Predicting Opportune Moments to Deliver Notifications in Virtual Reality

Kuan-Wen Chen
National Yang Ming Chiao Tung University
Taiwan
aa10402tw.cs07g@nctu.edu.tw

Yung-Ju Chang
National Yang Ming Chiao Tung University
Taiwan
armuro@nycu.edu.tw

Liwei Chan
National Yang Ming Chiao Tung University
Taiwan
liweichan@cs.nycu.edu.tw

ABSTRACT

Virtual reality (VR) has increasingly been used in many areas, and the need to deliver notifications in VR is also expected to increase accordingly. However, untimely interruptions could largely impact the experience in VR. Identifying opportune times to deliver notifications to users allows for notifications to be scheduled in a way that minimizes disruption. We conducted a study to investigate the use of sensor data available on an off-the-shelf VR device and additional contextual information, including current activity and engagement of users, to predict opportune moments for sending notifications using deep learning models. Our analysis shows that using mainly sensor features could achieve 72% recall, 71% precision and 0.86 area under receiver operating characteristic (AUROC); performance can be further improved to 81% recall, 82% precision, and 0.93 AUROC if information about activity and summarized user engagement is included.

CCS CONCEPTS

• **Human-centered computing** → *Virtual reality*; *Empirical studies in ubiquitous and mobile computing*.

KEYWORDS

Virtual Reality; Notifications; Interruptibility; Predictive Models;

ACM Reference Format:
Kuan-Wen Chen, Yung-Ju Chang, and Liwei Chan. 2022. Predicting Opportune Moments to Deliver Notifications in Virtual Reality. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA*. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3491102.3517529

1 INTRODUCTION

Virtual reality (VR) has gained popularity and has increasingly been used for many applications, such as in education, therapy, and entertainment [36]. Current commercial VR devices provide visual, audio, and even haptic feedback, providing an immersive experience for users. However, the feeling of presence and immersion in VR might also lead to disconnection from the real world [49], which could result in missing important messages. Being afraid of missing important information could cause stress and anxiety [68, 69]. Presenting notifications in VR allows users to stay informed about real-world events without removing their headsets, preserving the continuity and immersion of the VR experience. However, sending notifications during the use of VR, presumably when users are engaged, can cause feelings of disruption [29, 74]. After all, the timing of such interruptions can affect perceptions of disruptiveness in many contexts, such as on mobile phones [16, 17], or on desktop computers [1, 4]. VR is no exception. More importantly, it may not only annoy the user but may also lead to a decrease in task performance [5, 55]. To mitigate the disruption caused by the interruptions, deferring them to opportune moments, such as task breakpoints [34, 80] or rough boundaries during task execution [4] could largely mitigate the cost of interruption. In addition to finding factors which contribute to the impact of interruption, many recent studies have sought to predict opportune and interruptible moments for receiving notifications that leverage features including sensor data [39, 76] and contextual information such as the current activity [65, 95] and engagement [20, 64] of users. As VR is not yet prevalent in people's daily lives, only a few attempts to investigate the sending of real-life notifications in VR have been made (e.g. [24, 29, 74, 96]). To the best of our knowledge, there has not been any research work that has attempted to predict opportune moments for sending notifications in VR, which we deem important to reduce the likelihood of disrupting users during their VR experience.

In this paper, we present the first research attempt aimed at exploring the feasibility of predicting opportune moments for sending notifications in VR. We primarily focus on sensor features available on VR devices, including head-mounted displays (HMD), controllers and eye-trackers, for prediction since they offer easily accessible information open to developers and provide rich information about users. In addition, we also consider additional information that can be supplied by other parties, including metadata about VR activity at the time of data collection that can be supplied by VR content providers and information that may be collected via crowdsourcing (e.g., engagement labels provided by the crowd). Our primary objective is to examine the performance of prediction using only sensor data for prediction, as well as performance after including the aforementioned additional features. As far as sensors are concerned, since there are various sensors available on VR devices, our research questions are as follows. RQ1: *How well can sensors alone predict opportune moments for sending notifications in VR?* A follow-up sub-question is: *Which sensor type contributes the most to the prediction?* Furthermore, with the additional information

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3517529

about the VR activity and engagement information provided by users, our second research question (RQ2): *How much improvement can information about activity condition and user engagement make, respectively, in prediction performance?* Finally, for each of the explorations above, we examined whether a personalized model would outperform a general model. Thus, our RQ3 is: *Does a personalized model outperform a general model for predicting opportune moments in VR?*

We recruited 20 participants to participate in data collection. The participants underwent a total of six VR sessions composed of three different VR applications, during which we collected sensor information and used a cued retrospective method to collect engagement labels on VR sessions they had just experienced. Using these data, we built deep learning models for the prediction tasks.

Overall, our results, which are the main contribution of this paper, reveal that in predicting opportune moments for sending notifications, (1) personalized models outperformed general models; (2)using only sensor features achieved 71% precision, 72% recall, and 0.86 area under receiver operating characteristic (AUROC) in the personalized models. (3) no single sensor in the VR devices dominated, and different types of sensor were advantageous in predicting opportune moments in the different types of VR activity; and (4) with the inclusion of information about activity condition and user engagement information, the personalized model can achieve 82% precision, 81% recall, 0.93 AUROC; the general model can also achieve 70% precision, 66% recall, 0.81 AUROC.

2 RELATED WORK

2.1 Interruption and Notification Management

Many studies have suggested that interruptions during tasks have many negative effects, such as decreasing task performance and impacting an individual's emotional state [1, 5, 7, 55]. However, not all interruptions are equally unacceptable. Studies have shown that many factors, such as the type of primary task, the level of task engagement, and the timing of an interruption, could affect the disruptiveness of interruptions [34, 56, 64]. Various other studies have been conducted to reduce the impacts of interruptions; for example, scheduling interruptions at task breakpoints or transitions has been found to mitigate interruption cost [1, 4, 34, 80].

Today, since people receive numerous notifications every day from their mobile phones [45], many studies have been conducted to understand the impacts of notifications on mobile phone use, as well as mobile users' behaviors around phone notifications [10, 43, 52, 66, 75]. Because VR has gained popularity in the past decade, recent works have also explored the effect of notifications in VR. Ghosh et al. [24] investigated noticeability and perception in different VR interruption scenarios across different modalities. Zenner et al. [96] proposed a method for delivering notifications in an immersive, ambient way to preserve the immersion of VR. Rzayev et al. [74] examined the noticeability, distraction, and intrusiveness of four notification placements (Head-Up Display, On-Body, Floating, and In-Situ); they found that while showing notifications using a Head-Up Display placement decreased the response time and the number of missed notifications, it increased the noticeability, distraction, and intrusiveness of the notifications Hsieh et al. [29] investigated individuals' receptivity to message notifications delivered using three types of display during four VR activities;they found that affixing a notification to the upper left corner in the user's field of view could increase the recall rate by more than 20% compared to affixing a notification to a controller or making a notification a movable panel. All such notification research in VR has been conducted to investigate the modality, position or display design of notifications, rather than to predict the best timing for the delivery of notifications in VR.

2.2 Attention and Interruptibility Prediction

Many previous works have aimed to determine moments at which people's activities might be interruptible. Earlier works have focused on interruptibility and breakpoint prediction in desktop or workplace contexts. For example, Hudson et al. [31] found that a simple set of simulated sensors and manually coded features could construct an interruptibility prediction model with an overall accuracy of about 78%. Later, Fogarty et al. [19] further showed that sensor-based models of human interruptibility can provide robust estimates for a variety of office workers in a range of circumstances, with an accuracy of up to 79.5%. Horvitz et al. [28] utilized computer activity and users' environment to predict the cost of interruptions with an accuracy of up to 82.3%. Fogarty et al. [20] used low-level event logs to train a statistical model to differentiate interruptible situations from other situations (engaged or deeply engaged) with an overall accuracy of 71.8%. Iqbal et al. [33] utilized the characteristics of task structure to predict the costs of interruptions, and their model correctly predicted 53% of the costs of interruptions. Tanaka et al. [79] used head motion to predict interruptibility during PC work and non-PC work in office environments and obtained F-scores between 0.5 and 0.7.

In recent years, increasing attention has been given to leveraging wearable sensors and mobile devices. Kern et al. [39] showed that users' personal and social interruptibility could be determined with an accuracy of up to 91%. Ho et al. [27] used wireless accelerometers to detect postural and ambulatory activity transitions in real time with an average accuracy of 91.2%; their results showed that participants were more receptive to the delivered messages. Haapalainen et al. [26] used psycho-physiological sensors and found electrocardiogram median absolute deviation and median heat flux measurements were the most accurate at distinguishing between low and high levels of cognitive load, with an overall accuracy of 80% when used together. Zuger et al. [99] also demonstrated that using psycho-physiological sensors to classify the interruptibility of a software developer could achieve high accuracy (91.5% for a lab study and 78.6% for a field study). However, later, in a subsequent field study [100], they found that information from computer interactions was more accurate at predicting interruptibility during computer use than biometric data (74.8% vs. 68.3% accuracy) and that combining both ultimately yields the best results (75.7% accuracy).

Focusing on mobile phones, Hofte et al. [84] explored the use of context information that users provide to predict a person's availability for a phone call with an accuracy of 63.9%. Pejovic et al. [63] also showed that users' reported context information, such as activity, location, time of day, emotions, and engagement, could be leveraged to predict whether a user would respond to a notification

with a precision of 60%. However, when using such information to predict whether a user would perceive a moment as an opportune moment for receiving notifications, the precision ranged from 40% to 80%, depending on a cut-off threshold that was used for defining opportuneness. Despite the limited precision, their deployment study showed that notifications that were scheduled using their model were more favorably received, (i.e., 26.4% of them were marked as "very good" moment to interrupt, compared to 15.4% for randomly scheduled notifications.) Also aiming at predicting opportune moments but mainly based on sensor information, Poppinga et al.'s model achieved an accuracy of approximately 77% [70]. Okoshi et al. explored the use of both physical activity-based and UI event-based breakpoint detection to reduce workload perceptions caused by interruptive notifications on smart phone [58] and multi-devices mobile environments [59], with an 82.6% accuracy and 82.7% precision. Yuan et al. [95] explored the inclusion of personality traits and the use of a two-stage hierarchical interruptibility prediction model for smartphone interruption; their model achieved an overall accuracy of 66.1%. Pielot et al. [65], in contrast, aimed to predict if a user would engage with proactively recommended content, achieving a 66.6% better precision than a baseline model.

Opportune moment detection for interruption was also examined in specific contexts. For example, in the context of driving, Kim et al. [40] employed sensor and human-annotated data about drivers' states and driving situations to predict the drivers' interruptibility with an accuracy of 94%. In social context, Park et al. [61] utilized build-in sensors to detect social context and identified breakpoints for smartphone notifications; their model achieved a precision of 92.0% in a controlled social interaction setting and was measured against ground truth obtained from manually labeling captured videos.

Finally, other than opportune moment, Pielot et al. [67] predicted high attentiveness to mobile instant messages (i.e., seeing a message within a few minutes) using phone sensor data and users' interaction with mobile phones; their model achieved 70.6% overall accuracy and 81.2% precision.

To sum up, these studies indicate that using contextual information provided by sensors, software events, or self-reports has allowed researchers to build models for predicting interruptible and opportune moments for sending notifications. However, the wide range of prediction performances shown above suggests that the task of prediction can be challenging in some contexts and high accuracy cannot be guaranteed. A recent study shows that even human beings may falsely recognize interruptible moments for a person wearing an HMD simply by observing their movement and gestures [23]. Nevertheless, most opportune moment prediction studies have been conducted in a workplace or mobile setting, and no research has investigated whether opportune moment prediction in VR is feasible and can be performed effectively. After all, the sensory immersion experience and involvement of physical movement in VR applications may make the perception of opportune moments for receiving a notification different from that in a workplace or mobile environment.

In our study, we build upon prior work by using sensors in VR devices that capture users' movements and additional contextual information, including the current activity and engagement of users, to predict opportuneness for delivering notifications to users in three different types of VR activities. To the best of our knowledge, this is the first study to predict the opportune moment for sending notifications in VR settings.

3 DATA COLLECTION

To predict opportune moments for sending notifications in VR, we recruited 20 participants to experience three types of VR activities in a random order, and collected opportune and inopportune labels using a cued retrospective method, since retrospective methods with visual cues (e.g., image or video) have been proven to be effective in collecting accurate data for short-term studies [72, 73]. Specifically, participants were asked to annotate opportune and inopportune moments with the assistance of video replays that reminded them of their reactions during VR interaction. We did not use a simple in-situ prompt to collect momentary experience because much time is necessary to collect large amounts of momentary experience data to obtain sufficient data points to train a model, which could have easily caused fatigue and sickness for the study's participants with the current VR device. The participants of this study experienced three kinds of VR activities, each of which was divided into two sessions to avoid an excessively long timeframe for the study, not merely to prevent fatigue and sickness, but also to reduce bias in recalling perceived opportune moments. Each session was designed to be five minutes long, during which the participants saw four visual notifications. In the study, notification delivery was designed to create stimuli to elicit participants' feelings when seeing notifications in these VR applications. Since prior research has suggested that different modalities of notifications (e.g., visual, audio, and haptic) could impact the disruptiveness of a notifications [9, 24, 47, 52, 92], we chose to only use the modality of visual notifications in data collection, as this is most common in today's consumer VR platforms.[1] In particular, we used message notifications, which have been suggested to be the most pervasive type of notification users see in numerous notification studies (e.g., [9, 47, 52, 75, 92]). Each of the four notifications was randomly sent within a 25-second time window, and a 40-second interval was placed between any two such time windows. After each session, participants were asked to annotate their perceived opportune moment (opportune vs. inopportune) for receiving notifications using the interface we developed (Figure 3). We asked participants to label opportune moments as a binary outcome (i.e., opportune vs. inopportune) instead of rating using a scale for two reasons. First, prior research on interruptibility and opportune moments typically adopted a binary outcome as prediction targets. It was found that predicting an interruptible moment could effectively reduce the negative impact of interruption [50, 60, 86, 98]. In addition, the performance of interruptibility prediction could decrease significantly when predictions involved multiple levels of interruptibility [31, 67, 99, 100]. Finally, in addition to opportune moment, participants also described their engagement during the session. Below. we describe our study in more details.

[1] SteamVR and Oculus

3.1 VR Activity

We deemed it important to collect diverse VR experiences to make our prediction model more generalizable to different VR applications. To collect diverse VR experiences, we decided to collect data from different types of VR activities. We focused on two dimensions that are related to interruptibility. The first dimension was *body movement* because it has been suggested that moving one's body might produce high cognitive load [35], which is an important factor for interruptibility [1, 3]. The second dimension was *visual attention* because Hsieh et al. [29] found that their participants preferred not to receive visual notifications during a VR activity that required a high degree of visual attention. Therefore, we designed three kinds of VR activities that varied across these two dimensions—namely, 360 video, VR fitness, and the rhythm game—all of which are common activity types in today's consumer VR platform.[1]. In particular, to make the participants feel more engaged and perceive the latter VR activities as realistic, participants' performance were scored and displayed after the activity in both VR fitness and the rhythm game. Further details about each activity are given below.

Figure 1: Three type of VR activities in the data collection (a) Watching a 360° Video (b) Doing a workout move in VR Fitness (c) Playing a Rhythm Game

3.1.1 360° Video. 360° video is a common application of VR. Since users do not need to move when watching a 360° video, different video content might attract different levels of visual attention. We therefore considered this activity to require a low level of body movement and various levels of visual attention. We selected videos which varied in their arousal score from an open 360° video dataset [46], because an arousal level has also been linked to interruptibility [25]. The dataset provided 360° videos varying in arousal ratings from 1.57 to 7.42 on a 9-point rating scale. We sorted the videos in the dataset by their arousal scores. With diverse arousal scores, we aimed to also ensure the diversity of visual attention in the videos. To do so, we downloaded 18 pre-selected videos and trimmed each one to 50 seconds. We then asked participants to label preselected videos in pilot studies. We filtered out videos that were labeled by the participants as demanding high and low visual attention throughout the entire video because these videos featured a low degree of diversity in drawing visual attention. Eventually, this process resulted in our final selection of twelve videos: four videos from the lowest third as low-arousal (average arousal = 1.80), four videos from the middle third as medium-arousal (average arousal = 3.95), and four videos from the highest third as high-arousal (average arousal = 6.22). In each session, participants watched two low, medium, and high arousal videos, respectively, in a random order. The videos did not require participants' interaction.

3.1.2 VR Fitness. Exergaming,[2] or doing exercise in VR, has gained popularity in recent years. To collect data which featured diverse amounts of body movement with relatively low levels of visual attention, we designed an application called VR fitness. In VR fitness, participants were in a virtual fitness room and followed the workout moves from videos displayed on the wall. At the beginning of each move, participants would need to pay some degree of visual attention to watch a demonstration and learn how to correctly perform the move. Over time, we assumed that the need for visual attention would gradually decrease since participants simply needed to repeat the same moves. To ensure diversity of body movements, we designed different types and different intensities of movements, assuming that they might require different levels of body movement. Specifically, we designed two types of movements and arranged for a break after each one. The first type of movement was rotational, requiring participants to rotate their torso as well as their head (e.g., bend waist and toe touch). The second type of movement was translational, requiring participants to move their entire body (e.g., jumping jack and kick squat). We also designed two intensity levels for each movement, namely low-intensity and high-intensity—where the speed of the low-intensity workout videos was half of that of the high-intensity videos. In each session, participants performed four types of movements; each type of movement started at a low intensity for 20 seconds, followed by high-intensity ones for 35 seconds and a 15-second break time before the next movement.

3.1.3 Rhythm Game. Rhythm games are a popular genre of VR game.[3] To collect data that were diverse in both visual attention and body movement, we designed a game that refers to a popular VR rhythm game called Beat Saber.[4] In this game, participants used controllers to hit incoming beat cubes at the right time to earn points. The difficulty of the game (e.g., number of beat cubes per second, CPS) and the different stages of the game (e.g., in-game and break) might affect the level of visual attention and body movement. To ensure diversity of visual attention and body movement, we selected songs and beat maps of different difficulties adopted from [6] and arranged different game stages. Specifically, we trimmed each song to 85 seconds; in each session, we selected songs from three difficulty levels, including easy (average CPS = 1.38), normal (average CPS = 2.03), and hard (average CPS = 2.88). For each song, we arranged 10-second and 15-second breaks as in-game and between-game breaks, respectively.

3.2 Notification Display

The participants received four visual notifications in each session. As previously mentioned, the purpose of this was to allow participants to experience receiving notifications at different moments in these VR activities so that they could recall their feelings when annotating the opportuneness of the moments for receiving notifications during these sessions. To make the experiences realistic, when the notifications appeared, the system also allowed participants to respond to them through the controllers if they wanted to, including clicking the touchpad on the controller, sliding the

[2] https://en.wikipedia.org/wiki/Exergaming
[3] https://store.steampowered.com/sale/2018_so_far_top_vr_titles
[4] https://beatsaber.com/

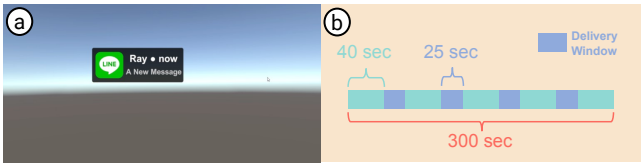


Figure 2: We presented notifications in VR activities as realistic stimuli and providing a reference for annotating interruptibility afterward: (a) Notifications were fixed in the upper region of field of view; (b) Notifications were delivered during delivery windows.

touchpad, or both, similar to the interaction experience of swiping away smartphone notifications. We did not use these inputs for the prediction task; this was because the absence of a response does not necessarily mean an inopportune moment, as participants might simply forget to respond. Instead, these inputs were intended to provide participants with a reference during annotation so that they could observe their reactions to those notifications in the moment. Each notification disappeared whenever a participant responded to it or lasted 10 seconds if the participant did not respond.

Regarding notification placement, we chose to display notifications in the upper region of a user's field of view (Figure 2 (a)), as research has indicated that notifications placed at this position would be more likely to be noticed than those attached to a controller or floating in the air [29, 74]. Given that placing notifications in a noticeable spot is also likely to cause visual disturbance [74, 83], we assumed that if participants perceived a moment as opportune when they saw notifications appear at this position, it was likely that they would also perceive the moment as opportune if the notifications had appeared in other positions.

Regarding notification content, a notification displayed a sender name and a summary "A New Message." As to the former, prior research (e.g. [43]) has suggested that the senders of notifications could affect users' receptivity. Thus, we asked participants to provide us with the names of three contacts they most frequently exchanged messages with. We used these names as the senders of the messages they received during the study to make the appearance of these notifications more realistic. As to the latter, since notification content has also been found to affect users' receptivity [53], we did not display the content but only showed "A New Message" to minimize the influence of the notification content. This short notification summary has also been incorporated in some instant messaging (IM) services, one of which is Line messenger, the most popular IM service in our country. Thus, we assumed that our participants would have been familiar with seeing such a summary.

3.3 Video Annotation of Opportune Moment and Engagement

After participants experienced each session, they took off their headsets and then labeled notifications in the VR activity they just experienced as opportune or inopportune according to their own perceptions. Specifically, each VR session was recorded through our program, and the participants used the interface we designed to

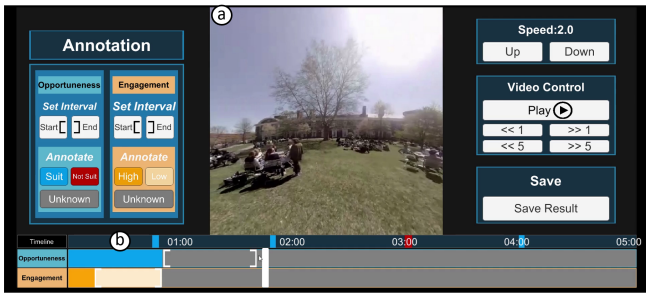


Figure 3: Interface of annotation program. Participants label the opportuneness and engagement during the activity assisted by (a) first-person-view video replay and (b) responses to notification

perform annotations (Figure 3). In addition to labeling opportuneness, the participants also labeled their engagement throughout the session, which, as mentioned earlier, was used as additional information to help the prediction task. We assumed that this information could be obtained by content providers who adopted crowdsourcing to obtain the information in advance, an increasingly common practice in processing multimedia data [8, 37, 82, 97]. Other than this, we did not use any subjective scales to provide information for the task of prediction, as we assumed that these pieces of information would not be available in real time.

During labeling, participants watched the video replay of the session they had just experienced, including the appearance of the four notifications. In the annotation interface, the participants used sliders to set time windows and label time windows as opportune or not and to describe their level of engagement. For opportuneness, participants could characterize timing as opportune, inopportune, or unknown. For engagement, they could mark the activity as low engagement, high engagement, or unknown. All of these labels as well as the language on the interface were in the participants' native language (i.e., Mandarin Chinese). Prior to annotation, the researchers also ensured that all of the participants understood these two concepts.

3.4 Participants

We recruited 20 participants to participate in the data collection. All were graduate or undergraduate students, consisting of 10 female and 10 male participants between the ages of 20–28 (M = 22.65, SD =1 .62). Four (20%) had never used a VR device before, while 11 (55%) had used VR between one and five times. (25%) had used VR more than five times.

3.5 Study Procedure

We used HTC Vive Pro Eye, which consists of two controllers and an HMD equipped with an eye tracker, as the VR device for this study. The participants were informed of the goals of data collection and then given a tutorial on HTC Vive Pro Eye and its controllers. They then wore the HMD and over-the-ear headphones, and we performed inter-pupillary distance adjustment and eye-tracking calibration. Before the main phase of the study, the participants performed a warm-up task, in which they experienced the three VR

Table 1: Features used in our study grouped by type and references to prior related works on these features.

Feature Type	Description	Reference
Sensor	VR device (HTV Vive Pro Eye)	[13, 14, 27, 39, 41, 70, 76, 88]
	HMD : velocity, angular velocity, rotation **Controller** : velocity, angular velocity **Eye-Tracker** : gaze angle, gaze-shift speed	
Activity	Predefined Categories (13 Categories)	[51, 65, 80, 90, 95, 100]
	360° video : low-arousal, mid-arousal, high-arousal **VR Fitness** : rest, low-intensity rotation, high-intensity rotation, low-intensity translation, high-intensity translation **Rhythm Game** : in-game break, ending screen, easy, normal, hard	
Engagement	User-provided label (2 levels)	[20, 43, 57, 63, 64, 100]
	Low engagement, high engagement	

activities, notifications, and the annotation program to ensure that they understood how to operate the equipment and the programs. We did not inform participants of the length and timing of the breakpoints in each of the VR activities, but expected them to learn about how long and when these would appear after the warm-up task naturally.

During formal data collection, the participants experienced the three VR activities in a random order, and each activity was split into two sessions, with each session lasting approximately five minutes long. The content of each activity (i.e., the videos, movements, and songs within the 360° video, VR fitness, and rhythm game activities, respectively) was presented in a random order. At the start of each session, the participants were asked to recalibrate the eye tracker to ensure that the collected gaze data were accurate. After each session, the participants took off the headset and annotated opportune moments and engagement for the activity they had just experienced. Then, they took part in a short interview about their perceptions of opportuneness and engagement during the activity.

3.6 Features and Data Processing

We recorded the sensor data from the VR device at 60 Hz and made note of the predefined current activity category in the programs during each session. The opportune moments and engagement labels were obtained through the participants' annotations after each session.

For the sensor data, we focused on the features that captured users' movement in VR. Previous research found that interruptibility could be predicted through velocity [27, 39, 41, 76] and the rotation of the device users wear [70, 88]; as such, we used the velocity and angular velocity of the HMD and the controllers, as well as the rotation of the HMD, as features. We also used the angle between the HMD and gaze, as well as gaze-shift speed over time, as features since previous research has found that different tasks affect eye-head dynamics [13] and gaze-shift dynamics [14].

The gaze angle and gaze-shift speed were computed as follows:

$$\theta_t = \arccos(\hat{g_t} \cdot \hat{h_t})$$

$$\omega_t = \arccos(\hat{g_t} \cdot \hat{g_{t-1}})/dt$$

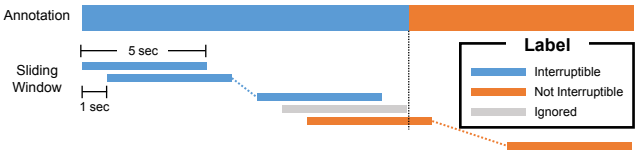


Figure 4: Sample data points from continuous labeled data using sliding window approach

The gaze angle θ_t (in radian) at time t is the vector angle between normalized gaze direction vector $\hat{g_t}$ and normalized head forward direction vector $\hat{h_t}$ at time t, where the vector angle is computed using the inverse cosine of the dot product of the normalized vectors. The gaze-shift speed ω_t (rad/sec) at time t is the angle between normalized gaze direction $\hat{g_t}$ and $\hat{g_{t-1}}$ at time t and $t-1$ divided by the difference between time t and $t-1$, which is $\frac{1}{60}$ seconds in our case. All of the above features from the sensor data were smoothened to remove noise using a moving average filter with a window size of 60 time frames (1 second).

To answer the second research question, we also included the predefined activity category and the level of engagement marked by participants as features. Note that we assumed that these types of information would be supplied by the VR content supplier and were are obtained beforehand, unlike the sensor information obtained in real time. As a result, given that the engagement label data were obtained from the participants, later, in training the model, we used summarized engagement information (i.e., considering the majority of the engagement annotation for every moment; for example, if 16 participants considered the fifth second of a roller coaster 360° video as high engagement and four participants considered it as low engagement, we considered that moment high engagement). Given that prior research has found that interruptibility is related to contextual information such as the current activity [51, 65, 80, 90, 95, 100] or engagement [20, 43, 57, 63, 64, 100] of users, we assumed that these features could improve the prediction.

As sensor data were obtained in real time, to predict opportune moments in real time, we used the sensor data five seconds prior to predict a perceived opportune moment. We chose five seconds as the threshold because it resulted in the best performance among

five candidate thresholds (1.0 s, 2.5 s, 5.0 s, 7.5 s, 10 s) in our results. Activity and engagement information was obtained in advance for every moment in the activity. Thus, we used the activity and engagement information associated with a particular "present" moment as the features.

Our labeling method gave us continuously labeled data. As a result, we used a sliding window to sample data points, where the size of the window was five seconds (300 frames) with an offset of 1 second (60 frames). For example, sensor data from the 600th frame to the 900th frame were used to predict whether the moment at the 901st frame was opportune for receiving a notification. In sampling the data points, we ignored data points that would contain noise, which were points that satisfied either of the following conditions. First, we removed data points that were at transitions between periods of opportune moments and inopportune moments (i.e., data points that were within 0.5 seconds before and after the boundary), to avoid labeling errors. Second, we removed data points associated with the period where a notification appeared because a notification itself was an interruption and its presence might have changed the participants' eye gaze and movement data. Finally, we removed data points that participants labeled as unknown for either opportune moment or engagement. This sampling method produced a dataset containing 33,219 total samples, consisting of 13,236 positive samples (suitable) and 19,983 negative samples (not suitable). We present our evaluation of the model in the next section.

4 ANALYSIS AND RESULT

4.1 Machine Learning Model

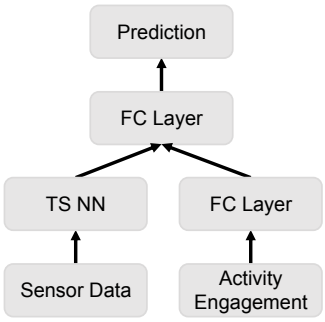
To examine which classification technique worked the best in predicting opportune moments in VR, we started with several traditional classifiers commonly used in previous interruptibility research, including Naive Bayes, Logistic Regression, SVM, and Random Forest [87]. We used scikit-learn [11], a widely used machine learning library, to implement those classifiers. Previous works have often extracted statistical features (e.g., mean, min, and max) from time-series data as features [2], and we also found that using the statistical features of time-series data was suitable for traditional classifiers with our dataset. Therefore, we extracted these pieces of information from our time-series data as features to train the classifiers mentioned above.

Next, since approaches based on neural networks (NN) have achieved success in many tasks involving time-series classification, such as anomaly detection [93], human activity recognition [44], and gaze pattern recognition [78], in comparison with the traditional classifiers, we also used a NN-based approach to process time-series sensor data. The pipeline of our model is shown in Figure 5. Specifically, sensor data (time-series data) were first fed into a time-series NN-based model to output a 64-dimensional feature vector. Additional contextual information (activity category and engagement) were categorical data, and we thus encoded them as a one-hot vector and fed them into a fully connected layer to output another 64-dimensional feature vector. Finally, these two feature vectors were concatenated and processed by another fully connected layer with a softmax activation function that performed binary classification. For the time series NN architecture that processes time-series sensor data, we initially explored two commonly used architectures for time series classification [15], including one-dimensional convolutional neural networks (1D-CNN), and a variant of the recurrent neural network named long short term memory (LSTM). Our comparison further included a more recent architecture [38]—namely, multivariate long short term memory fully convolutional network (MLSTM-FCN)—which combines the use of 1D-CNN and LSTM to achieve state-of-the-art results on many time-series classification tasks. The details of NN layers are described in Appendix A.

We implemented the NN-based models using PyTorch [62], a widely used deep-learning framework for Python. We trained the models end-to-end using standard cross-entropy loss and SGD optimizer with an initial learning rate of 10^{-1}, a momentum of 0.5, and a batch size of 64. The learning rate was scheduled to decrease to 10^{-5} over time.

For our dataset, the NN-based approach outperformed traditional classifiers mentioned above. Therefore, in the remainder of this paper, we present the results of the NN-based approach. Meanwhile, we note that all the NN-based models could achieve real-time prediction; the inference times of all the NN-based models on our 1080Ti machine were less than 10 ms.

Figure 5: Pipeline of our NN-based approach. We explore 1D-CNN, LSTM and MLSTM-FCN for processing sensor data. ***Legend: "TS NN":Time-Series Neural Net,"FC Layer": Fully Connected Layer.***



4.2 Validation Method

To investigate how predictive the features were for building a personalized model vs. a general model, we first trained personalized models for each individual participant. Since the data we collected was time-series data, we split the data into a training and a test dataset according to time interval to avoid overlapping between the two sets. For example, for 300 seconds of VR activity recording data, we took 0 to 30 seconds as test data and the remaining data as training data. Similarly to Zuger et al. [100], we applied 10-fold cross-validation for each participant, for which 10% of the data were used as testing data and the remaining 90% were used as training data. The result was averaged over 10 runs. This method provides a metric for how personalized models performed on similar but unseen data.

Next, to gather further insights on the generalizability of the features across individual participants, we performed a leave-one-participant-out cross-validation to build a general model. That is,

Table 2: Comparison of models that used sensor-only features (HMD, controller and gaze). We report accuracy, recall, precision and F1-score which are calculated under classification threshold 0.5, and AUROC for each model. *Legend: "Con.": Controller, "Prec.": Precision., "F1.": F1-Score, "AUC": AUROC*

	Personalized Model					General Model				
Sensors-only	Acc	Recall	Prec.	F1.	AUC	Acc	Recall	Prec.	F1.	AUC
Baseline	0.6103	0.2390	0.5240	0.3282	0.5000	0.6016	0.000	NaN	NaN	0.5000
1D-CNN	0.7616	0.6985	0.7018	0.7002	**0.8572**	**0.6727**	0.5308	**0.6012**	0.5638	**0.7257**
LSTM	0.7444	0.6918	0.6748	0.6832	0.8199	0.6581	0.5709	0.5710	0.5710	0.7014
MLSTM-FCN	**0.7738**	**0.7202**	**0.7144**	**0.7173**	0.8559	0.6715	**0.5880**	0.5878	**0.5879**	0.7253

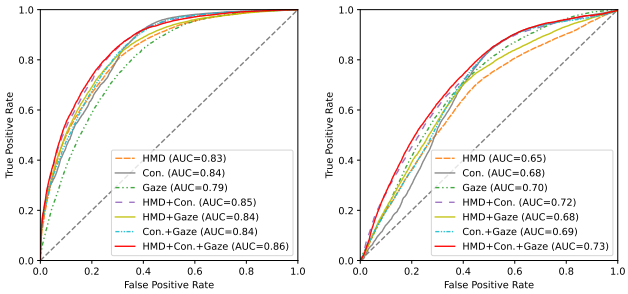
one participant's data were used as test data, and the data from the remaining 19 participants were used as the training data. The results were averaged over 20 runs. This method provides a metric of how a general model might perform for new users. To evaluate the models, we first adopted accuracy, which was the fraction of correct predictions and often used to evaluate the performance of classifiers in previous interruptibility works [87]. However, while accuracy treats all misclassified data equally, some misclassifications might be more undesirable in certain situations. In our study, the aim was to not disturb VR users with notifications; thus, a false positive (being predicted as an opportune moment but in fact being an inopportune one) was more undesirable and costly than a false negative (being predicted as an inopportune moment but in fact being an opportune one). Thus, in addition to accuracy, we also evaluated the model using metrics including recall, precision, F1-score, and the area under the receiver operating characteristic curve (AUROC). Recall measured how many of the actual opportune moments were predicted as opportune. A high recall means that most of the users' opportune moments are successfully detected. Precision measured how many of the moments being predicted as opportune were truly opportune moments. As a result, given our goal of reducing disruption, we deemed a model that achieved high precision to be more desirable than a model that achieved high recall. On the other hand, F1-score was the harmonic mean of precision and recall, providing a balance measure between precision and recall. Finally, the receiver operating characteristic (ROC) curve shows the true positive rate (TPR) and false positive rate (FPR) at different classification thresholds, which is beneficial for optimizing the model for different preferences regarding TPR and FPR. AUROC then measures the overall performance of a classification model at all classification thresholds.

4.3 Model Performance

4.3.1 Performance Comparison of Sensor-Only Models. We first examined the performance of only using sensor data for prediction tasks. The results of each model that used only sensor features are presented in Table 2 for both a personalized and a general model. For the baseline, we report the AUROC from a random classifier commonly used for this metric; for other metrics, we report the results from a majority classifier that always predicts classes containing more samples in the training dataset, often used as a baseline in works on interruptibility prediction [87].

Overall, our results showed that personalized models outperformed general models for all combinations of sensor types. This implies that individual variances in sensor data patterns among the participants might be so large so that the information learned from a group of individual participants was not sufficient to be predictive for a new user. In addition, Table 2 shows that all NN-based models performed significantly better than the baseline according to all metrics; this indicated that sensor features were helpful for predicting opportune moments for sending notifications in VR. For a personalized model, among the three NN-based models, MLSTLM-FCN achieved the best overall performance—highest accuracy (0.77), precision (0.71), recall (0.72), and F1-score (0.72), with only a slightly lower AUROC (0.86) than that of 1D-CNN. Notably, the recall of MLSTLM-FCN was noticeably higher than the other models for both the personalized model and the general model, respectively.

Figure 6: ROC curve of models that using different sensors. (Left) MLSTM-FCN personalized model. (Right) MLSTM-FCN general model



Next, we were interested in exploring which combination of sensor features best predicted the opportune moment for sending notifications in VR. Figure 6 shows an ROC curve of all combinations, of which we only displayed the performance of the MLSTLM-FCN models to maintain the readability of the figure. The detailed results for each model can be found in Appendix B.

4.3.2 Sensor Importance in Different Activities. Generally speaking, the results show that combining more types of sensor achieved better AUROC for both the personalized model and the general model and that no single sensor seemed to dominate the others. This suggests that the three types of sensor features complemented each other quite well in making predictions; perhaps they captured different aspects of participants' conditions, which were indicative of different aspects of a moment being perceived as opportune for receiving notifications in VR. Because combining three sensors yielded the best performance, when comparing the performances of the sensor-only models with the models using additional features (i.e., activity information and engagement), we used this configuration as the basis for making comparisons.

Table 3: Performance comparison of the models using all the features (sensors, activity condition, and summarized engagement). We report their accuracy, recall, precision, F1-score, and AUROC. *Legend: "Con.": Controller, "Prec.": Precision, "F1.": F1-score, "AUC": AUROC*

	Personalized Model					General Model				
All-Features	Acc	Recall	Prec.	F1.	AUC	Acc	Recall	Prec.	F1.	AUC
1DCNN	0.8503	0.8032	**0.8178**	0.8104	**0.9304**	0.7440	0.6460	0.6913	0.6679	0.8004
LSTM	0.8457	0.8049	0.8072	0.8061	0.9240	**0.7549**	**0.6644**	**0.7039**	**0.6836**	**0.8101**
MLSTM-FCN	**0.8530**	**0.8127**	0.8174	**0.8150**	0.9273	0.7212	0.6246	0.6583	0.6410	0.7724

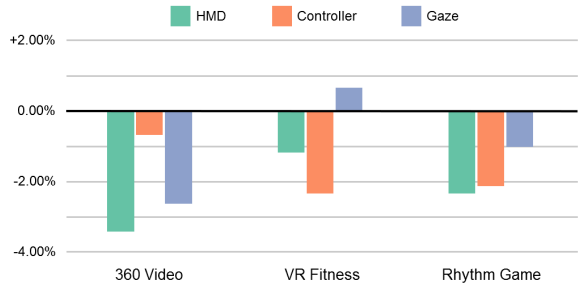
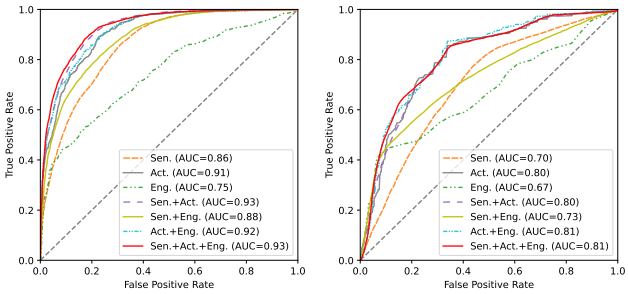


Figure 7: Feature importance of each sensor in the personalized MLSTM-FCN model for the different VR activities, quantified by mean loss of accuracy when removing the features of individual sensors.

Because each sensor seemed to capture different aspects of the participants' conditions, we were also interested in the feature importance of each type of sensor in different VR activities (360° video, VR fitness, and the rhythm game). Because we could not directly compute the feature importance in a neural network, we measured each sensor's importance by comparing the performance difference between the model trained with and without the features of that sensor type. This approach is commonly adopted to examine the necessity of certain features or modules in neural networks (e.g., [54, 94]). We particularly examined the performance difference in different activities, as these activities entailed different kinds and intensities of eye and physical movement. Figure 7 shows the decline in accuracy when removing the features of each sensor type in each activity. The figure clearly shows that removing each of them resulted in differing performance drops in each activity, suggesting that each sensor contributed differently to prediction in the different types of VR activity. For instance, the HMD and gaze sensors seemed to contribute more to the prediction than the controller did in 360° videos. Since participants did not need to interact with the video, this result was not surprising. Moreover, the HMD and controller seemed to contribute to prediction more than the gaze sensor did in VR fitness. Additionally, all three contributors to prediction in the rhythm game. These results resonate with the aforementioned observation that different sensors capture different aspects of users' conditions. Thus, when participants experienced VR activities that incorporated those aspects, dropping features that captured information about those aspects from consideration harmed the prediction of opportune moments in those particular VR activities.

Figure 8: ROC curves of models using different combinations of feature type. (Left) Personalized 1D-CNN model. (Right) General LSTM model.



4.3.3 Performance Improvements with Additional Information. Next, we examined the performances of the models using all the feature types (sensors, activity conditions, and summarized engagement). The results for both personalized models and general models are presented in Table 3. For the personalized models, 1D-CNN- and MLSTM-FCN-based models performed better than LSTM. Specifically, whereas 1D-CNN achieved better precision and AUROC, MLSTM-FCN achieved better accuracy, recall, and F1-score. However, in general models, LSTM-based models performed the best across all performance metrics.

Next, we compare the performance of the models using different combinations of feature types, shown in Table 4; Figure 8 shows their ROC curves. Similarly, the table and the figure only show the results from the models that performed the best among the three NN-based models (1D-CNN for personalized model and LSTM for general model, respectively). Overall, using all the features—sensor, activity, and summarized engagement—to predict opportune moments in VR achieved an accuracy of 85% and an AUROC of 0.93 for a personalized model and an accuracy of 75% and an AUROC of 0.81 for a general model, comparable to the interruptibility prediction results on desktop computers (e.g., [20, 31, 79, 81, 100]) and mobile phones (e.g. [57, 70, 76, 88, 90, 95]).

Interestingly, the results showed that models that leveraged activity condition information seemed to achieve the best AUROC, compared to sensor and engagement information; even using only the activity information feature alone achieved an AUROC of 0.91. Combining this information with either sensor or engagement information further improved performance. This indicates that the provision of activity information from a VR content provider to help such prediction may be worthwhile. It was unexpected, however, that using engagement information alone, which involved

Table 4: Comparison of feature types. Sensors combine features of HMD, controller, and gaze sensor. Activity is a label for predefined categories and subcategories of VR activities. Engagement is based on the majority of participants' engagement labels. Legend: "Sen.": Sensors, "Act.": Activity, "Eng.": Engagement, "Acc.": Accuracy, "Prec.": Precision, "F1.": F1-Score, "AUC": AUROC

	1D-CNN Personalized Model					LSTM General Model				
	Acc.	Recall	Prec.	F1.	AUC	Acc.	Recall	Prec.	F1.	AUC
Sen.	0.7616	0.6985	0.7018	0.7002	0.8572	0.6581	0.5709	0.5710	0.5710	0.7014
Act.	0.8210	0.7887	0.7683	0.7784	0.9101	0.7504	**0.6947**	0.6839	**0.6893**	0.8022
Eng.	0.7253	0.4207	0.7926	0.5497	0.7454	0.7253	0.4090	**0.8061**	0.5427	0.6667
Sen.+Act.	0.8436	**0.8038**	0.8037	0.8037	0.9257	0.7447	0.6946	0.6744	0.6844	0.8011
Sen.+Eng.	0.7983	0.7165	0.7629	0.7390	0.8831	0.7180	0.4092	0.7777	0.5362	0.7264
Act.+Eng.	0.8287	0.7795	0.7881	0.7838	0.9166	0.7505	0.6590	0.6980	0.6779	**0.8140**
Sen.+Act.+Eng.	**0.8503**	0.8032	**0.8178**	**0.8104**	**0.9304**	**0.7549**	0.6644	0.7039	0.6836	0.8101

summarized information by the participants, would obtain the lowest AUROC (0.75). We found that this was because it only achieved low recall (nearly 40%) in both the personalized and general models. In contrast, it achieved particularly high precision (nearly 80%) in both the personalized and general models.

In examining the performances of the models that combined features against those which used only one type of features, we found that combining either two types of features generally improved the performance of the personalized models. Notably, while engagement alone achieved relatively low recall, adding either sensor information or activity information to it significantly raised the recall (i.e., raising the likelihood of capturing actual opportune moments), especially in a personalized model. Finally, combining all three types of features achieved the best performance in nearly all metrics for the personalized model. However, it did not achieve any of the best performance metrics for the general model.

4.3.4 Additional Insight : Activity, Engagement and Opportuneness. To gather further insights into the prediction results, we examined the proportion of opportune and inopportune moments in different VR activities. Here, we provide some explanations from the participants' perspectives gained from interviews we conducted with them. Figure 9 shows the percentages of opportune moments against engagement—making, in total, four combinations—in different activity conditions. Blue and red represented opportune and inopportune moments, respectively. Considering the rightmost bar, the rhythm game at a difficulty level of hard, for example, nearly all moments associated with this activity conditions were associated with high-engagement and inopportune moments for notifications. In the difficulty level of easy, however, a few more moments were associated with opportune for notifications (thus a greater proportion of blue), among which nearly half were associated with high engagement and the rest were associated with low engagement. Overall, the chart explains the strong predictive power of activity conditions of opportune moments; the ratios of red to blue varied among activity conditions. For example, the length of opportune moments in 360° videos of high arousal was 2.9 times and 3.8 times shorter than that in 360° videos of medium and low arousal, respectively. The length of the opportune moments in VR fitness with high-intensity translational exercises was 3.4 times and 4.7 times shorter than with VR fitness of low intensity of translation and low intensity of rotation, respectively. It was even 5.9 times shorter than in resting in VR fitness. Most participants supported these observations in the end-of-study interviews, mentioning that whether or not notifications were timed well depended on the content of the activity they were engaged in at the time. Many users reported examples of this: *"While the video content is rapidly changing, I do not want to receive a notification"* (P6) and *"I feel that notifications are very annoying when I am playing the game"* (P10).

Interestingly, while all of the levels of gameplay in the rhythm game (i.e., easy, normal, and hard) were considered to be mostly inopportune for notifications, nearly 40% of the break times were also considered to be inopportune. This implies that break times are not necessarily considered to be timed well. It is possible that perceptions of opportuneness still depend on the pace and intensity of an activity near the break time. Perhaps in a more static game, break time would be more likely considered opportune for notifications. Another notable finding is that in all of the scenarios, nearly 20% of the time was labeled as inopportune, suggesting possible disagreement among participants on opportuneness in the various scenarios. In the interviews, we learned different opinions about the opportuneness of these moments. For example, some participants stated that even when they were having a break, they did not want to receive a notification in the rhythm game; as P14 stated, *"I want to prepare for the following game, so the notifications are not welcome, even during the break of the game."* However, some participants considered it as suitable; P15 stated, *"I am not busy in the break, so it is OK to receive a notification."* The diversity of these perceptions seems to be reflected in the prediction results, which may largely explain why activity information performed better in a personalized model than in a general model.

Another interesting and noteworthy finding is that while participants tended to associate low-engagement moments with opportune moments, Figure 9 clearly illustrated a difference between engagement and opportune moments. There were moments of low engagement with which participants associated inopportune moments (light red on the bottom) and also moments of high engagement with which participants associated opportune moments (dark blue on the top).

Overall, the former situation (i.e., associations between low engagement and inopportune moments) were less prevalent, but was relatively higher in VR fitness during rest (8%) and high intensity of translation (12%). In these moments, participants considered

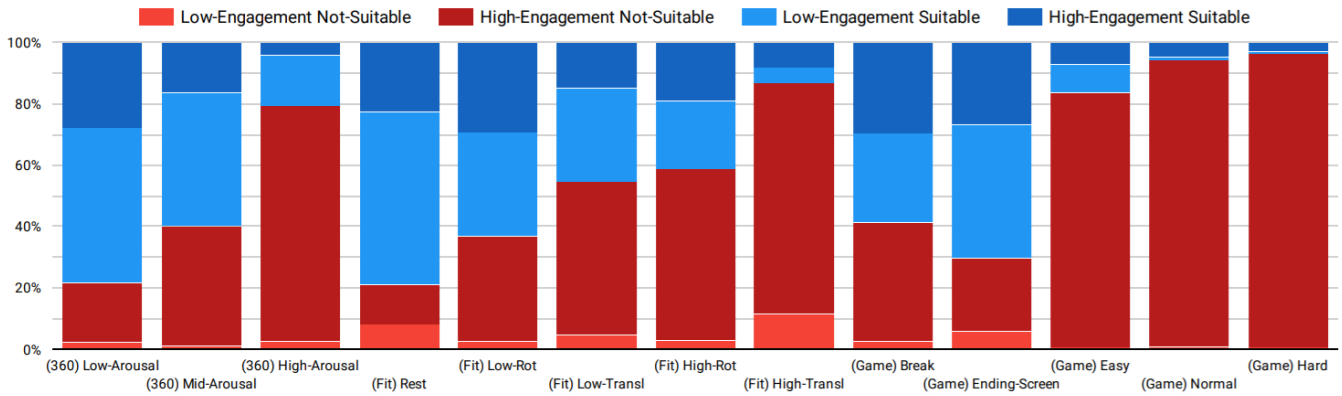


Figure 9: Activity, engagement, and opportuneness. ***Legend: " 360°": 360° Video, "Fit": VR Fitness, "Game": Rhythm Game, "Low-Rot": Low-Intensity-Rotation, "Low-Transl":Low-Intensity-Translation, "High-Rot": High-Intensity-Rotation, "High-Rot": High-Intensity-Translation.***

themselves not engaged in the activity, but they were reluctant to receive notifications during these moments because they felt tired; as P9 mentioned, *"That was when I was physically tired. I was not engaged, but I did not want to receive notifications."* Situations in which participants associated high engagement with opportune moments were more frequent than associations between low engagement and inopportune times. These moments were more prevalent when participants were watching low-arousal 360° videos (28%), were resting (22%), were performing low-intensity rotation exercises (29%) in VR fitness, were on a break (30%), or during the end screen (27%) in the rhythm game. Participants reported different reasons why they thought these moments were opportune for receiving notifications even though they were engaged: *"I was engaged in exploring in a static scene in the 360° video, but I was willing to receive notifications because I could explore the scene later"* (P9); *"I was engaged when working out, but I could read notifications as long as the move was not too hard"* (P16); and *"At the break of the rhythm game, I still feel engaged because of the visual and audio content was immersive. But I can respond to a notification because I am not busy"* (P18). The fact that participants felt it was opportune to receive notifications in many high-engagement situations explained why the model using the engagement feature alone missed these moments in its predictions, resulting in low recall (about 40%).

5 DISCUSSION

In this section, we discuss our results, including the implications of the results for the personalized and general models, sensor comparison, and model performance when including activity information and summaries about engagement.

5.1 Sensor Comparison and Importance

Our results show that using off-the-shelf VR devices can achieve an accuracy of 77.38% and an AUROC of 0.86 for the personalized model and an accuracy of 67.15% and AUROC of 0.73 in the general model, respectively. These results are comparable with the prediction performance in previous similar works that focused on the workplace and mobile context (e.g., [20, 21, 70, 89, 100]), demonstrating the feasibility of predicting opportune moments for sending notification in VR.

From the results of feature comparison, we found that no single sensor dominated the others and each sensor was important in different types of VR activities. For example, each sensor type captured different aspects of the participants' activity (e.g., head movement, hand movement, gaze movement), which were demanded to different extents in different VR activities. As a result, we observed differing performance declines when removing each sensor type from the prediction in each VR activity. For instance, information from the gaze sensors were important in activities that required various levels of visual attention; information from controller sensors was important in activities that required various degrees of body movement; and information from HMD sensors was important in both types of activities. Moreover, the sensors seemed to complement each other well, such that combining all sensors yielded the best performance metrics, including precision, which is the preferred metric for reducing the number of false positives (i.e., identifying an inopportune moment as an opportune one) and ultimately reducing interruptions in which a user would find notifications disruptive. All in all, these results show that it is feasible to use machine learning techniques to identify opportune moments for notification delivery, mainly based on the sensors from off-the-shelf VR devices. Furthermore, to build a model that can be used in various activities, it is important to include different sensors because they capture different aspects of users' movement.

5.2 Personalized vs. General Model

Several prior works have found interruptibility estimation models difficult to generalize to new users due to individual differences [21, 90, 99, 100]. Our results resonate with this observation: overall, personalized models outperformed general models. This might indicate that individual differences in movement patterns and perceptions of timing for receiving notifications made it difficult for a general model trained from a group of participants to identify a new participant's perceptions of such timing. As a result, when

only using sensor information from VR devices (including eye gaze data), the precision was only 60.12%, meaning that nearly 40% of the moments predicted as opportune for notifications were actually inopportune. However, the precision increased to 70% when building a personalized model. Nevertheless, while a personalized model sounds appealing because of its superior performance in nearly all aspects to a general model in our predicting task, building a personalized model requires more time and effort, as more individual data are needed. One possible way to resolve this dilemma is to start with a general model and then to use machine learning techniques to reduce the amount of personalized data required to achieve a high performance. That said, in our study, collecting each individual participant's data took, on average, 30 minutes per activity, and the time and effort to collect personalized data may be further reduced through techniques such as user clustering (e.g., [90]) and active learning (e.g., [18]). Developers can consider the trade-off between cost and performance based on their own needs.

5.3 Sensor-Based Only or Including Metadata for Opportune Moment Prediction?

Assuming that activity conditions and summarized engagement information could potentially be supplied from VR content providers and crowdsourcing [37], we examined whether the inclusion of these two types of information improved the performance of the prediction of opportune moments for receiving notifications. Our results consistently show that the inclusion of information about activity conditions and engagement improved the performance (about 5 to 13% of improvement in all metrics in both the personalized and the general models). Figure 9 also shows that activity conditions and engagement information both significantly influence and relate to participants' perceptions of opportune moments for receiving notifications. Considering these results together, it seemed most ideal to obtain all three types of information to train a predictive model.

However, because obtaining both of these pieces of information requires more effort and time, the key question is whether it is worth the effort to collect this information. For example, if developers want to focus on using real-time sensor information to make such predictions, a personalized model with a nearly 0.86 AUROC and 72% F1 score should suffice. However, with slightly more effort from content providers or developers to supply metadata about VR activity at the time of participation, including even one piece of information (in our case, activity condition), the prediction performance can be considerably improved (nearly 5 to 9% across all metrics in a personalized model and nearly 5 to 13% across all metrics in a general model). As for engagement information, video annotation was not as common as other crowdsourcing tasks on a somewhat smaller scale (e.g., image tagging, translation, handwriting recognition) that have been prevalent on some online crowdsourcing platforms (e.g., Google Crowdsource[5]); these other crowdsourcing tasks have appealed to a large number of crowd workers who have performed these tasks, but video annotation has increasingly been gaining in popularity in fields related to computer vision [91], multimedia [85], and HCI [12]. In our study, we used annotations from only 20 people and could already see improvement in the performance of the personalized model. Given that many participants in our study produced similar descriptions of engagement, it is likely that even several people's engagement descriptions would help improve the performance of the models. Furthermore, with the activity conditions and engagement information combined, developing a general model may become even more appealing because it saves the effort required to access sensor information and convert them into features. If a practitioner favors precision over recall in predicting opportune moments, even using engagement information alone to build a general model may be a good starting point. However, if a practitioner favors recall over precision, it should be noted that a model with the engagement information alone may likely miss the majority of actually opportune moments because users are likely to feel engaged because of the immersive experience of VR while still feeling receptive to notifications.

We deem that there is no right answer regarding whether it is worthwhile to obtain each type of information to help the prediction of opportune moments because there is a choice of who bears the burden. Using off-the-shelf sensors is convenient because they are all accessible. It also makes development work and prediction models independent from content providers and more portable across VR platforms because the prediction would mainly rely on real-time sensor information, which would presumably be available on most VR products. Even without engagement information from a crowd, many works have demonstrated that engagement level can be inferred from EEG [22] and video [71]. However, taking such a direction would mean that the burden is mainly on the developers of the VR platform (presuming that the platform takes the responsibility of delivering real-world notifications across various VR applications). If using metadata, it is likely that when VR becomes pervasive among consumers, crowdsourcing video annotation tasks will also become more common. Moreover, if content providers commonly supply metadata from their VR content, developers of VR platforms may find it easier to access and process sensor information to build an appropriate prediction model. However, the downside of this scenario is that prediction may not work in applications where no metadata are supplied.

In hoping that users will encounter fewer disruptions caused by notifications regardless of which VR application they are using, we deem that developers of the VR platform ought to take responsibility for leveraging sensors to identify opportune moments for notification delivery but meanwhile encourage (or instruct) content providers to provide more metadata from the activities to further improve their predictions. Given that VR platforms may soon become prevalent, delivering notifications from the real world to users in VR may be unavoidable, and preventing these notifications from disrupting the immersion experience is crucial. Our study has shed some light on useful information for identifying opportune moments for delivering notifications. We hope the results can be a useful reference for VR platform developers and content providers when making decisions.

[5] https://crowdsource.google.com/

6 LIMITATIONS AND FUTURE WORK

The current study was subject to a number of limitations. The first limitation relates to generalizability. Although we tried to collect diverse data from three activities that varied in terms of body movement and visual attention, these VR activities could not be representative of all kinds of VR applications. The number of participants was also small, and the participants were skewed toward a younger population. Therefore, their experience might also not be generalizable to the broader population. In terms of notification presentation and position, we only considered visual notifications fixed at the upper region of the field of view; however, different display designs might affect the perceived opportuneness of the timing (e.g., notifying the user visually might be more disruptive than notifying them haptically in activities that require high visual attention). Given these limitations, we encourage future research to explore this topic with more diverse populations, notification presentations and modalities, and VR activities.

Second, in terms of ecological validity, the participants experienced the VR activities in a lab setting. Thus, their perceptions of opportune moments for notification delivery may be affected by other factors, such as recent context, content of the message, and sender-recipient relationships in a real-life setting. Moreover, we did not tell participants how long each activity and break would take. However, participants' awareness of the length of each scenario could influence their perceived interruptibility. For example, some of the participants told us that their uncertainty about the length of the breaks in the rhythm game affected their willingness to receive notifications during these moments. Although participants could establish some rough idea about how long the duration of each break and activity session would be through the warm-up tasks, such an awareness could be different from real-life VR experiences, which could be reinforced through additional experiences with a VR particular activity. We encourage future research to examine the prediction performance, as well as how much a system that equips the prediction, can reduce interruptions from notifications in a field study.

Third, as VR is not yet widely used in daily life as desktop computers and mobile phones, individual differences in VR experience might also influence participants' categorization of both opportune moments and engagement. For example, it is likely that participants with less VR experience might find themselves more engaged due to the novelty effect and report less opportune moments. Although we did not observe such pattern in our study, future research can also explore if differences in VR experiences can affect perceptions of opportune moments for notifications.

Fourth, instead of using the experience sampling method, we used the retrospective method to collect the data in this study. Since the retrospective method might involve a memory bias problem, we kept each session short (five minutes). As a result, most users reported that it was easy and clear to label the data. However, we recognized that experience after-the-fact might still be slightly different from the experience in-the-moment. In addition, the post hoc continuous labeling method in our study simplified the dynamics and rapidness of change in people's perceptions of opportune moments, which is also likely to be on a continuum of choice rather than a binary choice. Using additional tools to provide reference points (e.g., using EEG to detect changes in brain waves) could help to collect more precise and fine-grain labeled data.

Sixth, in our work, we only considered sensors that captured users' movement in this study. However, many prior studies found the predictive power of biometric features (e.g., pupil size [21], EEG [48], EDA [99] and so on) and interactional features (e.g., click events [20] and keystroke [100]) to predict interruptibility. We also did not use computer vision techniques to extract features of the visual elements that participants saw in the VR activities. This was because we aimed to use off-the-shelf sensors directly available on the VR device, which would be more accessible and require less specific knowledge of vision to build this model. Nevertheless, we believe that the inclusion of these features would further improve predictions using only real-time information. We deem this work to be the first step in the prediction of opportune moments for notification delivery in VR. Future research can consider combining more biometric, interactional features, and even vision features to further improve predictions.

7 CONCLUSION

To reduce the potential disruption to VR users caused by real-life notifications, we aimed to build prediction models using sensors on VR devices to predict opportune moments for real-life notification delivery in VR. We additionally examined if adding metadata from activities and engagement information potentially obtainable via crowdsourcing could improve the prediction performance. Our results show that using sensors from off-the-shelf VR devices could achieve an AUROC of 0.86 for personalized models and 0.73 for a general models. In addition, we found that no single sensor in the VR devices was superior to the others, and each sensor was important in different types of VR activities. Because each sensor seemed to complement the others well, combining all sensors achieved the best performance among the other combinations. Finally, we showed that by including the activity conditions and summarized engagement information, the prediction achieved an AUROC of up to 0.93 for the personalized model and 0.81 for the general model. These results not only demonstrate the feasibility of identifying opportune moments for notifications in VR but also indicates that such predictions could achieve great performance. Nevertheless, to examine whether such prediction tasks can also achieve similar performance in the real world requires more investigation and validation in field studies.

ACKNOWLEDGMENTS

We greatly thank all the study participants. This project is supported by the Higher Education Sprout Project of the National Yang Ming Chiao Tung University and Ministry of Education (MOE), Taiwan, and the Ministry of Science and Technology, Taiwan (MOST109-2628-E-009-010-MY3, MOST110-2222-E-A49-008-MY3)

REFERENCES

[1] Piotr D Adamczyk and Brian P Bailey. 2004. If not now, when? The effects of interruption at different moments within task execution. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 271–278. https://doi.org/10.1145/985692.985727

[2] Christoph Anderson, Isabel Hübener, Ann-Kathrin Seipp, Sandra Ohly, Klaus David, and Veljko Pejovic. 2018. A Survey of Attention Management Systems in Ubiquitous Computing Environments. *Proc. ACM Interact. Mob. Wearable*

Ubiquitous Technol. 2, 2, Article 58 (July 2018), 27 pages. https://doi.org/10.1145/3214261

[3] Brian P. Bailey and Shamsi T. Iqbal. 2008. Understanding Changes in Mental Workload during Execution of Goal-Directed Tasks and Its Application for Interruption Management. *ACM Trans. Comput.-Hum. Interact.* 14, 4, Article 21 (Jan. 2008), 28 pages. https://doi.org/10.1145/1314683.1314689

[4] Brian P Bailey and Joseph A Konstan. 2006. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in human behavior* 22, 4 (2006), 685–708. https://doi.org/10.1016/j.chb.2005.12.009

[5] Brian P Bailey, Joseph A Konstan, and John V Carlis. 2001. The Effects of Interruptions on Task Performance, Annoyance, and Anxiety in the User Interface.. In *Interact*, Vol. 1. 593–601.

[6] BEASTSABER. 2019. *Beat Saber Custom Map / Song Reviews.* Retrieved July 22, 2020 from https://bsaber.com/

[7] Deborah A Boehm-Davis and Roger Remington. 2009. Reducing the disruptive effects of interruption: A cognitive framework for analysing the costs and benefits of intervention strategies. *Accident Analysis & Prevention* 41, 5 (2009), 1124–1129.

[8] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2334–2346. https://doi.org/10.1145/3025453.3026044

[9] Yung-Ju Chang, Yi-Ju Chung, and Yi-Hao Shih. 2019. I Think It's Her: Investigating Smartphone Users' Speculation about Phone Notifications and Its Influence on Attendance. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–13.

[10] Yung-Ju Chang and John C. Tang. 2015. Investigating Mobile Users' Ringer Mode Usage and Attentiveness and Responsiveness to Communication. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Copenhagen, Denmark) *(MobileHCI '15)*. Association for Computing Machinery, New York, NY, USA, 6–15. https://doi.org/10.1145/2785830.2785852

[11] David Cournapeau. 2007. *scikit-learn, Machine Learning in Python.* Retrieved July 22, 2020 from https://scikit-learn.org/

[12] Federico Cruciani, Mark P Donnelly, Chris D Nugent, Guido Parente, Cristiano Paggetti, and William Burns. 2010. DANTE: a video based annotation tool for smart environments. In *International Conference on Sensor Systems and Software*. Springer, 179–188. https://doi.org/10.1007/978-3-642-23583-2_13

[13] Anup Doshi and Mohan M. Trivedi. 2012. Head and eye gaze dynamics during visual attention shifts in complex environments. *Journal of Vision* 12, 2 (02 2012), 9–9. https://doi.org/10.1167/12.2.9

[14] Julie Epelboim, Robert M Steinman, Eileen Kowler, Zygmunt Pizlo, Casper J Erkelens, and Han Collewijn. 1997. Gaze-shift dynamics in two kinds of sequential looking tasks. *Vision research* 37, 18 (1997), 2597–2607. https://doi.org/10.1016/s0042-6989(97)00075-8

[15] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2019. Deep learning for time series classification: a review. *Data mining and knowledge discovery* 33, 4 (2019), 917–963. https://doi.org/10.1007/s10618-019-00619-1

[16] Joel E. Fischer, Chris Greenhalgh, and Steve Benford. 2011. Investigating Episodes of Mobile Phone Activity as Indicators of Opportune Moments to Deliver Notifications. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services* (Stockholm, Sweden) *(MobileHCI '11)*. Association for Computing Machinery, New York, NY, USA, 181–190. https://doi.org/10.1145/2037373.2037402

[17] Joel E. Fischer, Nick Yee, Victoria Bellotti, Nathan Good, Steve Benford, and Chris Greenhalgh. 2010. Effects of Content and Time of Delivery on Receptivity to Mobile Interruptions. In *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services* (Lisbon, Portugal) *(MobileHCI '10)*. Association for Computing Machinery, New York, NY, USA, 103–112. https://doi.org/10.1145/1851600.1851620

[18] Robert Fisher and Reid Simmons. 2011. Smartphone interruptibility using density-weighted uncertainty sampling with reinforcement learning. In *2011 10th international conference on machine learning and applications and workshops*, Vol. 1. IEEE, 436–441.

[19] James Fogarty, Scott E. Hudson, and Jennifer Lai. 2004. Examining the Robustness of Sensor-Based Statistical Models of Human Interruptibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vienna, Austria) *(CHI '04)*. Association for Computing Machinery, New York, NY, USA, 207–214. https://doi.org/10.1145/985692.985719

[20] James Fogarty, Andrew J. Ko, Htet Htet Aung, Elspeth Golden, Karen P. Tang, and Scott E. Hudson. 2005. Examining Task Engagement in Sensor-Based Statistical Models of Human Interruptibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, USA) *(CHI '05)*. Association for Computing Machinery, New York, NY, USA, 331–340. https://doi.org/10.1145/1054972.1055018

[21] Thomas Fritz, Andrew Begel, Sebastian C Müller, Serap Yigit-Elliott, and Manuela Züger. 2014. Using psycho-physiological measures to assess task difficulty in software development. In *Proceedings of the 36th international conference on software engineering*. 402–413. https://doi.org/10.1145/2568225.2568266

[22] Federico Cirett Galán and Carole R. Beal. 2012. EEG Estimates of Engagement and Cognitive Workload Predict Math Problem Solving Outcomes. In *Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization* (Montreal, Canada) *(UMAP'12)*. Springer-Verlag, Berlin, Heidelberg, 51–62. https://doi.org/10.1007/978-3-642-31454-4_5

[23] Ceenu George, Philipp Janssen, David Heuss, and Florian Alt. 2019. Should I Interrupt or Not? Understanding Interruptions in Head-Mounted Display Settings. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. 497–510. https://doi.org/10.1145/3322276.3322363

[24] Sarthak Ghosh, Lauren Winston, Nishant Panchal, Philippe Kimura-Thollander, Jeff Hotnog, Douglas Cheong, Gabriel Reyes, and Gregory D. Abowd. 2018. NotifiVR: Exploring Interruptions and Notifications in Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics* 24, 4 (April 2018), 1447–1456. https://doi.org/10.1109/TVCG.2018.2793698

[25] Nitesh Goyal and Susan R. Fussell. 2017. Intelligent Interruption Management Using Electro Dermal Activity Based Physiological Sensor for Collaborative Sensemaking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 52 (Sept. 2017), 21 pages. https://doi.org/10.1145/3130917

[26] Eija Haapalainen, SeungJun Kim, Jodi F Forlizzi, and Anind K Dey. 2010. Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. 301–310. https://doi.org/10.1145/1864349.1864395

[27] Joyce Ho and Stephen S. Intille. 2005. Using Context-Aware Computing to Reduce the Perceived Burden of Interruptions from Mobile Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, USA) *(CHI '05)*. Association for Computing Machinery, New York, NY, USA, 909–918. https://doi.org/10.1145/1054972.1055100

[28] Eric Horvitz, Paul Koch, and Johnson Apacible. 2004. BusyBody: Creating and Fielding Personalized Models of the Cost of Interruption. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work* (Chicago, Illinois, USA) *(CSCW '04)*. Association for Computing Machinery, New York, NY, USA, 507–510. https://doi.org/10.1145/1031607.1031690

[29] Ching-Yu Hsieh, Yi-Shyuan Chiang, Hung-Yu Chiu, and Yung-Ju Chang. 2020. Bridging the Virtual and Real Worlds: A Preliminary Study of Messaging Notifications in Virtual Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376228

[30] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

[31] Scott Hudson, James Fogarty, Christopher Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny Lee, and Jie Yang. 2003. Predicting Human Interruptibility with Sensors: A Wizard of Oz Feasibility Study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 257–264. https://doi.org/10.1145/642611.642657

[32] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *CoRR* abs/1502.03167 (2015). arXiv:1502.03167 http://arxiv.org/abs/1502.03167

[33] Shamsi T. Iqbal and Brian P. Bailey. 2006. Leveraging Characteristics of Task Structure to Predict the Cost of Interruption. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montréal, Québec, Canada) *(CHI '06)*. Association for Computing Machinery, New York, NY, USA, 741–750. https://doi.org/10.1145/1124772.1124882

[34] Shamsi T. Iqbal and Brian P. Bailey. 2008. Effects of Intelligent Notification Management on Users and Their Tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 93–102. https://doi.org/10.1145/1357054.1357070

[35] Katherine Isbister and Florian "Floyd" Mueller. 2015. Guidelines for the design of movement-based games and their relevance to HCI. *Human–Computer Interaction* 30, 3-4 (2015), 366–399.

[36] Jason Jerald. 2015. *The VR book: Human-centered design for virtual reality.* Morgan & Claypool.

[37] Aditya Kamath, Aradhya Biswas, and Vineeth Balasubramanian. 2016. A crowdsourced approach to student engagement recognition in e-learning environments. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–9.

[38] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Samuel Harford. 2019. Multivariate LSTM-FCNs for time series classification. *Neural Networks* 116 (2019), 237–245. https://doi.org/10.1016/j.neunet.2019.04.014

[39] Nicky Kern, Stavros Antifakos, Bernt Schiele, and Adrian Schwaninger. 2004. A Model for Human Interruptability: Experimental Evaluation and Automatic Estimation from Wearable Sensors. In *Proceedings of the Eighth International*

Symposium on Wearable Computers (ISWC '04). IEEE Computer Society, USA, 158–165. https://doi.org/10.1109/ISWC.2004.3

[40] SeungJun Kim, Jaemin Chun, and Anind K. Dey. 2015. Sensors Know When to Interrupt You in the Car: Detecting Driver Interruptibility Through Monitoring of Peripheral Interactions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 487–496. https://doi.org/10.1145/2702123.2702409

[41] Kyohei Komuro, Yuichiro Fujimoto, and Kinya Fujita. 2017. Relationship Between Worker Interruptibility and Work Transitions Detected by Smartphone. In *"Human-Computer Interaction. Interaction Contexts"*, Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 687–699. https://doi.org/10.1007/978-3-319-58077-7_53

[42] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. 2018. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering* 15, 5 (Jul 2018), 056013. https://doi.org/10.1088/1741-2552/aace8c

[43] Hao-Ping Lee, Kuan-Yin Chen, Chih-Heng Lin, Chia-Yu Chen, Yu-Lin Chung, Yung-Ju Chang, and Chien-Ru Sun. 2019. Does Who Matter? Studying the Impact of Relationship Characteristics on Receptivity to Mobile IM Messages. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300756

[44] Song-Mi Lee, Sang Min Yoon, and Heeryon Cho. 2017. Human activity recognition from accelerometer data using Convolutional Neural Network. In *2017 ieee international conference on big data and smart computing (bigcomp)*. IEEE, 131–134. https://doi.org/10.1109/BIGCOMP.2017.7881728

[45] Uichin Lee, Joonwon Lee, Minsam Ko, Changhun Lee, Yuhwan Kim, Subin Yang, Koji Yatani, Gahgene Gweon, Kyong-Mee Chung, and Junehwa Song. 2014. Hooked on Smartphones: An Exploratory Study on Smartphone Overuse among College Students. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 2327–2336. https://doi.org/10.1145/2556288.2557366

[46] Benjamin J. Li, Jeremy N. Bailenson, Adam Pines, Walter J. Greenleaf, and Leanne M. Williams. 2017. A Public Database of Immersive VR Videos with Corresponding Ratings of Arousal, Valence, and Correlations between Head Movements and Self Report Measures. *Frontiers in Psychology* 8 (2017), 2116. https://doi.org/10.3389/fpsyg.2017.02116

[47] Tzu-Chieh Lin, Yu-Shao Su, Emily Helen Yang, Yun Han Chen, Hao-Ping Lee, and Yung-Ju Chang. 2021. "Put it on the Top, I'll Read it Later": Investigating Users' Desired Display Order for Smartphone Notifications. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[48] Santosh Mathan, Stephen Whitlow, Michael Dorneich, Patricia Ververs, and Gene Davis. 2007. Neurophysiological estimation of interruptibility: Demonstrating feasibility in a field context. In *In Proceedings of the 4th International Conference of the Augmented Cognition Society*. 51–58.

[49] Mark McGill, Daniel Boland, Roderick Murray-Smith, and Stephen Brewster. 2015. A Dose of Reality: Overcoming Usability Challenges in VR Head-Mounted Displays. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2143–2152. https://doi.org/10.1145/2702123.2702382

[50] Abhinav Mehrotra and Mirco Musolesi. 2017. Intelligent notification systems: A survey of the state of the art and research challenges. *arXiv preprint arXiv:1711.10171* (2017).

[51] Abhinav Mehrotra, Mirco Musolesi, Robert Hendley, and Veljko Pejovic. 2015. Designing Content-Driven Intelligent Notification Mechanisms for Mobile Applications. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 813–824. https://doi.org/10.1145/2750858.2807544

[52] Abhinav Mehrotra, Veljko Pejovic, Jo Vermeulen, Robert Hendley, and Mirco Musolesi. 2016. My phone and me: understanding people's receptivity to mobile notifications. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 1021–1032.

[53] Abhinav Mehrotra, Veljko Pejovic, Jo Vermeulen, Robert Hendley, and Mirco Musolesi. 2016. My Phone and Me: Understanding People's Receptivity to Mobile Notifications. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 1021–1032. https://doi.org/10.1145/2858036.2858566

[54] Richard Meyes, Melanie Lu, Constantin Waubert de Puiseau, and Tobias Meisen. 2019. Ablation studies in artificial neural networks. *arXiv preprint arXiv:1901.08644* (2019).

[55] Christopher A Monk, Deborah A Boehm-Davis, and J Gregory Trafton. 2002. The attentional costs of interrupting task performance at various stages. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 46. SAGE Publications Sage CA: Los Angeles, CA, 1824–1828.

[56] Christopher A Monk, J Gregory Trafton, and Deborah A Boehm-Davis. 2008. The effect of interruption duration and demand on resuming suspended goals. *Journal of experimental psychology: Applied* 14, 4 (2008), 299.

[57] Griffin D. Romigh Nia Peters, Bhiksha Raj. 2017. Topic and Prosodic Modeling for Interruption Management in Multi-User Multitasking Communication Interactions. In *2017 AAAI Fall Symposia, Arlington, Virginia, USA, November 9-11, 2017*. AAAI Press, 45–53. https://aaai.org/ocs/index.php/FSS/FSS17/paper/view/15974

[58] Tadashi Okoshi, Julian Ramos, Hiroki Nozaki, Jin Nakazawa, Anind K Dey, and Hideyuki Tokuda. 2015. Attelia: Reducing user's cognitive load due to interruptive notifications on smart phones. In *2015 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 96–104. https://doi.org/10.1109/PERCOM.2015.7146515

[59] Tadashi Okoshi, Julian Ramos, Hiroki Nozaki, Jin Nakazawa, Anind K. Dey, and Hideyuki Tokuda. 2015. Reducing Users' Perceived Mental Effort Due to Interruptive Notifications in Multi-Device Mobile Environments. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 475–486. https://doi.org/10.1145/2750858.2807517

[60] Tadashi Okoshi, Kota Tsubouchi, Masaya Taji, Takanori Ichikawa, and Hideyuki Tokuda. 2017. Attention and engagement-awareness in the wild: A large-scale study with adaptive notifications. In *2017 ieee international conference on pervasive computing and communications (percom)*. IEEE, 100–110.

[61] Chunjong Park, Junsung Lim, Juho Kim, Sung-Ju Lee, and Dongman Lee. 2017. Don't Bother Me. I'm Socializing! A Breakpoint-Based Smartphone Notification System. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 541–554. https://doi.org/10.1145/2998181.2998189

[62] Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. 2016. *PyTorch*. Retrieved July 22, 2020 from https://pytorch.org/

[63] Veljko Pejovic and Mirco Musolesi. 2014. InterruptMe: Designing Intelligent Prompting Mechanisms for Pervasive Applications. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Seattle, Washington) *(UbiComp '14)*. Association for Computing Machinery, New York, NY, USA, 897–908. https://doi.org/10.1145/2632048.2632062

[64] Veljko Pejovic, Mirco Musolesi, and Abhinav Mehrotra. 2015. Investigating The Role of Task Engagement in Mobile Interruptibility. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct* (Copenhagen, Denmark) *(MobileHCI '15)*. Association for Computing Machinery, New York, NY, USA, 1100–1105. https://doi.org/10.1145/2786567.2794336

[65] Martin Pielot, Bruno Cardoso, Kleomenis Katevas, Joan Serrà, Aleksandar Matic, and Nuria Oliver. 2017. Beyond Interruptibility: Predicting Opportune Moments to Engage Mobile Phone Users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 91 (Sept. 2017), 25 pages. https://doi.org/10.1145/3130956

[66] Martin Pielot, Karen Church, and Rodrigo de Oliveira. 2014. An In-Situ Study of Mobile Phone Notifications. In *Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices & Services* (Toronto, ON, Canada) *(MobileHCI '14)*. Association for Computing Machinery, New York, NY, USA, 233–242. https://doi.org/10.1145/2628363.2628364

[67] Martin Pielot, Rodrigo de Oliveira, Haewoon Kwak, and Nuria Oliver. 2014. Didn't You See My Message? Predicting Attentiveness to Mobile Instant Messages. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3319–3328. https://doi.org/10.1145/2556288.2556973

[68] Martin Pielot and Luz Rello. 2015. The Do Not Disturb Challenge: A Day Without Notifications. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI EA '15)*. Association for Computing Machinery, New York, NY, USA, 1761–1766. https://doi.org/10.1145/2702613.2732704

[69] Martin Pielot and Luz Rello. 2017. Productive, anxious, lonely: 24 hours without push notifications. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–11.

[70] Benjamin Poppinga, Wilko Heuten, and Susanne Boll. 2014. Sensor-Based Identification of Opportune Moments for Triggering Notifications. *IEEE Pervasive Computing* 13, 1 (Jan. 2014), 22–29. https://doi.org/10.1109/MPRV.2014.15

[71] Ognjen Rudovic, Hae Won Park, John Busche, Björn Schuller, Cynthia Breazeal, and Rosalind W Picard. 2019. Personalized estimation of engagement from videos using active learning with deep reinforcement learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 217–226. https://doi.org/10.1109/CVPRW.2019.00031

[72] Daniel Russell and Ed Chi. 2014. *Looking Back: Retrospective Study Methods for HCI*. 373–393. https://doi.org/10.1007/978-1-4939-0378-8_15

[73] Daniel M Russell and Mike Oren. 2009. Retrospective cued recall: a method for accurately recalling previous user behaviors. In *2009 42nd Hawaii International Conference on System Sciences*. IEEE, 1–9.

[74] Rufat Rzayev, Sven Mayer, Christian Krauter, and Niels Henze. 2019. Notification in VR: The Effect of Notification Placement, Task and Environment. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (Barcelona, Spain) *(CHI PLAY '19)*. Association for Computing Machinery, New York, NY, USA, 199–211. https://doi.org/10.1145/3311350.3347190

[75] Alireza Sahami Shirazi, Niels Henze, Tilman Dingler, Martin Pielot, Dominik Weber, and Albrecht Schmidt. 2014. Large-scale assessment of mobile notifications. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 3055–3064.

[76] Hillol Sarker, Moushumi Sharmin, Amin Ahsan Ali, Md. Mahbubur Rahman, Rummana Bari, Syed Monowar Hossain, and Santosh Kumar. 2014. Assessing the Availability of Users to Engage in Just-in-Time Intervention in the Natural Environment. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Seattle, Washington) *(UbiComp '14)*. Association for Computing Machinery, New York, NY, USA, 909–920. https://doi.org/10.1145/2632048.2636082

[77] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15, 56 (2014), 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html

[78] Mikhail Startsev, Ioannis Agtzidis, and Michael Dorr. 2019. 1D CNN with BLSTM for automated classification of fixations, saccades, and smooth pursuits. *Behavior Research Methods* 51, 2 (2019), 556–572. https://doi.org/10.3758/s13428-018-1144-2

[79] Takahiro Tanaka, Ryosuke Abe, Kazuaki Aoki, and Kinya Fujita. 2015. Interruptibility estimation based on head motion and pc operation. *International Journal of Human-Computer Interaction* 31, 3 (2015), 167–179. https://doi.org/10.1080/10447318.2014.986635

[80] Takahiro Tanaka and Kinya Fujita. 2011. Study of User Interruptibility Estimation Based on Focused Application Switching. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work* (Hangzhou, China) *(CSCW '11)*. Association for Computing Machinery, New York, NY, USA, 721–724. https://doi.org/10.1145/1958824.1958954

[81] Takahiro Tanaka and Kinya Fujita. 2011. Study of User Interruptibility Estimation Based on Focused Application Switching. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work* (Hangzhou, China) *(CSCW '11)*. Association for Computing Machinery, New York, NY, USA, 721–724. https://doi.org/10.1145/1958824.1958954

[82] Alexey Tarasov, Sarah Jane Delany, and Charlie Cullen. 2010. Using crowdsourcing for labelling emotional speech assets. In *W3C workshop on Emotion ML*.

[83] Dan Tasse, Anupriya Ankolekar, and Joshua Hailpern. 2016. Getting Users' Attention in Web Apps in Likable, Minimally Annoying Ways. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 3324–3334. https://doi.org/10.1145/2858036.2858174

[84] G. H. (Henri) ter Hofte. 2007. Xensible Interruptions from Your Mobile Phone. In *Proceedings of the 9th International Conference on Human Computer Interaction with Mobile Devices and Services* (Singapore) *(MobileHCI '07)*. Association for Computing Machinery, New York, NY, USA, 178–181. https://doi.org/10.1145/1377999.1378003

[85] Vincent S Tseng, Ja-Hwung Su, Jhih-Hong Huang, and Chih-Jen Chen. 2008. Integrated mining of visual features, speech features, and frequent patterns for semantic video annotation. *IEEE Transactions on Multimedia* 10, 2 (2008), 260–267.

[86] Liam D Turner, Stuart M Allen, and Roger M Whitaker. 2015. Interruptibility prediction for ubiquitous systems: conventions and new directions from a growing field. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 801–812.

[87] Liam D. Turner, Stuart M. Allen, and Roger M. Whitaker. 2015. Interruptibility Prediction for Ubiquitous Systems: Conventions and New Directions from a Growing Field. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 801–812. https://doi.org/10.1145/2750858.2807514

[88] Liam D. Turner, Stuart M. Allen, and Roger M. Whitaker. 2015. Push or Delay? Decomposing Smartphone Notification Response Behaviour. In *Proceedings of the 6th International Workshop on Human Behavior Understanding - Volume 9277*. Springer-Verlag, Berlin, Heidelberg, 69–83. https://doi.org/10.1007/978-3-319-24195-1_6

[89] Gašper Urh and Veljko Pejović. 2016. TaskyApp: Inferring Task Engagement via Smartphone Sensing. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct* (Heidelberg, Germany) *(UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 1548–1553. https://doi.org/10.1145/2968219.2968547

[90] Aku Visuri, Niels van Berkel, Chu Luo, Jorge Goncalves, Denzil Ferreira, and Vassilis Kostakos. 2017. Predicting Interruptibility for Manual Data Collection: A Cluster-Based User Model. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vienna, Austria) *(MobileHCI '17)*. Association for Computing Machinery, New York, NY, USA, Article 12, 14 pages. https://doi.org/10.1145/3098279.3098532

[91] Carl Vondrick, Donald Patterson, and Deva Ramanan. 2013. Efficiently scaling up crowdsourced video annotation. *International journal of computer vision* 101, 1 (2013), 184–204.

[92] Dominik Weber, Alexandra Voit, Huy Viet Le, and Niels Henze. 2016. Notification dashboard: enabling reflection on mobile notifications. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. 936–941.

[93] Zhaohan Xiong, Martin K Stiles, and Jichao Zhao. 2017. Robust ECG signal classification for detection of atrial fibrillation using a novel neural network. In *2017 Computing in Cardiology (CinC)*. IEEE, 1–4. https://doi.org/10.22489/CinC.2017.066-138

[94] Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong Sun, Zhiru Shi, Jingyi Yu, and Shenghua Gao. 2018. Gaze prediction in dynamic 360 immersive videos. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5333–5342.

[95] Fengpeng Yuan, Xianyi Gao, and Janne Lindqvist. 2017. How Busy Are You? Predicting the Interruptibility Intensity of Mobile Users. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5346–5360. https://doi.org/10.1145/3025453.3025946

[96] André Zenner, Marco Speicher, Sören Klingner, Donald Degraen, Florian Daiber, and Antonio Krüger. 2018. Immersive Notification Framework: Adaptive & Plausible Notifications in Virtual Reality. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3170427.3188505

[97] Jing Zhang, Xindong Wu, and Victor S Sheng. 2016. Learning from crowdsourced labeled data: a survey. *Artificial Intelligence Review* 46, 4 (2016), 543–576.

[98] Manuela Züger, Christopher Corley, André N Meyer, Boyang Li, Thomas Fritz, David Shepherd, Vinay Augustine, Patrick Francis, Nicholas Kraft, and Will Snipes. 2017. Reducing interruptions at work: A large-scale field study of flowlight. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 61–72. https://doi.org/10.1145/3025453.3025662

[99] Manuela Züger and Thomas Fritz. 2015. Interruptibility of software developers and its prediction using psycho-physiological sensors. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2981–2990. https://doi.org/10.1145/2702123.2702593

[100] Manuela Züger, Sebastian C. Müller, André N. Meyer, and Thomas Fritz. 2018. Sensing Interruptibility in the Office: A Field Study on the Use of Biometric and Computer Interaction Sensors. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174165

A DETAIL OF NEURAL NETWORK LAYERS

We follow the network architecture of EEGNet [42] as our 1D-CNN because of its compact architecture, and only modify layers' configuration to better fit the dataset. For the LSTM model, we use 1 layer LSTM with 64 hidden cells. Similar to 1D-CNN, we follow the architecture of MLSTM-FCN [38] and only modify layers' configuration. In Table 5, we specify the configuration of each layer in our 1D-CNN and MLSTM-FCN, including kernel size k, stride s of convolutional and pooling layer, probability p of dropout [77] layer, reduction r of Squeeze-and-Excitation block (SE-Block) [30], hidden size h of LSTM, and output dimension d_{out} of fully-connected layer.

Table 5: Layer details of our 1D-CNN and MLSTM-FCN model. C is the number of channels in time series data. *Legend: "Conv1D":, 1D Convolution Layer, "BN":Batch-Normalization [32] "DepConv": Depthwise Convolution Layer, "SepConv": Separable Convolution Layer, "AvgPool" : Average Pooling Layer., "DimShuffle": Dimension Shuffle*

1D-CNN Layer		Input Shape
Conv	$k = (1, 51)$	$1 \times C \times 300$
BN	-	$16 \times C \times 300$
DepConv	$k = (C, 1)$	$16 \times C \times 300$
BN+ReLU	-	$32 \times 1 \times 300$
AvgPool	$k = (1, 4)$, $s = (1, 4)$	$32 \times 1 \times 300$
Dropout	$p = 0.25$	$32 \times 1 \times 75$
SepConv	$k = (1, 15)$	$32 \times 1 \times 75$
BN+ReLU	-	$32 \times 1 \times 75$
AvgPool	$k = (1, 8)$, $s = (1, 8)$	$32 \times 1 \times 75$
Dropout	$p = 0.25$	$32 \times 1 \times 9$
Dense	d_{out}=64	$32 \times 1 \times 9$

MLSTM-FCN Layer		Input Shape
Conv1D	$k = 31$	$C \times 300$
BN+ReLU	-	32×300
SE-Block	$r = 16$	32×300
Conv1D	$k = 17$	32×300
BN+ReLU	-	64×300
SE-Block	$r = 16$	64×300
Conv1D	$k = 9$	64×300
BN+ReLU	-	32×300
GlobalAvgPool	-	32×300
DimShuffle		$C \times 300$
LSTM	$h = 8$	$300 \times C$
Dropout	$p = 0.25$	8
Concate		32, 8
Dense	d_{out}=64	40

B DETAIL METRICS OF EACH MODEL

The detail of recall, precision, and AUC of each personalized model and general model based on 1D-CNN, LSTM, and MLSTM-FCN, are presented in Table 6 and Table 7. Note that these models are only different while using sensor features, therefore the result without using sensor feature are the same across three models.

Table 6: Detail of sensors and features comparison of each personalized NN-based model. ***Legend: "Prec.": Precision., "AUC": AUROC, "Con.": Controller, "Sen.": Sensors (HMD+Con.+Gaze), "Act.": Activity Category, "Eng.": Summarized Engagement***

	1D-CNN			LSTM			MLSTM-FCN		
	Recall	Prec.	AUC	Recall	Prec.	AUC	Recall	Prec.	AUC
HMD	0.6363	0.6838	0.8178	0.5508	0.6405	0.7554	0.6559	0.6938	0.8267
Con.	0.7314	0.6484	0.8300	0.6839	0.6353	0.7916	0.6955	0.6657	0.8351
Gaze	0.5834	0.6505	0.7796	0.4651	0.5793	0.6813	0.5962	0.6587	0.7917
HMD+Con.	0.7194	0.6750	0.8483	0.6708	0.6712	0.8153	0.7037	0.7084	0.8541
HMD+Gaze	0.6482	0.6968	0.8322	0.6139	0.6468	0.7730	0.6731	0.7036	0.8357
Con.+Gaze	0.6946	0.6573	0.8322	0.7027	0.6397	0.7943	0.7161	0.6795	0.8403
Sen.	0.6985	0.7018	0.8572	0.6918	0.6748	0.8199	0.7202	0.7144	0.8559
Act.	0.7887	0.7683	0.9101	0.7887	0.7683	0.9101	0.7887	0.7683	0.9101
Eng.	0.4207	0.7926	0.7454	0.4207	0.7926	0.7454	0.4207	0.7926	0.7454
Sen.+Act.	**0.8038**	0.8037	0.9257	0.8014	0.7848	0.9186	0.8016	0.8033	0.9224
Sen.+Eng.	0.7165	0.7629	0.8831	0.7040	0.7358	0.8583	0.7414	0.7580	0.8815
Act.+Eng.	0.7795	0.7881	0.9166	0.7795	0.7881	0.9166	0.7795	0.7881	0.9166
Sen.+Act.+Eng.	0.8032	**0.8178**	**0.9304**	**0.8049**	**0.8072**	**0.9240**	**0.8127**	**0.8174**	**0.9273**

Table 7: Detail of sensors and features comparison of each general NN-based model. ***Legend: "Prec.": Precision., "AUC": AUROC, "Con.": Controller, "Sen.": Sensors (HMD+Con.+Gaze), "Act.": Activity Category, "Eng.": Summarized Engagement***

	1D-CNN			LSTM			MLSTM-FCN		
	Recall	Prec.	AUC	Recall	Prec.	AUC	Recall	Prec.	AUC
HMD	0.4222	0.5609	0.6787	0.3343	0.4791	0.6212	0.3976	0.5429	0.6485
Con.	0.5406	0.5336	0.6790	**0.7110**	0.5479	0.6864	0.5943	0.5365	0.6776
Gaze	0.5228	0.5622	0.6964	0.2038	0.5195	0.5911	0.3877	0.5802	0.6976
HMD+Con.	0.5188	0.5813	0.7185	0.5246	0.5451	0.6889	0.5311	0.5891	0.7173
HMD+Gaze	0.4478	0.5981	0.7060	0.2136	0.5135	0.6029	0.4340	0.5582	0.6772
Con.+Gaze	0.5647	0.5528	0.6995	0.6065	0.5338	0.6805	0.6063	0.5470	0.6924
Sen.	0.5308	0.6012	0.7257	0.5709	0.5710	0.7014	0.5880	0.5878	0.7253
Act.	**0.6947**	0.6839	0.8022	0.6947	0.6839	0.8022	**0.6947**	0.6839	0.8022
Eng.	0.4090	**0.8061**	0.6667	0.4090	**0.8061**	0.6667	0.4090	**0.8061**	0.6667
Sen.+Act.	0.6757	0.6682	0.8040	0.6946	0.6744	0.8011	0.6322	0.6463	0.7635
Sen.+Eng.	0.4843	0.7138	0.7656	0.4092	0.7777	0.7264	0.5885	0.6226	0.7477
Act.+Eng.	0.6590	0.6980	**0.8140**	0.6590	0.6980	**0.8140**	0.6590	0.6980	**0.8140**
Sen.+Act.+Eng.	0.6460	0.6913	0.8004	0.6644	0.7039	0.8101	0.6246	0.6583	0.7724