# Exploring Model Design Spaces for Cognitive and Affective State Recognition in Immersive Virtual Reality Using Multimodal Neuro-physiological Signals

Anonymous Author(s)*


**Figure 1: Screenshots of immersive Virtual Reality (iVR) environment used in our experiments: (a) blank screen presented during the VR baseline phase; (b) Speech Trainer application used during the Speech Delivery Phase; (c) and (d) Virtual Buffet used during Food Selection Phase.**

## ABSTRACT

Immersive virtual reality (iVR) environments and wearable devices enable studying human behavior and cognitive and affective processes at a fine-grained level. Those advancements support the development of behavior modification or training programs. It is well known that developing accurate inverse inference models that map neuro-physiological signals to cognitive and affective states in iVR is crucial for the success of adaptive and personalized interventions. However, researchers building those inference models often face complex decisions like modality, feature set, and personalization. To add to the complexity, these choices may depend on the specific inference task of interest. In this paper, we explored this complex model design space through a secondary analysis of a rich neuro-physiological multimodal data set collected from 10 participants during a multi-phase iVR experiment. The experiment comprised two rest phases (one without and another with iVR), a phase on a go/no-go cognitive task to measure effortful control, a phase in which we induced stress and cognitive load by asking participants to prepare and deliver a surprise speech about their strengths and weaknesses before a virtual audience, and a phase where participants selected food in a virtual buffet environment. With these clearly defined phases serving as natural labels, we conducted a series of machine learning-based evaluations to investigate the impact of design decisions on the model's capacity to discriminate among the six phases. We further shed light on the between-person variation through non-linear manifold embedding methods on the multimodal feature set, which underscore the importance of building models adapting to individual characteristics.

## CCS CONCEPTS

• **Computer systems organization → Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

## KEYWORDS

virtual reality, neuro-physiological signals, multimodal

**ACM Reference Format:**
Anonymous Author(s). 2018. Exploring Model Design Spaces for Cognitive and Affective State Recognition in Immersive Virtual Reality Using Multimodal Neuro-physiological Signals . In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

In recent years, there has been growing interest in the adoption of immersive virtual reality (iVR) environments to study human behavior and underlying cognitive processes, supported by various validation studies demonstrating that human behavior in these environments closely mirrors real-life scenarios [4, 9] The increasing accessibility of wearable devices, such as fNIRS, GSR, and HR monitors, has made iVR an attractive tool for examining human behavior and cognitive and affective processes at a granular level, with the ultimate goal of providing adaptive support for behavior modification or health education, for example. To achieve this goal,


*Conference acronym 'XX, June 03–05, 2018, Woodstock, NY*

ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/XXXXXXX.XXXXXXX

it is crucial to develop an accurate “inverse inference" model that maps the neuro-physiological signals to cognitive and affective states that unfold in the iVR environment. The precision level of these models will significantly influence the success of adaptive and personalized interventions.

When constructing these models, however, researchers face a complex decision space, which includes factors such as modality selection (choosing sensors/devices), feature set (deciding which features to include), and personalization decisions (e.g., whether or not to incorporate personalized features). It is hypothesized that these choices may depend on the specific inference task targeted of interest, such as recognizing specific cognitive states, affective states, or a combination of both. In this study, we conducted a secondary analysis of a rich mulitmodal neuro-physiological dataset from 10 participants during a multi-phase iVR experiment. The experiment comprised two rest phases (one without and another with iVR), a phase on a go/no-go cognitive task, a phase in which participants prepared and delivered a surprise speech about their weaknesses before a virtual audience, and a phase where participants selected food in a virtual buffet environment. These phases served as natural labels for conducting machine learning experiments to evaluate the data and the model's ability to distinguish between the phases. We investigated the impact of model design decisions in this context. We also conducted an in-depth analysis of between-person variation in responses to the given stimuli via non-linear manifold embedding methods, which shed light on the necessity of personalization.

The following section will present a brief overview of the related work in this field. We will then explain the virtual environment and study protocol, including the sensing devices utilized, descriptions of the phases, and the data collection process. We will then outline the data analysis and modeling pipeline, encompassing preprocessing, feature extraction, and construction steps. Next, we will present the machine learning framework of the inverse inference problem and describe the experimental setup. In the results section, we will report on the effects of design decisions on the model's ability to discriminate between various phases, followed by an in-depth exploration of the between-person variation. The paper concludes by discussing the implications of wearable-based adaptive interventions in iVR and possible future avenues for exploration.

# 2 RELATED WORK

## 2.1 Immersive Virtual Environment as a Promising Tool for Health Education and Behavioral Intervention

In recent years, virtual reality (VR) technology has been increasingly utilized to study food-related decisions. A study demonstrated that VR effectively elicits emotional responses similar to real-life situations in the context of food-related decisions [14]. Furthermore, VR has been successfully employed to treat body image disturbances in patients with eating disorders and shows promise as a technology for cue exposure therapy in this domain [13]. Immersive virtual reality (iVR) has also been investigated for interactive food portion-size education [3], treating eating disorders [27], and training inhibitory control to reduce binge eating [24]. The application of VR in marketing research and health education demonstrates its potential to influence food selection and purchasing decisions, encourage healthier choices, and address diet-related diseases [6, 28, 34]. Additional studies have explored using VR to enhance emotion regulation [7] and treat mental disorders [33]. The ultimate success of those applications in real life necessitates a robust framework that can accurately model the complex cognitive and affective processes as the basis to provide adaptive and personalized interventions.

## 2.2 Modelling Cognitive and Affective Processes Using Neuro-Physiological Sensing Data

iVR technology is rapidly being utilized to investigate cognitive and affective processes in realistic surroundings [2]. To achieve a better understanding of these dynamic responses in iVR contexts, researchers have employed physiological signals such as EEG, fNIRS, and GSR [11, 15, 21, 32], for example, to characterize cognitive load [29], attentional states [17], stress levels [18], and brain activity patterns [2]. In the context of fNIRS, research has explored methods for predicting attentional states [16], quantifying the relationship between exercise and cognition [19], and classifying mental tasks. [5]. Likewise, wearable EEG devices has be used in virtual reality environments for real-time monitoring of cognitive functioning, particularly decision-making [25]. There are also studies that explored the integration of GSR and ECG signals for emotion recognition in virtual reality environments using wearable sensors and deep learning techniques[8].

## 2.3 Multimodal and Personalization Considerations

*Multimodal Fusion.* Multimodal fusion can handle complex or ambiguous situations where a single modality may not be sufficient[1, 10, 20, 26]. Combining multiple physiological markers provides more accurate insights into cognitive and emotional responses in immersive VR (iVR) settings [22]. For instance, Sun et al. (2020) employed functional near-infrared spectroscopy (fNIRS) and electrocardiogram (ECG) to classify affective states and revealed that the multimodal model outperformed individual modalities [31].

*Personalization.* The use of personalized data has increased in recent years to improve the accuracy of predictions and decision-making. Personalization has been found to enhance multimodal classification accuracy [12, 30]. To create a personalized baseline, data from an individual or a specific circumstance is gathered and examined to ascertain what is "normal" or usual for that individual or environment. Personalization eliminates the need to use a population-based average for comparing changes and allows comparisons according to each person's own baseline. For example, Harrivel et al. (2016) created a personalized model that predicts attentional states in a VR environment using fNIRS signals and discovered that it was more accurate than a non-personalized model [16].

These studies show the potential benefits of combining numerous physiological markers with machine learning classifiers to provide more precise insights into cognitive and affective states in

VR environments. Personalized models could also improve classification accuracy and tailor VR experiences to individuals based on physiological reactions.

# 3 OVERVIEW OF IMMERSIVE VIRTUAL REALITY ENVIRONMENTS

Figure 1 illustrates the interactive VR environment used in our experiments. Virtual Buffet (subplots (c) and (d)) is a custom-developed environment with accurate architectural references created using photogrammetry. While participants interacted with the VR Buffet, they can pick food from the buffet trays and place it on the plate (as in subplot (d)). Previous validation study (citation removed for anonymity) shows that participants' food choices in a VR environment are comparable to those in real life. This application was used during the Food Selection Phase. In addition, we used a third-party application Speech Trainer[1] which is used during the Speech Delivery phase (Subplot (b)). Subplot (a) is a screenshot of a blank screen presented to participants during the VR baseline phase.

# 4 STUDY PROTOCOL AND DATA COLLECTION

## 4.1 Participants

As part of the primary study, we enrolled 10 healthy participants from the university, including 3 males and 7 females, with an average age of 20.5 years (SD = 1.58 years). The sample was diverse in ethnicity, with 5 participants identified as White, 4 as Asian, and 1 as Black. The study was approved by the Institute Research Board from the university (name blinded for review), and informed consent was obtained from all participants.

## 4.2 Experimental Phases

The research team designed six experimental phases to understand participants' neuro-physiological responses to a variety of stimuli that may trigger interesting cognitive and affective processes in the iVR environment as related to food-related decision-making. These tasks include a baseline task to establish a neuro-physiological baseline, a Go/no-Go Task to assess participants' effortful control in a non-VR setting, and an Emotion Induction Task in which participants are surprised and asked to do an impromptu speech to a virtual audience about their strengths and weaknesses to elicit stress in participants. Finally, a VR food selection task is administered. Each task's start and end timestamps are recorded, enabling the building of supervised machine-learning models using those natural labels of experiment phases (as described below) derived from those marked timestamps.

***Phase 1: Non-VR Baseline Phase.*** After participants were equipped with ECG, GSR, and fNIRS sensors, they began the initial phase, during which they sat and rested for 5 minutes without engaging in any physical activities to acquire a neuro-physiological baseline.

***Phase 2: Go/No-Go Task Phase.*** The Go/No-Go task, a widely used psychological test, measures response effortful controls. Participants will undergo one practice and three experimental blocks with low energy-density food, high energy-density food, and neutral cues as NO-GO signals and household objects or plants as GO cues. Each block consists of 100 trials, starting with a fixation cross and followed by a GO or NO-GO cue. Participants must press the spacebar for GO cues and withhold for NO-GO cues, focusing on speed and accuracy. A 25-trial practice block provides feedback, while the experimental trials do not. A 30-second rest period separates each block, and block order is randomized. Stimuli presentation follows a pseudo-randomized pattern.

***Phase 3: VR Baseline.*** After being fitted with the VR headset and learning how to navigate the virtual environment using the controller, participants entered the VR baseline phase, during which they viewed an empty VR environment (Figure 1(a)) and rested for 2 minutes without engaging in any activities to acquire a baseline for the neuro-biophysical data in the VR environment.

***Phase 4 & 5: Emotion Induction with Impromptu Speech.*** In this phase, participants were prompted to give an unexpected task involving a 2-minute impromptu speech about their strengths and weaknesses in a VR room with virtual avatar audiences (Figure 1 (b)) They had 1 minute to prepare (Speech Preparation Phase) before delivering the speech (Speech Delivery Phase). The main objective of this phase was to induce stress and examine its potential impact on food choices in the subsequent phase. While this setup was motivated by the primary research question, the current secondary data analysis considers these phases as having distinct cognitive and affective features, which will be analyzed alongside other phases.

***Phase 6: Food Selection.*** During this phase, participants selected food items from the virtual buffet (Figure 1 (c, d)) by picking up items and placing them on trays. Participants could take as long as they wanted until they explicitly informed the experimenter that they had finished food selection. This phase lasted approximately 1 to 5 minutes, varying across participants.

## 4.3 Neuro-physiological Data Acquisition

We collected GSR, ECG, and fNIRS signals from each participant during all six phases described above. We used Shimmer3 ECG and Shimmer3 GSR sensors [2] to collect Electrocardiogram (ECG) and Galvanic Skin Response (GSR) signals. Both the ECG sensor and the GSR sensor are recorded at 512 Hz. We used an OctaMon fNIRS device from Arinis [3] to capture continuous fNIRS data waves using two wavelengths (760 & 840 nm) at a 50 Hz sampling rate. The fNIRS setup included 8 channels, each comprising a pair of detectors and sources. The distance between the receiver and transmitter, known as the source-detector or inter-optode distance, measures 35mm. Figure 2 illustrates the sensors used in this study and Figure 3 is a timeseries plot of one of the fNIRS measuree HbDiff overlaid the the 6 phases described above.

# 5 METHODS

Figure 4 outlines the steps involved in data preprocessing and feature engineering. The output of those features will be used in the machine learning experiment described below.

[1] $https://store.steampowered.com/app/552770/Speech_Trainer/$

[2] https://shimmersensing.com/product/shimmer3-gsr-unit/

[3] https://www.artinis.com/

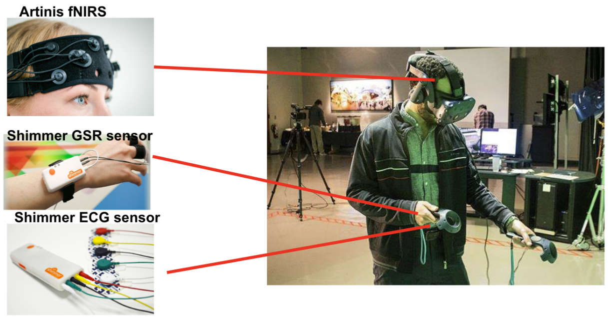


**Figure 2: The sensors used in this study.**

### 5.1 Data Preprocessing

At the start of the pipeline, we manually inspect the data streams to ensure the signals are clean and have no obvious anomalies. For fNIRS signals, we first process the raw data using the Oxysoft software from Artinis, the device provider. Four measures from each of the eight channels were derived, including O2Hb (oxygenated hemoglobin), HHb (deoxygenated hemoglobin), HbDiff (hemoglobin difference, the difference between O2Hb and HHb), and tHb (total hemoglobin, the sum of O2Hb and HHb). For ECG data, we use the out-of-box Heart Rate measures available from the Shimmer 3 Unit. For GSR signals, we use the open-source Python toolbox NeuroKit2 [23] to extract the phasic and tonic components of the GSR signals.

### 5.2 Feature Extraction and Engineering

After the preprocessing steps, we then apply a similar feature extraction pipeline for all data streams, and statistical features were extracted from a sequence of 20s-moving windows with 50% overlap. The details of features extracted from each data dream are outlined in the grid in Figure 5. All those features are derived from the values within the 20s window. We experimented with two versions of the features, a simple feature set that includes only the mean values commonly used by practitioners and is easy to understand; in addition, we calculated a less commonly used feature set that includes other statistics. As shown in the empirical experimental result (section 6.2), we identify a parsimonious set of features that captures the majority of the information. This feature set includes 32 mean features from fNIRS and 19 features from HR and GSR, which were highlighted in blue in the grid (Figure 5).

### 5.3 Personalized Feature Set

For comparison, we created a personalized feature set. To create this feature set, we first calculate the person-specific mean values of the given features derived from the first phase Non-VR Baseline as described in section 4.2. We then subtract these mean values from the original features. This step is done for each person and each feature.

### 5.4 Multimodal Feature Set

Due to the complex nature of cognitive and affective states and their interaction, a single modality may not provide robust or optimal measurements. Sensors are often affected by various noise sources (e.g., instrumental and environmental), which can increase their sensitivity to error. In fNIRS measurements, several confounds, such as skin and skull hemodynamics or motion artifacts, can affect results. This uncertainty restricts analysis. Studies have shown that physiological responses from different measurements can vary significantly between subjects [24]. As a result, a single sensor may not scale well for estimating physiological signals. Utilizing a multimodal or fusion approach may provide an effective solution to enhance the robustness and reliability of measurements and address current challenges. By combining information from multiple sources, a more comprehensive picture of the underlying processes can be obtained, leading to more accurate and reliable results.

Following an early fusion paradigm, in addition to the unimodal feature set derived from each modality, we created a multimodal feature set which is a concatenation of features derived from all modalities fusing fNIRS, GSR and ECG.

### 5.5 Machine Learning Experiments

Our primary focus in this study is determining whether the model can differentiate between the six experimental phases outlined in section 4.2. Therefore, we conducted 15 independent sets of machine learning experiments; each formulated as a binary classification problem to distinguish between phases (e.g. Phase 5: Speech Delivery and Phase 6: Food Selection). We use the Area Under Curve (AUC) metric, which ranges between 0 and 1 and is widely used to measure classifier performance. A higher AUC score indicates greater separability or difference between the two paired phases given the modeled neuro-physiological signals.

We have experimented with two popular machine learning models: Random Forest (RF), which can model non-linear relationships, and non-regularized Logistic Regression (LR), a linear model. We found that Random Forest outperformed Logistic Regression consistently in all classification tasks, achieving an average 15% improvement in AUC measures (RF 78% vs. LR 68%). To streamline our discussion of other design space dimensions, we only report results from the Random Forest experiment.

A crucial aspect of affective and cognitive state detection modeling is its capacity to generalize, meaning the model must perform well when applied to new individuals in the future. To objectively evaluate model performance in this context, we utilized a leave-one-person-out (LOPO) experiment paradigm. We trained the model on data from all participants except one and assessed the performance of the excluded participant. We report the average performance as the mean performance across all excluded participants in the test set.

## 6 RESULTS

### 6.1 Overview

In this section, we present the results demonstrating the impact of three key model design decisions on the performance of classification tasks, as previously described. Specifically, we explore the

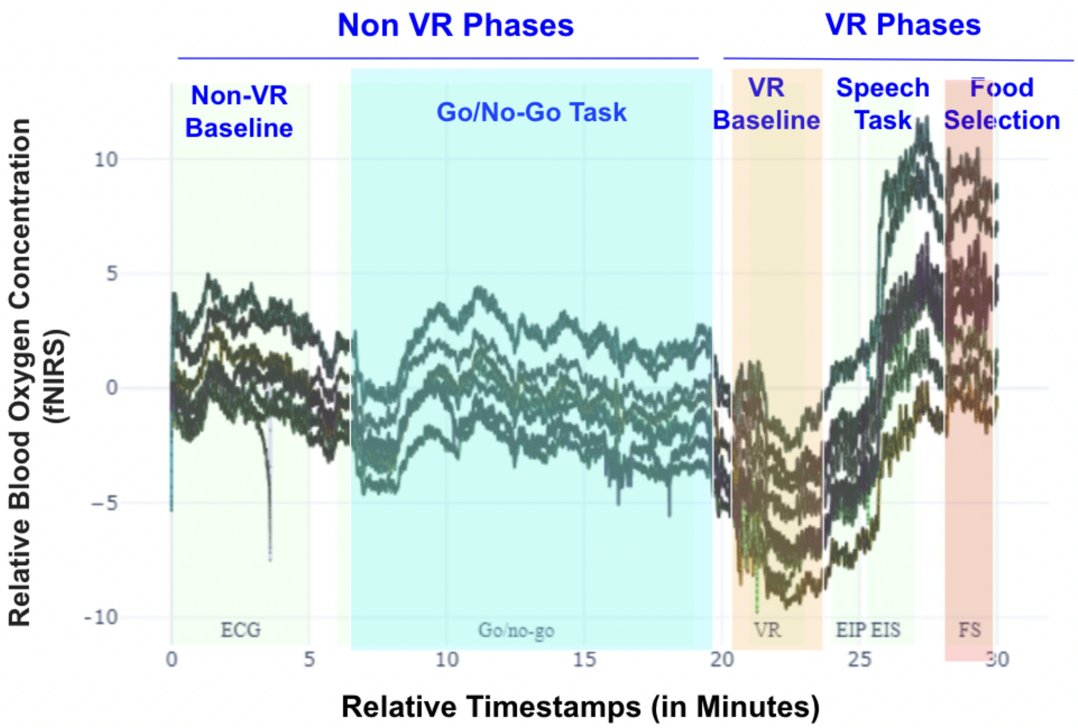


**Figure 3: An illustration of 6 phases during the experiments, overlaid with a plot of 8-channel fNIRS measures of relative blood oxygenation concentration level (i.e. HbDiff) for a given participant. Those time series is plotted against the relative timestamps, which mark the starting of the first phase (Non-VR baseline) as timestamp 0.**

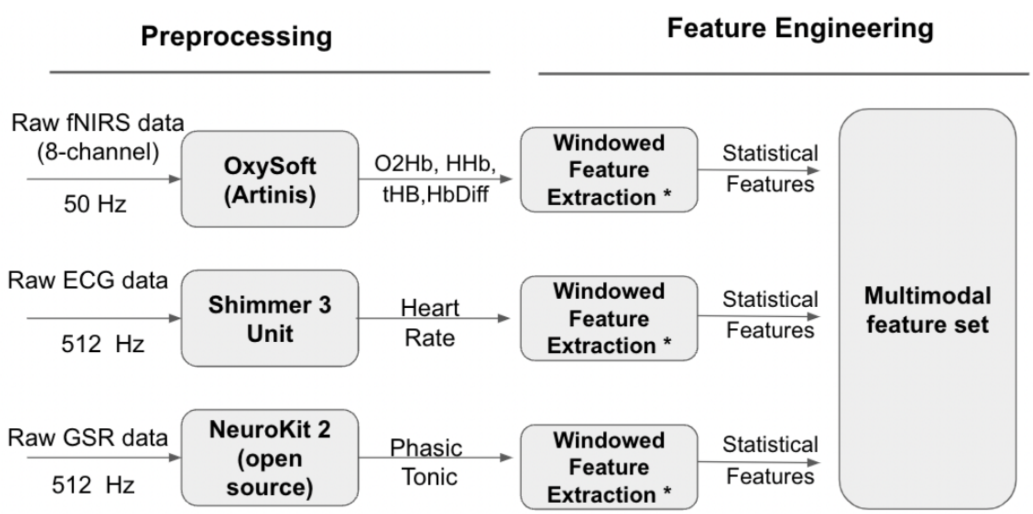


**Figure 4: Pipeline for data preprocessing and feature engineering. For the Windowed Feature Extraction step, a similar feature extraction approach was applied for all data streams to derive statistical features which were extracted from a sequence of 20s-moving windows with 50% overlap.**

following three design choices: (1) Modality, in which we investigated five different options - fNIRS alone, GSR alone, HR alone, GSR and HR combined, and a multimodal approach that includes all three modalities; (2) Feature Complexity, where we examined

| Signals | Mean Features | Count | Additional Features | Count | Total Count |
|---|---|---|---|---|---|
| fNIRS | *Simple means for each of four measures(O2Hb, HHb, tHB, and HbDiff) from eight channels* | *32* | standard_deviation<br>variances<br>median<br>variation_coefficient<br>skewness<br>kurtosis<br>root_mean_square<br>maximum<br>absolute_maximum<br>minimum | 320 (10*4*8) | 352 |
| ECG/HR | Mean HR | 1 | *Standard deviation, min and max* | *3* | 4 |
| GSR | Mean Phasic<br>Mean Tonic | 2 | *Standard deviation, number of peaks, min, max, slope, AUC (area under the curve), mean Peak (only for phasic *)* | *13 (6*2+1)* | 15 |
| Feature Count | | 35 | | 336 | 371 |

**Figure 5: Feature set derived from each of the 20s moving window. Blue-highlighted features are a parsimonious feature set found to be effective.**

the use of basic mean-level features, which are easily interpretable by humans, and compared them to less intuitive features (such as fNIRS skewness or GSR phasic mean AUC); and (3) Personalization: which refers to features derived from individualized data streams, normalized against mean values from each person's Non-VR baseline as described above in section

## 6.2 Effect of Feature Complexity

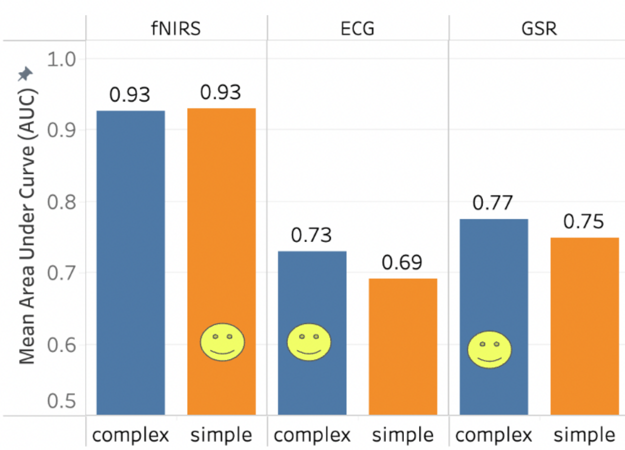


**Figure 6: Mean model performance from all pairs of experiments, comparing complex features against simple mean features for each sensing modality. The chosen feature sets are denoted by smiley faces.**

Figure 6 provides an overview of the model performance aggregated from all pairs of experiments (including those with personalized and non-personalized features). The figure compares models using simple features to those using complex features across various sensing modalities. The results indicate that for fNIRS, there is no discernable difference between complex and simple features. However, for ECG and GSR modalities, using simple features leads to a noticeable decline in model performance. As such, in subsequent evaluations, we only include simple features from fNIRS and complex features from ECG and GSR sensors. The multimodal feature set is a concatenation of these three sets of features, which includes a total of 51 features (32 fNIRS features, 15 GSR features, and 4 HR features, please refer to Figure 5 for details.)

## 6.3 Effect of Modality and Personalization

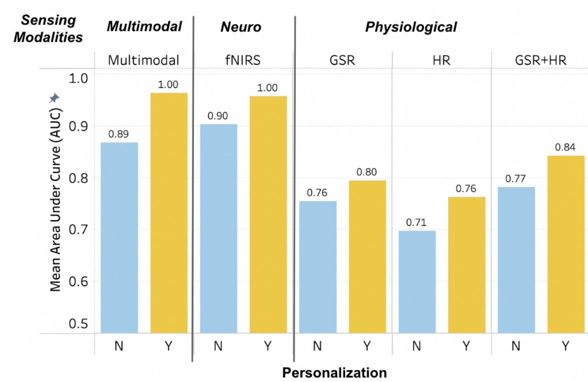


**Figure 7: Mean model performance across all 15 pairs of experiments, comparing personalized features to non-personalized features for each sensing modality.**

Figure 7 summarizes the average AUC score for model performance, comparing personalized and non-personalized features. As shown, personalized features consistently outperform non-personalized features across all modalities, individual or combined. This difference is more prominent in the neuro modality (i.e., fNIRS) compared to physiological modalities, including GSR and ECG, and their combination. Furthermore, we observed that the feature set derived from the fNIRS modality outperforms those from physiological modalities, and this pattern holds for both personalized and non-personalized features. When combining GSR and ECG signals, the resulting model performs better than those from individual modalities. However, when combining neuro and physiological modalities, we did not observe notable differences compared to fNIRS alone. However, the multimodal performance was notably better than the GSR and ECG or combined. This result suggests that the neuro modality provides sufficient information for discriminating among phases, possibly making GSR/ECG sensor-based physiological signals redundant.

## 6.4 Model Performance by Classification Tasks

Figure 8 provides a detailed comparison of model performance (measured by AUC score) across 15 classification tasks, each aiming to differentiate between two of the six phases. The results indicate that while most classification tasks achieve a high level of accuracy across modalities, distinguishing among the last three phases involving Speech Preparation, Speech Delivery, and Food Selection proves to be the most challenging. The most ambiguous pair is Speech Delivery and Food Selection. In this task, the multimodal model achieves an AUC score of around 72%, while GSR and HR

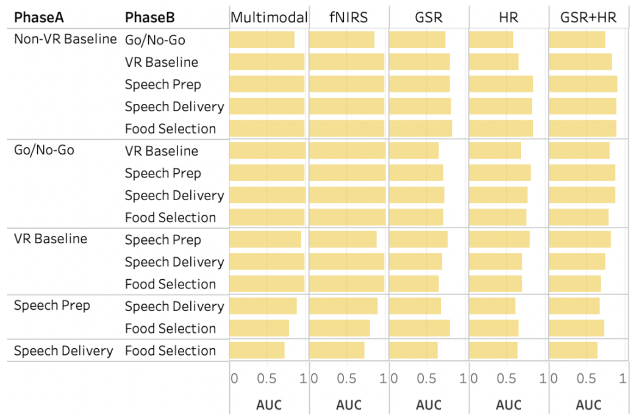


**Figure 8: A detailed comparison of model performance (measured by AUC score) across 15 classification tasks, each aiming to differentiate between two of the six phases, using personalized features.**

modalities achieve AUC scores of only 64%. A similar pattern is observed for fNIRS. This low level of discrimination suggests that the neuro-physiological signals between these two phases are more similar than those of other pairs.

On the other hand, given an AUC score higher than a random classifier of 50%, it implies that the cognitive/affective states (e.g., emotional stress and high cognitive load) experienced during speech preparation or delivery are notably different from those during food selection, which is possibly characterized by a mixture of emotional stress carried over from the previous phase and the cognitive process involved in food-related decisions.

## 6.5 Between-person Variation during the Last Three Phases in VR

Figure 9 further illustrates the between-individual variation in patterns of overlap among the last three phases in the VR environment, which includes Speech Preparation, Speech Delivery, and Food Selection. The plot was generated using UMAP (Uniform Manifold Approximation and Projection), a popular technique for dimensionality reduction and visualization of high-dimensional data, which was computed using the "umap" package in R.

Figure 10 summarizes the overlap patterns and corresponding participants belonging to each pattern. For participants belong to Pattern A, all three phases seem indistinguishable from one another, while for Pattern B, Speech Preparation is notably different from the other two phases, which are indistinguishable in between. Patterns A and B share a commonality in that the stress or arousal induced by the speech task possibly has a significant influence on the Food Selection task. This pattern of carrying over stress to Food Selection, however, is not observed in Pattern C. In this pattern, the signals from the two speech-related phases resemble each other but differ from those observed in the Food Selection phase. This could suggest that these participants recovered quickly from the stress induced by previous phases. Further studies could explore these overlapping patterns in relation to food choices (e.g. between healthy and unhealthy food items) as well as psychological measures such as effortful control that can be obtained from Go/No-Go tasks.

## 6.6 Effect of Personalization by Classification Tasks and by Modality

Figure 11 provides a detailed view of the influence of personalization on machine learning models' capacity to distinguish between any two of the six phases, analyzed by modalities. The effect is quantified by the percentage change in AUC scores. The result reveals variations in terms of both classification tasks and modalities. For example, the most considerable effect is observed when contrasting the effortful control cognitive task of Go/No-Go with other cognitive/affective tasks such as Speech Preparation, Speech Delivery, and Food Selection. This effect is especially pronounced for fNIRS and HR signals but not for GSR signals. Among the last three phases involving tasks in the VR environment, we see a notable personalization effect for GSR in discriminating between the Speech Preparation phase and Speech Delivery and Food Selection. However, this effect is not evident for other modalities, possibly because these tasks are related to arousal primarily detectable by GSR signals. Intriguingly, when the machine learning model differentiates between speech delivery and food selection tasks, the multimodal model exhibits the largest personalization effect, improving by about 20%. Overall, this analysis highlights the potential correlation between the personalization effect and detection tasks, which are moderated by the modalities used in the model.

# 7 DISCUSSION

In this secondary analysis of a rich multimodal neuro-physiological dataset collected from a multi-phase experiment in an immersive virtual reality (iVR) environment, we examined the impact of several critical machine learning-based model design decisions on a total of 15 classification tasks. These tasks aimed to discriminate among pairs of the six experimental phases, four of which occurred in the iVR environment. The design decisions we investigated included choices of modalities, feature sets, and personalization.

Our empirical analysis revealed that: (1) fNIRS, using only simple mean level features, outperformed other modalities; (2) while personalization generally improved results, its effect varied by modality and task and, in rare cases, even negatively impacted performance. We further observed that the personalization effect diminished when the classification task was more challenging.

By employing a non-linear manifold embedding method, we gained insights into between-person variation in terms of similarity/dissimilarity patterns among the last three phases, which proved relatively challenging to distinguish from a modeling perspective. These phases involved surprise speech preparation and delivery, and food selection. To some extent, the stress and cognitive load induced by the surprise speech task may have been carried over to the final phase. However, the "carry over" effect may vary from one individual to another, as evidenced by three distinct overlap patterns observed. This observation underscores the need to account for between-person variation in modeling neurophysiological responses.

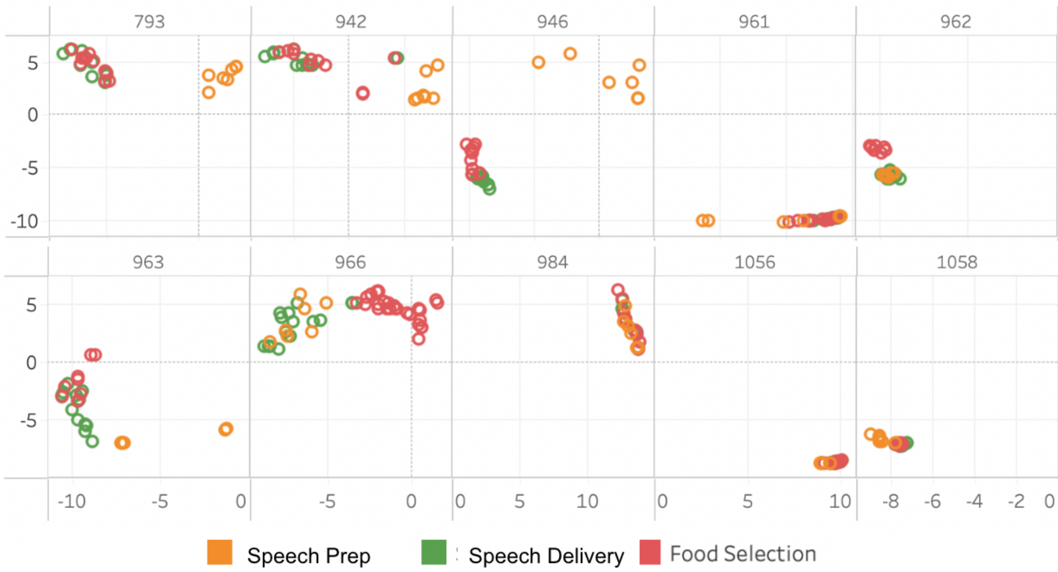


**Figure 9: The UMAP Embedding of multimodal features, with each grid representing data from one participant and colors denoting each of the last three phases.**

| Pattern ID | Overlap Patterns | Participant ID |
|---|---|---|
| A | [Speech Preparation, Speech Delivery, Food Selection] | 961, 984, 1056 and 1058 |
| B | Speech Preparation, [Speech Delivery, Food Selection] | 793, 942, 946 and 963 |
| C | [Speech Preparation, Speech Delivery], Food Selection | 962 and 966 |

**Figure 10: Overlappping Patterns summarized from the UMAP embedding and the list of participants belonging to each pattern. The square bracket groups similar phases together.**

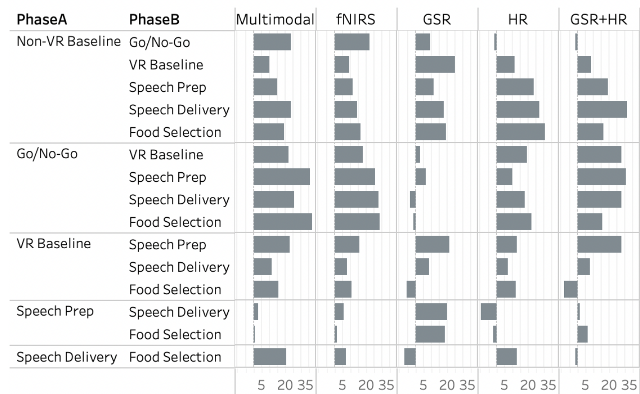


**Figure 11: Impact of personalization by classification tasks (in rows) and modalities (in columns), expressed as a percentage of improvement (+) or degradation (-) of AUC scores between models using personalized features versus non-personalized features.**

While personalizing feature sets is one possible solution, other avenues exist. For example, we could explicitly model individual differences using other personal-level characteristics, such as effortful control capacity, personality, or resilience. Such models would necessarily increase complexity and would be more suitable when we have access to a larger sample of participants with varying characteristics. This approach is planned as part of our future work.

## 8 CONCLUSION

As technology advances in both immersive virtual reality (iVR) and sensing capabilities, there is a growing need to develop practical guidelines for navigating complex model design decisions in order to create personalized and adaptive interventions for diverse users. Our study highlights the complexity and interactions within the design space when solving the reverse inference problem of predicting cognitive and affective states from neurophysiological signals. Although our results largely align with the literature on the benefits of multimodal and personalization approaches, they also shed light on the nuanced nature of these findings. This emphasizes the importance of meticulous experimentation when navigating the complex model space to address real-world problems, as demonstrated in this study.

## REFERENCES


[1] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems* 16 (2010), 345–379.

[2] Corey J. Bohil, Bradly Alicea, and Frank A. Biocca. 2011. Virtual reality in neuroscience research and therapy. *Nature Reviews Neuroscience* 12, 12 (Dec 2011), 752–762. https://doi.org/10.1038/nrn3122

[3] Ufuk Celikcan, Ahmed Şamil Bülbül, Cem Aslan, Zehra Buyuktuncer, Kübra Işgın, Gözde Ede, and Nuray Kanbur. 2018. The virtual cafeteria: an immersive environment for interactive food portion-size education. In *Proceedings of the 3rd International Workshop on Multisensory Approaches to Human-Food Interaction*. 1–5.

[4] Charissa SL Cheah, Salih Barman, Kathy TT Vu, Sarah E Jung, Varun Mandalapu, Travis D Masterson, Ryan J Zuber, Lee Boot, and Jiaqi Gong. 2020. Validation of a virtual reality buffet environment to assess food selection processes among emerging adults. *Appetite* 153 (2020), 104741.

[5] Cheng Chen, Yizhen Wen, Shaoyang Cui, Xiangao Qi, Zhenhong Liu, Linfeng Zhou, Mingyi Chen, Jian Zhao, and Guoxing Wang. 2020. A Multichannel fNIRS System for Prefrontal Mental Task Classification with Dual-level Excitation and Deep Forest Algorithm. *Journal of Sensors* 2020 (Jun 2020), 1–10. https://doi.org/10.1155/2020/1567567

[6] Damien Clus, Mark Erik Larsen, Christophe Lemey, and Sofian Berrouiguet. 2018. The use of virtual reality in patients with eating disorders: systematic review. *Journal of medical Internet research* 20, 4 (2018), e157.

[7] Desirée Colombo, Amanda Díaz-García, Javier Fernandez-Álvarez, and Cristina Botella. 2021. Virtual reality for the enhancement of emotion regulation. *Clinical Psychology & Psychotherapy* 28, 3 (2021), 519–537.

[8] Muhammad Najam Dar, Muhammad Usman Akram, Sajid Gul Khawaja, and Amit N Pujari. 2020. CNN and LSTM-based emotion charting using physiological signals. *Sensors* 20, 16 (2020), 4551.

[9] Sophie Melissa Clare Davison, Catherine Deeprose, and Sylvia Terbeck. 2018. A comparison of immersive virtual reality with traditional neuropsychological measures in the assessment of executive functions. *Acta Neuropsychiatrica* 30, 2 (2018), 79–89.

[10] Essam Debie, Raul Fernandez Rojas, Justin Fidock, Michael Barlow, Kathryn Kasmarik, Sreenatha Anavatti, Matt Garratt, and Hussein A Abbass. 2019. Multimodal fusion for objective assessment of cognitive workload: a review. *IEEE transactions on cybernetics* 51, 3 (2019), 1542–1555.

[11] Xue Deng, Chuyao Jian, Qinglu Yang, Naifu Jiang, Zhaoyin Huang, and Shaofeng Zhao. 2022. The analgesic effect of different interactive modes of virtual reality: A prospective functional near-infrared spectroscopy (fNIRS) study. *Frontiers in Neuroscience* 16 (Nov 2022), 1033155. https://doi.org/10.3389/fnins.2022.1033155

[12] Anna Ferrari, Daniela Micucci, Marco Mobilio, and Paolo Napoletano. 2020. On the Personalization of Classification Models for Human Activity Recognition. *IEEE Access* 8 (2020), 32066–32079.

[13] Marta Ferrer-García and José Gutiérrez-Maldonado. 2012. The use of virtual reality in the treatment of eating disorders. *Studies in Health Technology and Informatics* 181 (2012), 17–21.

[14] Marta Ferrer-García, José Gutiérrez-Maldonado, Alejandra Caqueo-Urízar, and Elena Moreno. 2009. The Validity of Virtual Environments for Eliciting Emotional Responses in Patients With Eating Disorders and in Controls. *Behavior Modification* 33, 6 (Nov 2009), 830–854. https://doi.org/10.1177/0145445509348056

[15] Kunal Gupta, Jovana Lazarevic, Yun Suen Pai, and Mark Billinghurst. 2020. AffectivelyVR: Towards VR Personalized Emotion Recognition. In *26th ACM Symposium on Virtual Reality Software and Technology*. ACM, Virtual Event Canada, 1–3. https://doi.org/10.1145/3385956.3422122

[16] Angela R. Harrivel, Daniel H. Weissman, Douglas C. Noll, Theodore Huppert, and Scott J. Peltier. 2016. Dynamic filtering improves attentional state prediction with fNIRS. *Biomedical Optics Express* 7, 3 (Mar 2016), 979. https://doi.org/10.1364/BOE.7.000979

[17] Angela R. Harrivel, Daniel H. Weissman, Douglas C. Noll, and Scott J. Peltier. 2013. Monitoring attentional state with fNIRS. *Frontiers in Human Neuroscience* 7 (2013). https://doi.org/10.3389/fnhum.2013.00861

[18] Sarah J. Heany, Nynke A. Groenewold, Anne Uhlmann, Shareefa Dalvie, Dan J. Stein, and Samantha J. Brooks. 2018. The neural correlates of Childhood Trauma Questionnaire scores in adults: A meta-analysis and review of functional magnetic resonance imaging studies. *Development and Psychopathology* 30, 4 (Oct 2018), 1475–1485. https://doi.org/10.1017/S0954579417001717

[19] Fabian Herold, Patrick Wiegel, Felix Scholkmann, and Notger Müller. 2018. Applications of Functional Near-Infrared Spectroscopy (fNIRS) Neuroimaging in Exercise–Cognition Science: A Systematic, Methodology-Focused Review. *Journal of Clinical Medicine* 7, 12 (Nov 2018), 466. https://doi.org/10.3390/jcm7120466

[20] Wenxing Hu, Xianghe Meng, Yuntong Bai, Aiying Zhang, Gang Qu, Biao Cai, Gemeng Zhang, Tony W Wilson, Julia M Stephen, Vince D Calhoun, et al. 2021. Interpretable multimodal fusion networks reveal mechanisms of brain cognition. *IEEE transactions on medical imaging* 40, 5 (2021), 1474–1483.

[21] Syem Ishaque, Alice Rueda, Binh Nguyen, Naimul Khan, and Sridhar Krishnan. 2020. Physiological Signal Analysis and Classification of Stress from Virtual Reality Video Game. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*. IEEE, Montreal, QC, Canada, 867–870. https://doi.org/10.1109/EMBC44109.2020.9176110

[22] Yi Li, Adel S. Elmaghraby, Ayman El-Baz, and Estate M. Sokhadze. 2015. Using physiological signal analysis to design affective VR games. In *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, Abu Dhabi, United Arab Emirates, 57–62. https://doi.org/10.1109/ISSPIT.2015.7394401

[23] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. 2021. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods* 53, 4 (feb 2021), 1689–1696. https://doi.org/10.3758/s13428-020-01516-y

[24] Stephanie M Manasse, Elizabeth W Lampe, Adrienne S Juarascio, Jichen Zhu, and Evan M Forman. 2021. Using virtual reality to train inhibitory control and reduce binge eating: A proof-of-concept study. *Appetite* 157 (2021), 104988.

[25] Tim R Mullen, Christian AE Kothe, Yu Mike Chi, Alejandro Ojeda, Trevor Kerth, Scott Makeig, Tzyy-Ping Jung, and Gert Cauwenberghs. 2015. Real-time neuroimaging and cognitive monitoring using wearable dry EEG. *IEEE Transactions on Biomedical Engineering* 62, 11 (2015), 2553–2567.

[26] Md Hasin Raihan Rabbani and Sheikh Md Rabiul Islam. 2021. Multimodal Decision Fusion of EEG and fNIRS Signals. In *2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*. IEEE, 1–6.

[27] Giuseppe Riva, Clelia Malighetti, and Silvia Serino. 2021. Virtual reality in the treatment of eating disorders. *Clinical psychology & psychotherapy* 28, 3 (2021), 477–488.

[28] Barb Ruppert. 2011. New directions in the use of virtual reality for food shopping: marketing and education perspectives. *Journal of Diabetes Science and Technology* 5, 2 (Mar 2011), 315–318. https://doi.org/10.1177/193229681100500217

[29] Yangming Shi, Yibo Zhu, Ranjana K. Mehta, and Jing Du. 2020. A neurophysiological approach to assess training outcome under stress: A virtual reality experiment of industrial shutdown maintenance using Functional Near-Infrared Spectroscopy (fNIRS). *Advanced Engineering Informatics* 46 (Oct 2020), 101153. https://doi.org/10.1016/j.aei.2020.101153

[30] Pekka Siirtola, Heli Koskimäki, and Juha Röning. 2019. Personalizing human activity recognition models using incremental learning. (May 2019). http://arxiv.org/abs/1905.12628 arXiv:1905.12628 [cs].

[31] Yanjia Sun, Hasan Ayaz, and Ali N Akansu. 2020. Multimodal affective state assessment using fNIRS+ EEG and spontaneous facial expression. *Brain Sciences* 10, 2 (2020), 85.

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

[33] Annika Wiebe, Kyra Kannen, Benjamin Selaskowski, Aylin Mehren, Ann-Kathrin Thöne, Lisa Pramme, Nike Blumenthal, Mengtong Li, Laura Asché, Stephan Jonas, et al. 2022. Virtual reality in the diagnostic and therapy for mental disorders: A systematic review. *Clinical Psychology Review* (2022), 102213.

[34] Chengyan Xu, Michael Siegrist, and Christina Hartmann. 2021. The application of virtual reality in food consumer behavior research: A systematic review. *Trends in Food Science & Technology* 116 (2021), 533–544.