F. Anwar, T. M. Madni, M. F. Cheema, and S. I. Janjua. Motion sickness in virtual reality: An empirical evaluation. *IEEE Access*, 8:130486–130499, 2020.

[10] D. Chicco, M. J. Warrens, and G. Jurman. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, 7:e623, 2021.

[11] M. S. Dennison, A. Z. Wisti, and M. D'Zmura. Use of physiological signals to predict cybersickness. *Displays*, 44:42–52, 2016.

[12] J. Dong, K. Ota, and M. Dong. Why vr games sickness? an empirical study of capturing and analyzing vr games head movement dataset. *IEEE MultiMedia*, pp. 1–1, 2022. doi: 10.1109/MMUL.2022.3176142

[13] J. Dong, K. Ota, and M. Dong. Why vr games sickness? an empirical study of capturing and analyzing vr games head movement dataset. *IEEE MultiMedia*, 2022.

[14] L. E. Garrido, M. Frías-Hiciano, M. Moreno-Jiménez, G. N. Cruz, Z. E. García-Batista, K. Guerra-Peña, and L. A. Medrano. Focusing on cybersickness: pervasiveness, latent trajectories, susceptibility, and effects on the virtual reality experience. *Virtual Reality*, pp. 1–25, 2022.

[15] C. Goutte and E. Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval*, pp. 345–359. Springer, 2005.

[16] J. G. Grandi, M. Ogren, and R. Kopper. An approach to designing next generation user interfaces for public-safety organizations. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 944–945. IEEE, 2019.

[17] A. Gulhane, A. Vyas, R. Mitra, R. Oruche, G. Hoefer, S. Valluripally, P. Calyam, and K. A. Hoque. Security, privacy and safety risk assessment for virtual reality learning environment applications. In *2019 16th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, pp. 1–9. IEEE, 2019.

[18] A. Hadadi, C. Guillet, J.-R. Chardonnet, M. Langovoy, Y. Wang, and J. Ovtcharova. Prediction of cybersickness in virtual environments using topological data analysis and machine learning. *Frontiers in Virtual Reality*, 3, 2022.

[19] A. Halbig, S. Babu, S. Gatter, M. Latoschik, K. Brukamp, and S. von Mammen. Opportunities and challenges of virtual reality in healthcare–a domain experts inquiry. front. *Virtual Real. 3: 837616. doi: 10.3389/frvir*, 2022.

[20] M. C. Howard and E. C. Van Zandt. A meta-analysis of the virtual reality problem: Unequal effects of virtual reality sickness across individual differences. *Virtual Reality*, 25(4):1221–1246, 2021.

[21] R. Islam. A deep learning based framework for detecting and reducing onset of cybersickness. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 559–560.

IEEE, 2020.

[22] R. Islam, S. Ang, and J. Quarles. Cybersense: A closed-loop framework to detect cybersickness severity and adaptively apply reduction techniques. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 148–155. IEEE, 2021.

[23] 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.

[24] 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 *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 31–40. IEEE, 2021.

[25] 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 *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 400–411. IEEE, 2020.

[26] D. Jeong and K. Han. Leveraging multimodal sensory information in cybersickness prediction. In *Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology*, pp. 1–2, 2022.

[27] D. Jeong, S. Yoo, and J. Yun. Cybersickness analysis with eeg using deep learning algorithms. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 827–835. IEEE, 2019.

[28] D. Jeong, S. Yoo, and J. Yun. Cybersickness analysis with eeg using deep learning algorithms. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 827–835, March 2019. doi: 10.1109/VR.2019.8798334

[29] W. Jia, M. Sun, J. Lian, and S. Hou. Feature dimensionality reduction: a review. *Complex & Intelligent Systems*, pp. 1–31, 2022.

[30] W. Jin, J. Fan, D. Gromala, and P. Pasquier. Automatic prediction of cybersickness for virtual reality games. In *2018 IEEE Games, Entertainment, Media Conference (GEM)*, pp. 1–9. IEEE, 2018.

[31] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology*, 3(3):203–220, 1993.

[32] B. Keshavarz and H. Hecht. Validating an efficient method to quantify motion sickness. *Human factors*, 53(4):415–426, 2011.

[33] B. Keshavarz, K. Peck, S. Rezaei, and B. Taati. Detecting and predicting visually induced motion sickness with physiological measures in combination with machine learning techniques. *International Journal of Psychophysiology*, 2022.

[34] H. K. Kim, J. Park, Y. Choi, and M. Choe. Virtual reality sickness questionnaire (vrsq): Motion sickness measurement index in a virtual reality environment. *Applied ergonomics*, 69:66–73, 2018.

[35] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10580–10589, 2019.

[36] Y. Y. Kim, E. N. Kim, M. J. Park, K. S. Park, H. D. Ko, and H. T. Kim. The application of biosignal feedback for reducing cybersickness from exposure to a virtual environment. *Presence: Teleoperators and Virtual Environments*, 17(1):1–16, 2008.

[37] R. Kottaimalai, M. P. Rajasekaran, V. Selvam, and B. Kannapiran. Eeg signal classification using principal component analysis with neural network in brain computer interface applications. In *2013 IEEE international conference on emerging trends in computing, communication and nanotechnology (ICECCN)*, pp. 227–231. IEEE, 2013.

[38] R. K. Kundu, R. Islam, P. Calyam, and K. A. Hoque. TruVR: Trustworthy Cybersickness Detection using Explainable Machine Learning. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 777–786. IEEE, 2022.

[39] J. J. LaViola Jr. A discussion of cybersickness in virtual environments. *ACM Sigchi Bulletin*, 32(1):47–56, 2000.

[40] T. M. Lee, J.-C. Yoon, and I.-K. Lee. Motion sickness prediction in stereoscopic videos using 3d convolutional neural networks. *IEEE transactions on visualization and computer graphics*, 25(5):1919–1927, 2019.

[41] C.-T. Lin, S.-F. Tsai, and L.-W. Ko. Eeg-based learning system for online motion sickness level estimation in a dynamic vehicle environment. *IEEE transactions on neural networks and learning systems*, 24(10):1689–1700, 2013.

[42] Y.-T. Lin, Y.-Y. Chien, H.-H. Wang, F.-C. Lin, and Y.-P. Huang. 65-3: The quantization of cybersickness level using eeg and ecg for virtual reality head-mounted display. In *SID Symposium Digest of Technical Papers*, vol. 49, pp. 862–865. Wiley Online Library, 2018.

[43] P. Lopes, N. Tian, and R. Boulic. Eye thought you were sick! exploring eye behaviors for cybersickness detection in vr. In *Motion, Interaction and Games*, pp. 1–10. 2020.

[44] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

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

[46] M. A. Mawalid, A. Z. Khoirunnisa, M. H. Purnomo, and A. D. Wibawa. Classification of eeg signal for detecting cybersickness through time domain feature extraction using naïve bayes. In *2018 International Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM)*, pp. 29–34. IEEE, 2018.

[47] M. Melo, J. Vasconcelos-Raposo, and M. Bessa. Presence and cybersickness in immersive content: Effects of content type, exposure time and gender. *Computers & Graphics*, 71:159–165, 2018.

[48] A. Munawar, Z. Li, P. Kunjam, N. Nagururu, A. S. Ding, P. Kazanzides, T. Looi, F. X. Creighton, R. H. Taylor, and M. Unberath. Virtual reality for synergistic surgical training and data generation. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 10(4):366–374, 2022.

[49] 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. *Cyberpsychology, Behavior, and Social Networking*, 25(2):135–139, 2022.

[50] H. Oh and W. Son. Cybersickness and its severity arising from virtual reality content: A comprehensive study. *Sensors*, 22(4):1314, 2022.

[51] S. Ooi, T. Tanimoto, and M. Sano. Virtual reality fire disaster training system for improving disaster awareness. In *Proceedings of the 2019 8th International Conference on Educational and Information Technology*, pp. 301–307, 2019.

[52] N. Padmanaban, T. Ruban, V. Sitzmann, A. M. Norcia, and G. Wetzstein. Towards a machine-learning approach for sickness prediction in 360 stereoscopic videos. *IEEE transactions on visualization and computer graphics*, 24(4):1594–1603, 2018.

[53] A. Panahi, R. Askari Moghadam, B. Tarvirdizadeh, and K. Madani. Simplified u-net as a deep learning intelligent medical assistive tool in glaucoma detection. *Evolutionary Intelligence*, pp. 1–12, 2022.

[54] E. S. Pane, A. Zuhriya Khoirunnisaa, A. D. Wibawa, and M. H. Purnomo. Identifying severity level of cybersickness from eeg signals using cn2 rule induction algorithm. In *2018 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, vol. 3, pp. 170–176, 2018. doi: 10.1109/ICIIBMS.2018.8549968

[55] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[56] A. S. Pillai and P. S. Mathew. Impact of virtual reality in healthcare: a review. *Virtual and augmented reality in mental health treatment*, pp. 17–31, 2019.

[57] T. Porcino, E. O. Rodrigues, F. Bernardini, D. Trevisan, and E. Clua. Identifying cybersickness causes in virtual reality games using symbolic machine learning algorithms. *Entertainment Computing*, p. 100473, 2021.

[58] C. Qu, X. Che, S. Ma, and S. Zhu. Bio-physiological-signals-based vr cybersickness detection. *CCF Transactions on Pervasive Computing and Interaction*, pp. 1–17, 2022.

[59] L. Rebenitsch and C. Owen. Review on cybersickness in applications and visual displays. *Virtual Reality*, 20(2):101–125, 2016.

[60] M. Recenti, C. Ricciardi, R. Aubonnet, I. Picone, D. Jacob, H. Á. Svansson, S. Agnarsdóttir, G. H. Karlsson, V. Baeringsdóttir, H. Pe-

tersen, et al. Toward predicting motion sickness using virtual reality and a moving platform assessing brain, muscles, and heart signals. *Frontiers in Bioengineering and Biotechnology*, 9:132, 2021.

[61] G. Reddy, C. A. Spencer, K. Durkee, B. Cox, O. Fox Cotton, S. Galbreath, S. Meyer, M. Natali, T. Seech, G. Severe-Valsaint, et al. Estimating cognitive load and cybersickness of pilots in vr simulations via unobtrusive physiological sensors. In *International Conference on Human-Computer Interaction*, pp. 251–269. Springer, 2022.

[62] P. Reyero Lobo and P. Perez. Heart rate variability for non-intrusive cybersickness detection. In *ACM International Conference on Interactive Media Experiences*, pp. 221–228, 2022.

[63] M. Satu, S. Rahman, M. Khan, M. Z. Abedin, M. S. Kaiser, M. Mahmud, et al. Towards improved detection of cognitive performance using bidirectional multilayer long-short term memory neural network. In *International Conference on Brain Informatics*, pp. 297–306. Springer, 2020.

[64] E. J. Schmitt and H. Jula. On the limitations of linear models in predicting travel times. In *2007 IEEE Intelligent Transportation Systems Conference*, pp. 830–835. IEEE, 2007.

[65] A. Sergeev and M. Del Balso. Horovod: fast and easy distributed deep learning in tensorflow. *arXiv preprint arXiv:1802.05799*, 2018.

[66] V. Sevinc and M. I. Berkman. Psychometric evaluation of simulator sickness questionnaire and its variants as a measure of cybersickness in consumer virtual environments. *Applied ergonomics*, 82:102958, 2020.

[67] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.

[68] A. Singla, S. Göring, D. Keller, R. R. Ramachandra Rao, S. Fremerey, and A. Raake. Assessment of the simulator sickness questionnaire for omnidirectional videos. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 198–206, 2021. doi: 10.1109/VR50410.2021.00041

[69] K. Stanney, C. Fidopiastis, and L. Foster. Virtual reality is sexist: but it does not have to be. *Frontiers in Robotics and AI*, 7:4, 2020.

[70] W. B. Stone III. *Psychometric evaluation of the Simulator Sickness Questionnaire as a measure of cybersickness*. PhD thesis, Iowa State University, 2017.

[71] N. Tian, P. Lopes, and R. Boulic. A review of cybersickness in head-mounted displays: raising attention to individual susceptibility. *Virtual Reality*, pp. 1–33, 2022.

[72] E. Tjoa and C. Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4793–4813, 2021. doi: 10.1109/TNNLS.2020.3027314

[73] E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. G. Zelaya, and A. Van Moorsel. The relationship between trust in ai and trustworthy machine learning technologies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 272–283, 2020.

[74] S. Valluripally, B. Frailey, B. Kruse, B. Palipatana, R. Oruche, A. Gulhane, K. A. Hoque, and P. Calyam. Detection of security and privacy attacks disrupting user immersive experience in virtual reality learning environments. *IEEE Transactions on Services Computing*, 2022.

[75] S. Valluripally, A. Gulhane, K. A. Hoque, and P. Calyam. Modeling and defense of social virtual reality attacks inducing cybersickness. *IEEE Transactions on Dependable and Secure Computing*, 2021.

[76] A. van Kasteren, K. Brunnström, J. Hedlund, and C. Snijders. Quality of experience of 360 video–subjective and eye-tracking assessment of encoding and freezing distortions. *Multimedia tools and applications*, 81(7):9771–9802, 2022.

[77] H. Ved and C. Yildirim. Detecting mental workload in virtual reality using eeg spectral data: A deep learning approach. In *2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pp. 173–178. IEEE, 2021.

[78] J. Wang, R. Shi, Z. Xiao, X. Qin, and H.-N. Liang. Resolution tradeoff in gameplay experience, performance, and simulator sickness in virtual reality games. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 542–543, 2022. doi: 10.1109/VRW55335.2022.00122

[79] S. Wang, T. Zhou, and J. Bilmes. Jumpout: improved dropout for deep neural networks with relus. In *International conference on machine learning*, pp. 6668–6676. PMLR, 2019.

[80] Y. Wang, J.-R. Chardonnet, F. Merienne, and J. Ovtcharova. Using fuzzy logic to involve individual differences for predicting cybersickness during vr navigation. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 373–381. IEEE, 2021.

[81] A. H. X. Yang, N. Kasabov, and Y. O. Cakmak. Machine learning methods for the study of cybersickness: a systematic review. *Brain Informatics*, 9(1):1–25, 2022.

[82] H. Yi, J. Hong, H. Kim, and W. Lee. Dexcontroller: Designing a vr controller with grasp-recognition for enriching natural game experience. In *25th ACM Symposium on Virtual Reality Software and Technology*, pp. 1–11, 2019.

[83] 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-laryngologica*, 125(3):280–285, 2005.

[84] D. Zielasko. Subject 001 - a detailed self-report of virtual reality induced sickness. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 165–168, 2021. doi: 10.1109/VRW52623.2021.00038

[85] D. Zielasko, A. Meißner, S. Freitag, B. Weyers, and T. W. Kuhlen. Dynamic field of view reduction related to subjective sickness measures in an hmd-based data analysis task.