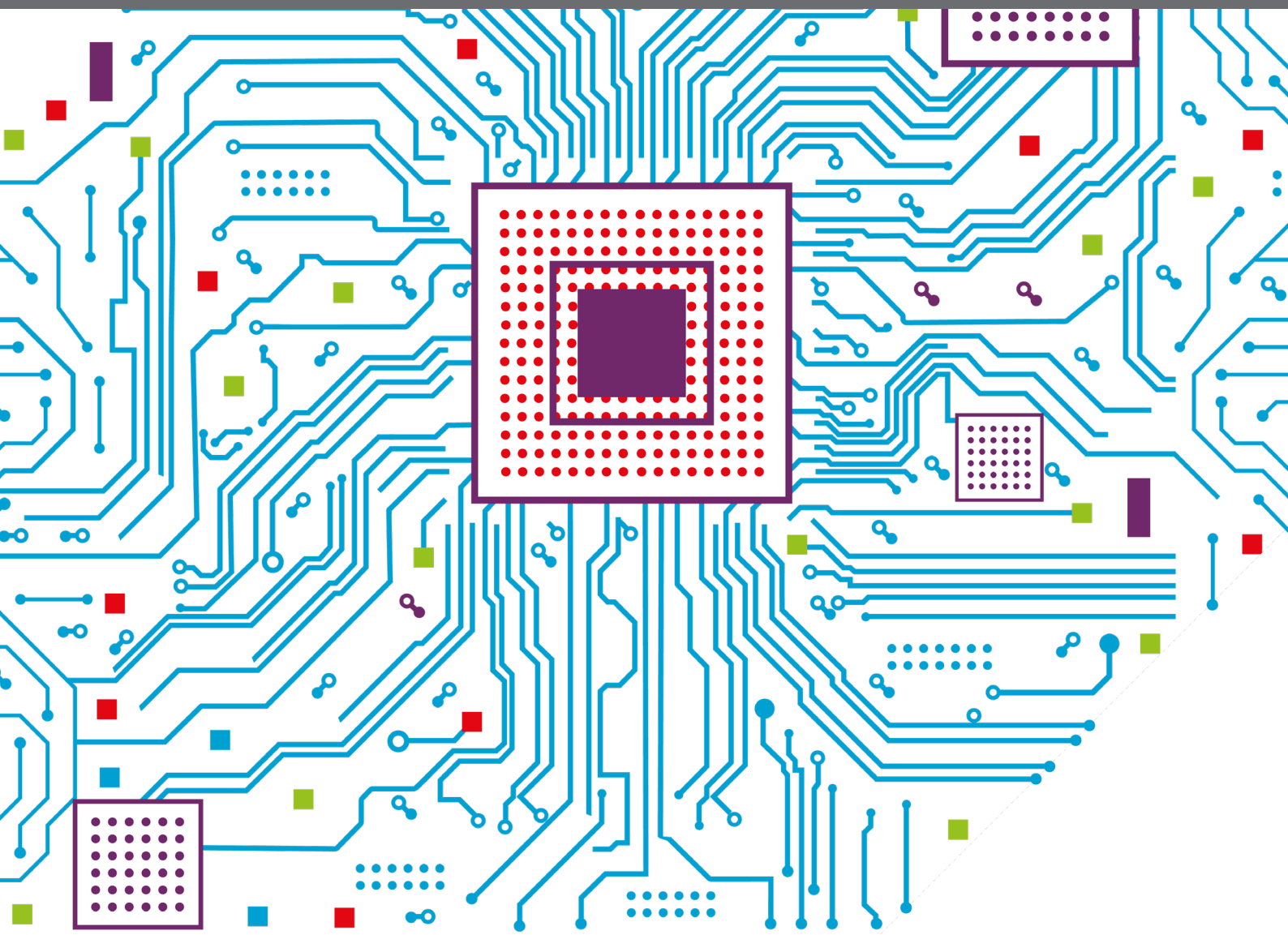
RECOGNIZING THE STATE OF EMOTION, COGNITION AND ACTION FROM PHYSIOLOGICAL AND BEHAVIOURAL SIGNALS

EDITED BY: Siyuan Chen, Youngjun Cho, Kun Yu, Laura M. Ferrari and Francois Bremond

PUBLISHED IN: Frontiers in Computer Science and Frontiers in Psychology



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88976-948-3
DOI 10.3389/978-2-88976-948-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

RECOGNIZING THE STATE OF EMOTION, COGNITION AND ACTION FROM PHYSIOLOGICAL AND BEHAVIOURAL SIGNALS

Topic Editors:
Siyuan Chen, University of New South Wales, Australia
Youngjun Cho, University College London, United Kingdom
Kun Yu, University of Technology Sydney, Australia
Laura M. Ferrari, Université Côte d'Azur, France
Francois Bremond, Institut National de Recherche en Informatique et en Automatique (INRIA), France

Citation: Chen, S., Cho, Y., Yu, K., Ferrari, L. M., Bremond, F., eds. (2022). Recognizing the State of Emotion, Cognition and Action From Physiological and Behavioural Signals. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88976-948-3

Table of Contents

frontiers | Frontiers in Computer Science

TYPE Editorial
PUBLISHED 03 August 2022
DOI 10.3389/fcomp.2022.998416

OPEN ACCESS

EDITED AND REVIEWED BY
Anton Nijholt,
University of Twente, Netherlands

*CORRESPONDENCE
Siyuan Chen
siyuan.chen@unsw.edu.au

SPECIALTY SECTION
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

RECEIVED 19 July 2022
ACCEPTED 20 July 2022
PUBLISHED 03 August 2022

CITATION
Chen S, Cho Y, Yu K, Ferrari LM and Bremond F (2022) Editorial: Recognizing the state of emotion, cognition and action from physiological and behavioral signals. *Front. Comput. Sci.* 4:998416. doi: 10.3389/fcomp.2022.998416

Editorial: Recognizing the state of emotion, cognition and action from physiological and behavioral signals

Siyuan Chen[1]*, Youngjun Cho[2], Kun Yu[3], Laura M. Ferrari[4] and Francois Bremond[5]

[1]Department of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW, Australia, [2]Department of Computer Science, University College London, London, United Kingdom, [3]The Data Science Institute, University of Technology Sydney, Sydney, NSW, Australia, [4]STARS Team, INRIA, Université Côte d'Azur, Nice, France, [5]STARS Team, INRIA, Sophia Antipolis, France

KEYWORDS

computational psychophysiology, affective computing, human computer interaction, human factors, assistive technology, signal processing, pattern recognition, machine learning

Editorial on the Research Topic
Recognizing the state of emotion, cognition and action from physiological and behavioral signals

Seamless blending of humans and technology for intelligent interaction is becoming more popular. One key aspect is to let machine understand users' state of emotion, cognition, and action. This Research Topic is a collection of ten papers where physiological and behavioral signals are exploited to recognize user states. In this collection, multiple techniques, systems, and applications are introduced, spanning from healthcare (e.g., dementia, disorientation in aged people, alpha waves asymmetry), workload, sleep monitoring and self-care assistive technology, to decision-making tasks (e.g., relevance of text read, relational communication, emotion classification). We highlight the main findings of these research studies.

A multidisciplinary research team from the UCL Interaction Centre (Holloway et al.) proposes a new cost-effective approach with Inertial Measurement Units (IMU) sensors to predict dementia. The results demonstrate state-of-the-art performance in classifying data from different dementia groups including typical Alzheimer's disease and posterior cortical atrophy. This approach paves the way for a simple clinical test to enable dementia screening in real-world.

Researchers at the University of Bremen (Steinert et al.) conduct a study on the prediction of activation ratings of people with dementia, which has been shown to be a possible cue of cognitive functioning. With an existing dataset that includes verbal and non-verbal cues of people with dementia, the team demonstrates the positive contribution of behavioral cues to the prediction and discusses unique challenges in the task.

Teipel et al. study the features of gait and accelerometry associated with disorientation events. The orientation ability of older and younger cognitively normal participants navigating on a treadmill is under investigation. Although the strength of the association of currently studied features is not sufficient for accurate real-time prediction of disorientation in a single individual, it paves the way for a future system that allows monitoring the orientation, the gait, the accelerometric and physiological data in a controlled environment.

To better understand and apply the theory of alpha asymmetry, Sabu et al. conduct a review on the role of affective stimuli in event-related frontal alpha asymmetry. They confirm that strongly engaging, salient and/or personally relevant stimuli are important to induce an approach-avoidance effect. Meanwhile, the selection of stimuli accounts for part of the diversity in alpha asymmetry research, where notably, multimodal stimuli and stimuli employing tasks induce approach-avoidance effects more strongly than images.

A collaborative team (Meteier et al.) from Switzerland investigates the use of physiological data to assess mental workload in the context of automated driving. The team confirms that respiratory indicators and heart rate variability are effective measures of mental workload and highlights the possible relationship between task performance and mental workload prediction.

The author Liang investigates the relationship between brain hemodynamics and stress in the first sleep cycle. Chemical biomarkers and novel wearables for near-infrared spectroscopy are coupled with machine learning in a new research paradigm. The study sheds light on the possible role of the left rostral and dorsolateral prefrontal cortex in stress responses.

Barz et al. conduct a study on estimating paragraph relevance from eye movement. They confirm that eye gaze can be used to estimate the perceived relevance of short news articles although there is no evidence to clearly show that the approach generalizes to multi-paragraph documents when users scroll down to see all text passages. It can be envisaged that the gaze-based relevance detection can be a part of future adaptive user interfaces that leverage multiple sensors for behavioral signal processing and analysis.

Vortmann et al. compare early, middle, and late fusion in a classification task to infer internal (e.g., thought, memories) or external (e.g., sensory input) attentional state. The dataset used in this study is multimodal and composed of EEG and eye tracking. The results indicate that middle or late fusion are better suited than early fusion approaches.

Burgoon et al. apply the Brunswikian lens model of relational communication, which measures linguistic, vocalic, and facial cues, to establish a perception of other people on relational attributes (dominance, affection, composure, involvement, similarity, trust) and quantify their perceived credibility while participants are interacting in game of Resistance. They find that the behavior elicited during the activity correlates with relational messages in a supportive manner, such as the correlations between affection and longer sentences and less hedging.

The research conducted by Menétrey et al. from University of Geneva and University of New South Wales aims to identify key components contributing to accurate emotion prediction. They highlight that emotion recognition requires the integration of various components (appraisal, motivation, expression, physiology, and feeling). In this study they extract mean and variance of the physiological data and show that emotional features are encoded within the other components.

We hope the readers enjoy this topic collection. These studies demonstrate a growing interest in empowering machine to understand user state and a multidisciplinary approach to improve human and machine collaboration in the best form.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

frontiers in Computer Science

ORIGINAL RESEARCH
published: 02 December 2021
doi: 10.3389/fcomp.2021.774949

What Does Sleeping Brain Tell About Stress? A Pilot Functional Near-Infrared Spectroscopy Study Into Stress-Related Cortical Hemodynamic Features During Sleep

Zilu Liang[1,2]*

[1]Ubiquitous and Personal Computing Laboratory, Faculty of Engineering, Kyoto University of Advanced Science, Kyoto, Japan, [2]Institute of Industrial Science, The University of Tokyo, Tokyo, Japan

People with mental stress often experience disturbed sleep, suggesting stress-related abnormalities in brain activity during sleep. However, no study has looked at the physiological oscillations in brain hemodynamics during sleep in relation to stress. In this pilot study, we aimed to explore the relationships between bedtime stress and the hemodynamics in the prefrontal cortex during the first sleep cycle. We tracked the stress biomarkers, salivary cortisol, and secretory immunoglobulin A (sIgA) on a daily basis and utilized the days of lower levels of measured stress as natural controls to the days of higher levels of measured stress. Cortical hemodynamics was measured using a cutting-edge wearable functional near-infrared spectroscopy (fNIRS) system. Time-domain, frequency-domain features as well as nonlinear features were derived from the cleaned hemodynamic signals. We proposed an original ensemble algorithm to generate an average importance score for each feature based on the assessment of six statistical and machine learning techniques. With all channels counted in, the top five most referred feature types are Hurst exponent, mean, the ratio of the major/minor axis standard deviation of the Poincaré plot of the signal, statistical complexity, and crest factor. The left rostral prefrontal cortex (RLPFC) was the most relevant sub-region. Significantly strong correlations were found between the hemodynamic features derived at this sub-region and all three stress indicators. The dorsolateral prefrontal cortex (DLPFC) is also a relevant cortical area. The areas of mid-DLPFC and caudal-DLPFC both demonstrated significant and moderate association to all three stress indicators. No relevance was found in the ventrolateral prefrontal cortex. The preliminary results shed light on the possible role of the RLPCF, especially the left RLPCF, in processing stress during sleep. In addition, our findings echoed the previous stress studies conducted during wake time and provides supplementary evidence on the relevance of the dorsolateral prefrontal cortex in stress responses during sleep. This pilot study serves as a proof-of-concept for a new research paradigm to stress research and identified exciting opportunities for future studies.

Keywords: ubiquitous computing, functional near-infrared spectroscopy (fNIRS), stress, prefrontal cortex, wearable computing, sleep

OPEN ACCESS

Edited by:
Laura M. Ferrari,
Université Côte d'Azur, France

Reviewed by:
Zhen Yuan,
University of Macau, China
Min Cheol Chang,
Yeungnam University, South Korea

***Correspondence:**
Zilu Liang
liang.zilu@kuas.ac.jp

Specialty section:
This article was submitted to Mobile and Ubiquitous Computing, a section of the journal Frontiers in Computer Science

Received: *13 September 2021*
Accepted: *29 October 2021*
Published: *02 December 2021*

Citation:
Liang Z (2021) What Does Sleeping Brain Tell About Stress? A Pilot Functional Near-Infrared Spectroscopy Study Into Stress-Related Cortical Hemodynamic Features During Sleep. Front. Comput. Sci. 3:774949. doi: 10.3389/fcomp.2021.774949

1 INTRODUCTION

There is abundant evidence that mental stress is often linked to reduced sleep quality, suggesting abnormalities in brain activity during sleep when people are stressed (Buysse et al., 2011). While our understanding into how stress affects brain activity when we are awake (and are engaged in lab-based stress induction tasks) has been greatly advanced in recent years (Alonso et al., 2015; Kramer et al., 2017; Chang and Yu, 2018; Rosenbaum et al., 2018; Rampino et al., 2019; Schaal et al., 2019; Rosenbaum et al., 2021), no study has looked at how stress modulates brain activity during sleep. Attempts to study stress during sleep face several challenges. First, traditional neuroimaging techniques for studying stress in daytime are not suited for in-sleep measurement due to various methodological restraints imposed by these techniques. Hemodynamic imaging methods such as functional magnetic resonance imaging (fMRI) can generate hemodynamic profiles at high spatial resolution, but they are invasive as people could hardly fall asleep in noisy fMRI scanners. Functional near-infrared spectroscopy (fNIRS) achieves better trade-off between convenience and spatial resolution, but traditional fNIRS systems still use many cables which make them unsuited for measurement during sleep. Electrophysiological neuroimaging techniques such as electroencephalography (EEG) is widely used to measure brain activity during sleep, but their spatial resolution is limited, and they are sensitive to motion artifacts. Second, laboratory-induced stress response is often temporary, and the effect could barely sustain until and throughout nocturnal sleep (Rosenbaum et al., 2021). Established methods for inducing social stress (e.g., the Trier Social Stress Test (Chang and Yu, 2018)), emotional stress (e.g., viewing scary pictures (Rampino et al., 2019)), and physical stress (e.g., sleep deprivation (Alonso et al., 2015)) may fail to mirror natural stress responses, as a laboratory setting often does not represent the typical conditions under which stress occurs in real life (Wolfram et al., 2013). Laboratory-induced stress responses often fade out in an hour (Rosenbaum et al., 2021; Rosenbaum et al., 2018), while real-life stress responses could last hours to days or even longer after the onset of the stressors (Joëls and Baram, 2009). Another pitfall of lab-based stress induction protocols is that they are unsuited for longitudinal repeated measurement from individual subjects as they are likely to cause response habituation especially in the hypothalamic–pituitary–adrenal (HPA) axis as indicated by the cortisol secretion level (Schommer et al., 2003; Kudielka et al., 2006; Jönsson et al., 2010; Gianferante et al., 2014). In addition, stress has been routinely treated as a dichotomous variable (i.e., stress is either present or absent) in many research studies. In real life, however, people may experience various levels of stress with different temporal profiles (Joëls and Baram, 2009). The dichotomous perspective of stress is also unnatural when biomarkers of stress responses such as cortisol is used as an indicator because it is difficult to set a universal cutoff line that accommodates interpersonal variability.

This pilot study is the first to look at how bedtime stress associates to brain activity during sleep. We aimed to explore which cortical areas demonstrate stress-related blood flow patterns during the first sleep cycle. Especially, we focused on answering the following two research questions:

- What hemodynamic features are significantly associated to each stress indicator?
- Which sub-regions in the PFC are significantly associated to each stress indicator?

The study design included addressing the limitations of the existing research paradigm. **Table 1** highlights the originality of the present study in comparison with previous studies. This study adopted the N-of-1 approach which is an idiographic research methodology that overcomes the pitfalls of the widely adopted large-sample approach. The large-sample approach requires stringent conditions such as cohort homogeneity—a condition difficult to meet no matter how large the sample size is. When the within-subject variability is much larger than the inter-subject variability, which is common in psychology and physiology studies, the large-sample approach often fails to provide us with findings that generalize well to individuals (Molenaar, 2004; Barlow and Nock, 2009; Mehl and Conner, 2012; van Ockenburg et al., 2015; Burg et al., 2017; Fishera et al., 2018; Piccirillo et al., 2019). In contrast, the N-of-1 approach embraces longitudinal repeated measurement on a single subject to generate the most relevant and reliable information for the specific person, which *represents a true scientific undertaking* (Barlow and Nock, 2009).

With respect to the experiment settings, we performed the measurement at the subject's home using a cutting-edge wearable fNIRS system together with non-invasive wearable and mobile devices to achieve the highest level of ecological validity. We also did not rely on lab-based stress induction protocols, as they require the subjects to be actively engaged in cognitive tasks and often fail to induce stress responses that sustain until bedtime. Instead, we tracked the stress indicators on a daily basis, and utilized the days of lower levels of measured stress as natural controls to the days of higher levels of measured stress. Stress responses in human may manifest in multiple physiological systems with varied temporal profile (Joëls and Baram, 2009). In this study, stress was quantified using both objective and subjective indicators. The objective indicators included two widely used stress biomarkers that reflect the hormonal and immunological responses to stress: salivary cortisol and secretory immunoglobulin A (sIgA). The rise of salivary cortisol reveals the stress-related changes in the hypothalamic–pituitary–adrenal (HPA) axis. Meanwhile, stress-associated immunological response could occur more rapidly compared to the HPA axis, characterized by a quick and temporal rise and then decrease in sIgA (Engeland et al., 2016). sIgA may also be a valuable indicator for differentiating between positive and negative stress effects or between successful and unsuccessful adaptation or coping with situational demands (Zeier et al., 1996). The subjective perception of stress can be measured using psychometric instruments ranging from as simple as a Likert scale to as complex as the 30-item Perceived Stress Questionnaire (PSQ) (Levenstein et al., 1993). In this study, the perceived stress level was rated on a 1–10 Likert scale that was

TABLE 1 | Research paradigm comparison between the current study and previous studies.

	Previous studies	Current study
Approach	Large-sample (nomothetic)	N-of-1 (idiographic)
Data collection	1. Experiment was performed in a lab using bulky equipment	1. Experiment was performed at the subject's home using wearable and mobile devices
	2. Lab-induced stress	2. Stress occurs in daily life setting
	3. Subjects were awake and engaged in cognitive tasks	3. Subject was sleeping
Data analysis	1. Limited features derived from hemodynamic signals (usually the mean)	1. A wide range of features derived to characterize the hemodynamic patterns
	2. Stress treated as dichotomous variable (either 1 or 0)	2. Stress treated as a continuous/ordinal variable
	3. Basic statistical test used to find inter-group differences	3. Original ensemble algorithm for feature ranking
Ecological validity	Low	High

implemented using a mobile application. Another significant difference of this study is that the measurement was mostly performed when the subject was sleeping, while in previous stress neuroimaging studies the subjects were all awake and were engaged in cognitive tasks.

In this study, stress response was treated as a continuous phenomenon in contrast to the traditional dichotomous perspective. The data analysis focused on finding significant associations between cortical hemodynamic features and each individual stress indicator. Previous studies mostly rely on one feature type—the mean of the concentration changes in oxyhemoglobin (ΔO_2Hb) and deoxyhemoglobin (ΔHHb). While this feature type has the merit of easy interpretation, it fails to fully capture the characteristics of the cortical hemodynamic signals. In search for the most useful stress-association features, we derived a wide range of time-domain, frequency-domain, and nonlinear features from the cortical hemodynamic signals. We also proposed an original ensemble feature ranking algorithm that leverages six different statistical and machine learning techniques to generate an average importance score for each feature.

This pilot study does not intend to generate conclusive findings, but rather serves as a proof-of-concept for a new research paradigm that can be implemented to study stress in unexplored settings (e.g., during sleep). Understanding the neurophysiological mechanism that underlies the relation between stress and sleep has the significance of giving hint to the development of brain activity markers of stress, which can be readily measured and monitored using wearable brain imaging technologies. Despite of being a small-scale pilot study, the data collection and analysis protocols are readily applicable to large-scale studies. The observations from this study serve as a foundation for future research to elucidate where the brain processes stress during sleep, based on which new stress indicators or stress coping strategies may be developed.

2 DATA COLLECTION

2.1 Measuring Stress

In this study, we quantified stress using both objective (i.e., cortisol and sIgA) and subjective indicators (i.e., perceived stress rating). Salivary cortisol and sIgA were measured using the SOMA Dual Analyte LFD test kits. These kits can be used for real-time measurement in a naturalistic setting. Saliva samples were collected using oral fluid collector (OFC) swabs and were incubated for 15 min in OFC buffers before being read. The participant was instructed not to eat, drink, or brush teeth 30 min prior to providing saliva samples. The calibration range of cortisol and sIgA were 1.25–40 nmol/L and 25–800 μg/ml, respectively (Dunbar et al., 2015). The validity of the SOMA kits has been examined in previous studies (Mitsuishi et al., 2019). The measured salivary cortisol and sIgA data were manually logged in a CSV file. Perceived stress was rated on a 1–10 Likert scale (1 = not stressed at all; 10 = extremely stressed) which was implemented using a mobile application named HealthLog.

2.2 Measuring Prefrontal Hemodynamics

A wearable functional near-infrared spectroscopy (fNIRS) (Brite 24; Artinis Medical Systems Co., Netherlands) was used to measure the concentration changes in oxyhemoglobin (ΔO_2Hb) and deoxyhemoglobin (ΔHHb) in the PFC. The fNIRS is a non-invasive brain imaging technique that strikes a good trade-off between temporal and spatial resolution (Tak and Ye, 2014). The advantage of the Brite 24 system—which weighs only 300 g—is that it permits the monitoring of ΔO_2Hb and ΔHHb without imposing constraints on the posture and movement of the subject, and thus is suited for studying cortical hemodynamics during sleep. In this study, the Brite 24 consists of 10 transmitters (Tx) and 8 receivers (Rx). The Txs take turns to emit light at wavelengths of 760 nm (dominantly absorbed by HHb) and 850 nm (dominantly absorbed by O_2Hb). They were fixed on a soft neoprene head cap, which ensures the alignment of optode placement across different measurements. The optodes were placed at an interoptode distance of 3 cm to achieve the maximum penetration depth of 1.5 cm and were configured into 27 channels as shown in the *Template DAQ state* at the bottom of **Figure 1**. All optodes were placed between the FpZ–F3–Cz–F4–FpZ regions in the PFC according to the international 10–20 EEG system. The sampling rate was set to 50 Hz. The Brite 24 device has a battery life of up to 2.5 h when it is used for continuous online measurement. While the battery life could be extended

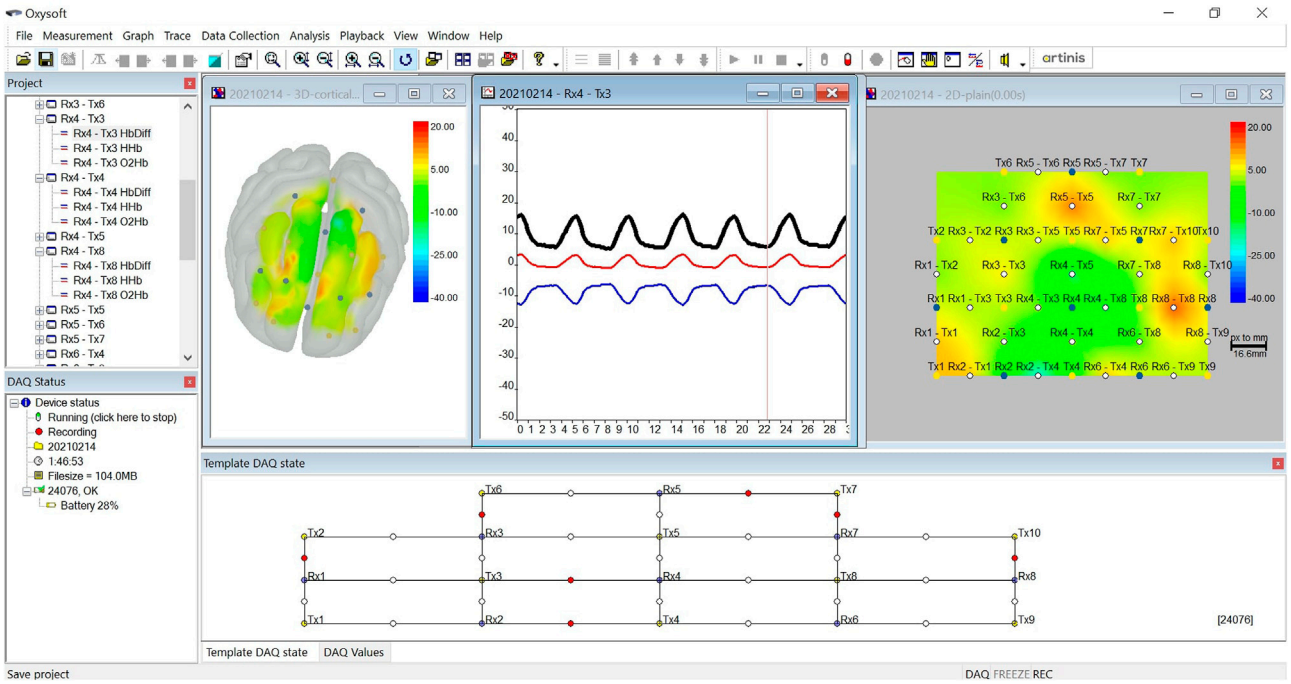


FIGURE 1 | Screenshot of the OxySoft.

by connecting the device to an external power supply, we decided not to use that strategy out of safety concern for the subject. The PFC was selected as the region of interest in this pilot study because previous studies performed during wake time have shed light on the role of the PFC in responding to acute and chronic stress (Cerqueira et al., 2007; Hains and Arnsten, 2008; Dedovic et al., 2009; Yuen et al., 2009; Arnsten et al., 2015; Nejati et al., 2021).

The Brite 24 system consists of companion software named OxySoft. The software allows the real-time inspection of the signal quality of each channel when the fNIRS device is paired up *via* Bluetooth connection. Channels with poor quality are marked by red dots, as shown at the bottom of **Figure 1**. The OxySoft supports several visualization formats of the ΔO_2Hb and ΔHHb signals, including time series plots, 2D heatmap, and 3D heatmap in a glass head. The data recorded by the Brite 24 device were synchronized with the software at regular time interval and were stored in temporal files. When a measurement was stopped, OxySoft processed the temporal files to generate a complete data file that contained raw optical density (OD) data.

2.3 Measuring Complementary Physiological Data

In addition to the Brite 24 system, we also used a Fitbit Sense together with the companion Fitbit app to collect complementary data of sleep, heart rate, and breath rate. These data were utilized in the data preprocessing pipeline to remove physiological artifacts, which is described in detail in the next section. Fitbit is well-suited to this study as it supports the collection of multiple streams of physiological signals without imposing additional burden to the subject. Despite that Fitbit devices may not offer medical-grade measurements, numerous validation studies have demonstrated that Fitbit devices can achieve reasonable accuracy and a better trade-off between accuracy and ecological validity (Menghini et al., 2020; Liang and Chapa-Martell, 2019; Liang and Chapa-Martell, 2018).

2.4 Data Collection Procedure

Grounded on the N-of-1 approach, a longitudinal data collection experiment was conducted with a healthy subject (male, 30 years). The principle of the N-of-1 method allows the exclusion of confounding factors pertaining interpersonal differences in health and physiological conditions. In comparison, the traditional large-sample approach requires stringent conditions such as cohort homogeneity and the findings often do not generalize well to individuals (Molenaar, 2004; Barlow and Nock, 2009; van Ockenburg et al., 2015; Fishera et al., 2018). The large-sample approach becomes especially problematic when the variability within subject is much larger than the variability across subjects (Mehl and Conner, 2012; Fishera et al., 2018). There has been increasing evidence that the large-sample approach may not provide us with information that generalizes well to individuals (Molenaar and Campbell, 2009; Burg et al., 2017; Piccirillo et al., 2019), and hence should not be deemed as more scientific than other approaches that explicitly address within-person variability (Mehl and Conner, 2012). On the other hand, the N-of-1 approach has been well-recognized to provide the

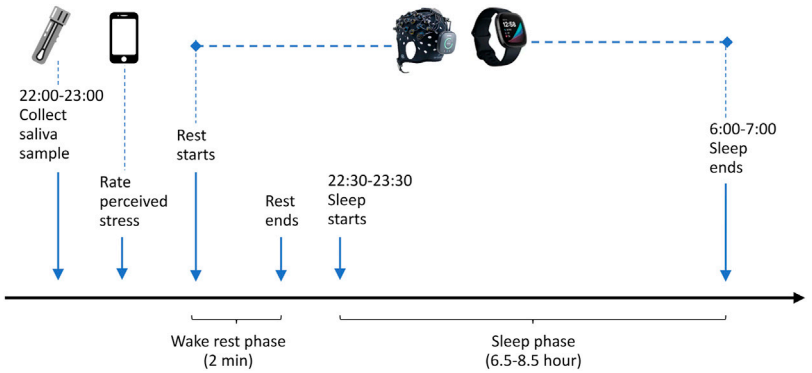


FIGURE 2 | Data collection procedure.

highest reliability at the individual level (Molenaar, 2004; Molenaar and Campbell, 2009; Mehl and Conner, 2012; van Ockenburg et al., 2015). The subject was recruited through personal connections. The inclusion criteria were 1) healthy subjects aged 18–65 years without chronic diseases, sleep disorders, and mental disorders, 2) has a smartphone, and 3) understand the contents of the informed consent. This study was approved by the Ethics Committee of the Kyoto University of Advanced Science. Written informed consent was obtained from the subject before the data collection experiment started.

The data collection procedure is illustrated in **Figure 2**. Saliva samples were collected before bedtime at night. We ensured that saliva sample collection was always done during a fixed time period 22:00–23:00 to control the confounding effect of the circadian hormonal rhythm (Oster et al., 2017). The subject was asked to rate how stressful he felt on the HealthLog app after a saliva sample was collected. The Brite 24 and Fitbit Sense were put on the subject when he was ready for sleep. The Brite 24 head cap was placed symmetrically on the subject's head, and the Fitbit Sense was worn on the non-dominant wrist. To reset the brain to a common baseline, the subject first went through a wake rest phase where he simply sat quietly for 2 min while staying awake. The wake rest phase was followed immediately by the sleep phase. The Brite 24 was left on until it ran out of battery. The subject was instructed to remove and stop the Brite 24 (simply by pressing the main button) when he needed to go to the restroom early morning or when he woke up, whichever happened first. The subject was asked to synchronize the Fitbit Sense with the companion mobile application after waking up.

3 DATA ANALYSIS

The objective of the data analysis was to identify the channel-wise features derived from the hemodynamic signals that are significantly associated to stress indicators. We first processed the raw OD signals to yield cleaned high quality ΔO_2Hb and ΔHHb signals, and then derived features from the cleaned signals at each channel. The data analysis pipeline was implemented using Python 3.8.8.

3.1 Data Preprocessing

We exported data from all the devices and instruments for preprocessing. Stress data and Fitbit data were aggregated at a 1-day resolution (i.e., one data point for each day during the experiment period). The perceived stress data were exported from the HealthLog app into a CSV file with a premium account subscription. Fitbit data of sleep, heart rate, and breath rate were exported using a web app that we developed in our previous study (Liang et al., 2016). All these data were then merged by matching date stamps.

The fNIRS data were collected at a high sampling rate of 50 Hz and required more complex preprocessing. The total raw signals measured by the Brite 24 consist of several components, and the ΔO_2Hb and ΔHHb related to neural activity is only a small portion. Noisy components are those related to breath, heartbeat, and movement, which need to be removed (Tak and Ye, 2014). The fNIRS data preprocessing pipeline is illustrated in **Figure 3**.

1) *Export raw OD signals.* Using the OxySoft, we exported raw OD signals in EDF format so that they were semi-compatible with the data formats supported by the MNE-NIRS Python library (Luke et al., 2021). Although the OxySoft also allows the export of ΔO_2Hb and ΔHHb signals that have been converted from raw OD signals, unfortunately the Artinis format of these signals is not supported by the MNE-NIRS library at the time of this study. It is worth noting that the MNE-NIRS library read in the EDF data as EEG signals by default; hence, additional processing was needed to convert the signal type to fNIRS after loading EDF files using the MNE-NIRS library.
2) *Trim the signals.* Since we were only interested in the first sleep cycle, we discarded the signal segments before the sleep start time and after the first sleep cycle. The sleep start time as recorded by the Fitbit Sense was used as the start time (T_s) of the effective data. The end time (T_e) of the effective data were set to $T_e = T_s + 90$ min as the average sleep cycle of healthy adults is 90 min (Feinberg and Floyd, 1979).
3) *Remove channels with poor signal quality.* The quality of the OD signals could be compromised by many factors during the measurement. While the OxySoft allows real-time inspection

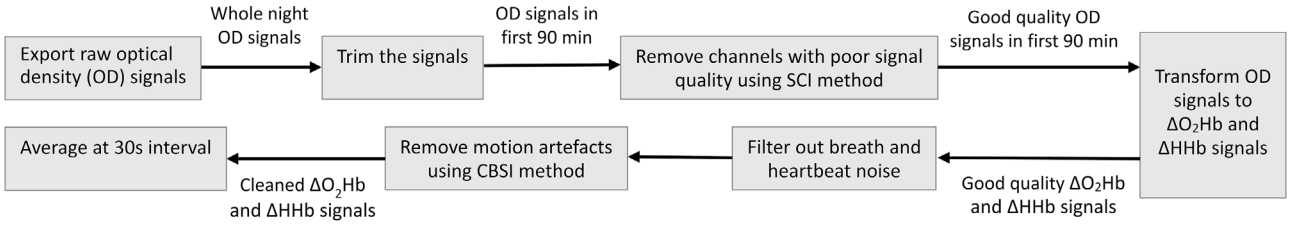


FIGURE 3 | Pipeline for fNIRS data preprocessing.

of signal quality, it does not provide computational tools to remove channels with poor signals. In our data preprocessing pipeline, we removed channels with poor signal quality using the Scalp Coupling Index (SCI) method (Pollonini et al., 2014). We first performed channel-wise filtering on the OD signals at both wavelengths using a band-pass filter (0.7–1.5 Hz) to preserve only the heartbeat components. The resulting signals were normalized to balance any difference between their amplitudes. The zero-lag cross-correlation between the resulting signals of the same channel—defined as the SCI—was computed and used as a quantitative measure of the signal-to-noise ratio of the channel. Channels with an SCI-value below 0.75 were regarded as poor channels and were removed from the subsequent analysis.

4) *Transform OD to* ΔO_2Hb *and* ΔHHb. The modified Beer–Lambert law (MBLL) was applied to convert the OD signals to ΔO_2Hb and ΔHHb signals (Delpy et al., 1988). As shown in **Eq. 1**, $\epsilon_{O_2HB}(\lambda_i)$ and $\epsilon_{HHB}(\lambda_i)$ are the extinction coefficients of O_2Hb and HHb at wavelength of λ_i, respectively. L denotes the interoptode distance. PPF(λ_i) denotes the partial pathlength factor, which represents the sensitivity of the measured optical density to the hemoglobin concentration change in a focal region (Steinbrink et al., 2001). In this study, L and PPF(λ_i) were set to 0.03 and 0.1 (Strangman et al., 2014).

$$\Delta OD(\lambda_i) = \left[\epsilon_{O_2HB}(\lambda_i)\Delta O_2HB + \epsilon_{HHB}(\lambda_i)\Delta HHB\right] \times L \times PPF(\lambda_i). \quad (1)$$

5) *Filter out physiological systemic responses.* The hemodynamic response due to neural activity has frequency content predominantly below 0.5 Hz (in many cases around 0.1 Hz). The ΔO_2Hb and ΔHHb signals were band-pass filtered to remove cardiac and respiratory noise. According to the Fitbit data, the subject typically had a breath and heart rate between 11–13 bmp and 55–80 bmp, respectively, during sleep. Hence, the cutoff frequency of the band-pass filter was set to 0.02–0.18 Hz.
6) *Remove motion artifacts.* Although fNIRS is considered more resilient to motion artifacts than EEG, abrupt head motion such as tossing and turning in sleep may still induce spikes that contaminate the true cortical hemodynamic signals. We used the correlation based signal improvement (CBSI) method (Cui et al., 2010) for motion artifacts removal. This method is based on the observation that the ΔO_2Hb and ΔHHb signals, which are typically strongly negatively correlated, will become more positively correlated when contaminated with motion artifacts. Correspondingly, the CBSI method removes motion artifact through recovering the negative correlation between the ΔO_2Hb and ΔHHb signals.
7) *Compute epoch-wise average.* The effective data of each measurement trial spanned over 90 min (generating 270,000 data points each night). The duration was significantly longer than that of traditional fNIRS studies where a measurement is usually at the scale of several minutes. To efficiently analyze such huge amount of data, we averaged the cleaned ΔO_2Hb and ΔHHb signals epoch-by-epoch at a 30-s interval. Each epoch contains 1,500 data points. This step was compliant with the standard procedure for sleep analysis (Iber et al., 2017). The output time series signals are denoted as $\{X_n: n = 1, \ldots, N\}$ where $N = 180$.

3.2 Feature Construction

We derived 36 features from the ΔO_2Hb signal and 36 features from the ΔHHb at each channel. These features fall into three groups. The first group contains 11 time-domain features. These features were directly extracted from the cleaned signals.

- Descriptive statistics: mean ($\bar{X}$), standard deviation (σ), maximum (X_{max}), and minimum (X_{min}).
- Skewness (*skew*): a normalized measure of the asymmetry of the probability distribution of a signal.
- Kurtosis (*kurt*): a normalized measure of the relative importance of tails versus shoulders in causing dispersion of a signal.
- The 5th-order moment (mmt_5): a measure of the relative importance of tails versus center in causing skew of a signal.
- Mean absolute value (*MAV*): the average of the absolute value of the signal amplitude.
- Root mean square (*RMS*): a measure of the average power of a signal.
- Zero crossing (*ZC*): the number of times the signal changes value from positive to negative and vice versa. It can be interpreted as a measure of the noisiness of a signal.
- Crest factor (*CF*): an indicator of how extreme the peaks are in a signal.

The second group contains two most typical frequency-domain features. Fast Fourier transform (FFT) was applied to convert $\{X_n: n = 1, 2, 3, \ldots, N\}$ from time domain to frequency domain to extract the following two features.

- Total power (*totalSpec*): the sum of the spectral components of a signal.
- Maximal power (*maxSpec*): the maximum amplitude of the spectral components of a signal.

The third group contains 23 features that characterize the nonlinear characteristics of the cortical hemodynamic signals. Human physiological systems are dynamical systems that often exhibit nonlinear characteristics (Goldberger and West, 1992; Cheffer et al., 2021). Previous studies found that nonlinearities are particularly present in the brain (Toyoda et al., 2008; Ma et al., 2018). To extract nonlinear features, we first used Takens' time-delay embedding to construct a phase space representation of the system as:

$$\vec{u}(i) = (x(i), x(i+\tau), \ldots, x(i+\tau(d-1))), \quad (2)$$

where τ is the time delay and d the embedding dimension. The optimal value of τ and d were decided by minimizing the time-delayed mutual information and by the false nearest neighbors method (Kantz and Schreiber, 2003), respectively. The search range was set to [1, 10] for τ and [2, 6] for d at an increment of 1. The signals were then embedded using the optimal τ_{opt} and d_{opt}. The maximal Lyapunov exponent (*MLE*), Hurst exponent (*HE*), and correlation dimension (*CD*) were computed from the embedded ΔO_2Hb and ΔHHb signals. Several nonlinear analysis techniques were also applied to derive features, including recurrence quantitative analysis (RQA), Poincaré plots (PP), and detrended fluctuation analysis (DFA). Different measures of entropy were also calculated.

The RQA computes several quantitative metrics from a recurrence plot (RP). A RP is a visualization of the recurrence behavior of the phase space trajectory $\vec{u}(i)$ of a dynamical system. Each element in the RP is calculated by the following equation:

$$R(i,j) = \Theta(\epsilon - \| \vec{u}(i) - \vec{u}(j) \|), \quad (3)$$

where $\Theta: \mathbb{R} \rightarrow (0, 1)$ is the Heaviside step function, ϵ is a cutoff distance, and $\|\blacksquare\|$ is the Euclidean norm. In this study, ϵ was set to 0.85. The metrics derived from a RP quantify the recurrence behavior of a dynamic system.

Poincaré plot (PP) is a special type of RP used to quantify self-similarity of a dynamical system. It is a scatter plot of each pair of consecutive data points in a time series signal (technogram), which is often in a shape of ellipse. The minor axis (or width) of the ellipse, denoted as SD_1, reflects the level of short-term instantaneous variability. The major axis (or length) of the ellipse, denoted as SD_2, reflects the long-term variability. PP has been widely used in ECG analysis to help diagnose cardio abnormalities (Hoshi et al., 2013).

The DFA method is often used to quantify the fractal scaling properties and is useful for revealing the statistical self-similarity of a signal (Peng et al., 1994). It has been proven particularly useful in neurology studies (Peng et al., 1994; Hardstone et al., 2012). The DFA first converts a signal to mean-centered cumulative sum. The output signal is then split into epochs, detrended, and the *RMS* is computed. This process is repeated over a range of epoch sizes n at different scale. A linear trend line is then fit to the log (*RMS*) – log(n) plot. The slope of the fitted trend line, denoted by α, is called scaling exponent.

The derived nonlinear features are summarized below.

- Optimal delay (τ_{opt}).
- Optimal embedding dimension (d_{opt}).
- Maximal Lyapunov exponent (*MLE*): a measure of separation rate of a signal's trajectories in the phase space. It indicates the predictability of a dynamic system. A positive maximum Lyapunov exponent is an indicator of the presence of chaos (Eckmann and Ruelle, 1985).
- Hurst exponent (*HE*): a measure of long-term memory (or long-range dependency) of a signal. A value of *HE* in the range 0–0.5 indicates long-term negative autocorrelation, while a value in the range 0.5–1.0 indicates long-term positive autocorrelation. A value of 0.5 can indicate a completely uncorrelated signal.
- Correlation dimension (*CD*): an indicator used to distinguish deterministic chaos from stochastic processes.
- Recurrence rate (*RR*): the number of black dots in a RP excluding the main diagonal line. It is a measure of the relative density of recurrence points in the entire RP.
- Percent determinism (*DET*): the fraction of recurrence points that form diagonal lines. It reports the percentage of recurrent points in diagonal structures.
- Maximal diagonal line length (D_{max}): the length of the single longest line in the diagonal direction within an entire RP. The smaller the D_{max}, the more divergent the trajectories.
- Average diagonal line length (D_{avg}): the average time that two segments of the phase space trajectory are close to each other.
- Entropy of diagonal lines lengths (ENT_{D}): the Shannon entropy of the frequency distribution of the diagonal line lengths. It reflects the complexity of the deterministic structure in the system.
- Laminarity (*LAM*): the histogram of lengths of vertical lines in a RP. It reports the percentage of recurrent points in vertical structures.
- Trapping time (*TT*): the average length of the vertical lines. It indicates the mean time the system will abide at a specific state.
- Longest vertical line length (V_{max}): the maximal length of the vertical lines in the entire RP.
- Entropy of vertical lines lengths (ENT_{V}): the Shannon entropy of the frequency distribution of the vertical line lengths.
- Standard deviation of the minor axis of a PP (SD_1).
- Standard deviation of the major axis of a PP (SD_2).
- Ratio of SD_1 and SD_2 (SD_{ratio}); computed as SD_1/SD_2.
- Area of the fitted ellipse (S_e): the area of the ellipse fitted into the PP. It is computed as $S_e = \pi \times SD_1 \times SD_2$.

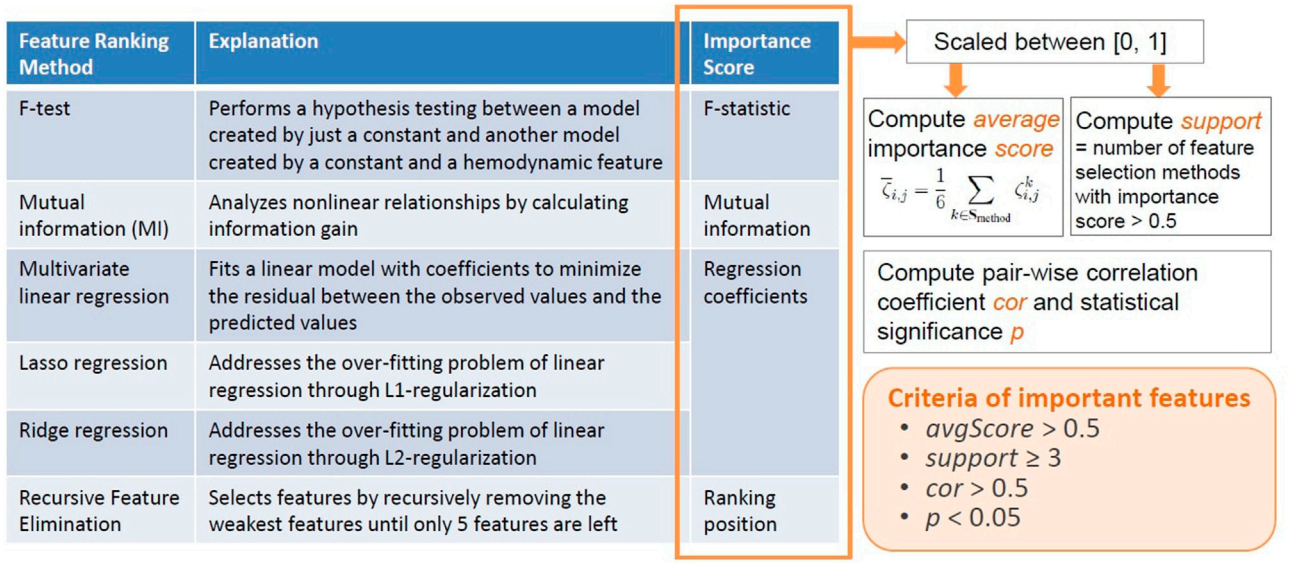


FIGURE 4 | Proposed ensemble method for selecting important hemodynamic features.

- Scaling exponent (α): the slope of the fitted trend line in DFA, where each epoch has no overlap.
- Scaling exponent with overlap (α_{OL}): the slope of the fitted trend line in DFA, where each epoch has 50% overlap.
- Sample entropy (*sampEn*): a measure of the negative natural logarithm of the probability that if two sets of data points of length m have Euclidean distance $D\,[X_m\,(n_1), X_m\,(n_2)] < r$ $(n_1 \neq n_2)$ then two sets of data points of $m + 1$ also have Euclidean distance $D\,[X_{m+1}\,(n_1), X_{m+1}\,(n_2)] < r$. In this study, r was set to 0.2σ. A lower value for the sample entropy corresponds to a higher probability indicating more self-similarity and less noise in the signal (Richman and Moorman, 2000).
- Permutation entropy (*perEn*): a complexity measure that captures the order relations between the values of a signal. Signals with smaller *perEn* are more regular and deterministic, and those with higher *perEn* are noisier and more random.
- Statistical complexity (*SC*): the product of the normalized permutation and a normalized version of the Jensen–Shannon divergence between the ordinal distribution and the uniform distribution (López-Ruiz et al., 1995).

3.3 Feature Ranking

We proposed an original ensemble approach to rank channel-wise hemodynamic features for each stress indicator. As outlined in **Figure 4**, this approach utilized six feature selection statistical and machine learning techniques to generate an average importance score for each feature $\bar{\zeta}$, calculated the correlation coefficient *cor* and the corresponding *p*-value between each feature and a target stress indicator, and performed feature pruning based on the specified criteria. Feature ranking was performed on ΔO_2Hb and ΔHHb features separately. The six feature selection techniques included F-test, mutual information, multivariate linear regression, least absolute shrinkage and selection operator (Lasso) regression, Ridge regression, and recursive feature elimination (RFE).

The F-test and mutual information (MI) are univariate methods that consider the relationship between each feature and a target stress indicator individually. The F-test method performs a hypothesis testing between a model created by just a constant and another model created by a constant and a hemodynamic feature, and hence reveals the significance of each feature in improving the model. The calculated F-statistic was used as the importance scores of features. While the F-statistic only reflects the linear relationship between a feature and a target stress indicator, the MI method analyzes nonlinear relationships by calculating information gain (Estevez et al., 2009; Ross, 2014). The MI between a hemodynamic feature and a target stress indicator reveals the reduction in uncertainty for the stress indicator given the known value of the feature (Ross, 2014), which was used as the importance score of features. The multivariate linear regression (LR) method fits a linear model with coefficients to minimize the residual between the observed values and the predicted values of the stress indicator. The Lasso regression and Ridge regression methods address the over-fitting problem of linear regression through L1-regularization and L2-regularization, respectively. In addition, the Lasso regression is considered a very useful technique in selecting a strong subset of features as it aggressively produces coefficients of 0 for some features. On the other hand, the Ridge regression is suited for data interpretation because useful features tend to have non-zero coefficients. The α parameters were all set to 0.5 for Lasso and Ridge regression. For the three linear regression methods, the estimated coefficients were used as the importance scores of the features. The RFE method selects features by recursively

TABLE 2 | Descriptive statistics of stress indicators.

	Mean	SD	Range	cor_{ps} [1]	p_{ps} [2]	cor_{sIgA} [3]	p_{sIgA} [4]
Cortisol (nmol)	3.4	3.3	1.5–13.2	−0.05	0.855	−0.38	0.144
sIgA (μg/ml)	295.3	163.0	89.7–674.1	0.17	0.528	—	—
Perceived stress	3.8	1.6	2.0–8.0	—	—	—	—

[1]*Correlation coefficient to perceived stress.*
[2]p-*value of the correlation coefficient to perceived stress.*
[3]*Correlation coefficient to sIgA.*
[4]p-*value of the correlation coefficient to sIgA.*

removing the weakest features until only five features are left. The ranking position of RFE was used as the importance score.

The cortical hemodynamic features were all scaled between [0, 1] before each feature selection technique was performed. For feature i ($i \in \mathbf{S}_{\text{feature}}$), the importance score generated by method k, denoted by ζ_i^k ($k \in \mathbf{S}_{\text{method}}$), were also scaled between [0, 1]. For a stress indicator j (j is either cortisol, sIgA, or perceived stress), the scores of all feature ranking methods for feature i were then averaged to produce an average score (denoted as $\bar{\zeta}_{i,j}$) of feature i. The computation of the $\bar{\zeta}_{i,j}$ is explained in **Eq. 4**. We also defined the *support* of feature i with respect to stress indicator j (denoted as $support_{i,j}$) as the number of feature selection methods that yielded an importance score above 0.50 for feature i. Since the $\bar{\zeta}$ only indicates the relative ranking of a feature, we calculated the linear correlation between individual feature and a target stress indicator (denoted as $cor_{i,j}$) to better interpret the quantitative relationships. The Pearson's correlation analysis was performed when cortisol and sIgA were used as the stress indicator, and Spearman's correlation analysis was performed when perceived stress was used as the stress indicator. The p-values (denoted as $p_{i,j}$) were calculated to indicate the significance of the correlation coefficients at a significance level of 0.05. The features that satisfied the following four criteria were selected as important features: 1) $\bar{\zeta}_{i,j} > 0.50$, 2) $support_{i,j} \geq 3$, and 3) $cor_{i,j} > 0.50$, and 4) $p_{i,j} < 0.05$. Feature ranking was performed in a channel-wise manner and for ΔO_2Hb and ΔHHb signals separately.

$$\bar{\zeta}_{i,j} = \frac{1}{6} \sum_{k \in \mathbf{S}_{\text{method}}} \zeta_{i,j}^k. \tag{4}$$

4 RESULTS

4.1 Descriptive Statistics

In total 15 days of data were collected from the subject. As shown in **Table 2**, the average level of salivary cortisol and sIgA were 3.4 nmol and 295.3 μg/ml, respectively. These values were within the normal ranges for healthy adults (Zeier et al., 1996; Oster et al., 2017). Perceived stress ranges from 2.0 to 8.0 with an average score of 3.8. Only 4 out of the 15 days were rated above 5, indicating that the subject did not perceive constant chronic stress during the data collection experiment. A correlation analysis found no significant linear relationship among the three stress indicators.

4.2 PFC Hemodynamic Features Associated to Stress

Channel-wise hemodynamic features associated to stress indicators are summarized in **Table 3–8**. The aggregated frequency of each feature type is illustrated in **Figure 5**. Visualization of the channels associated to each stress indicator is provided in **Figures 6**, **7**.

Stress was associated to features in both time and frequency domains as well as to nonlinear features. For cortisol, the top three associated cortical hemodynamic features of both the ΔO_2Hb and ΔHHb signals are the $\bar{X}$ of channel 16, the *sampEn* of channel 26, and the SD_{ratio} of channel 12. These three features were supported by all six feature selection techniques of the ensemble feature ranking algorithm. Higher cortisol level was strongly associated to increased mean of ΔO_2Hb but decreased mean of ΔHHb at channel 26, strongly associated to increased sample entropy of both the ΔO_2Hb and ΔHHb signals at channel 16, and moderately associated to decreased SD_{ratio} of both the ΔO_2Hb and ΔHHb signals at channel 12. With all channels counted in, the most frequently referred feature type was the time-domain feature $\bar{X}$. Five feature types were found to be associated only to cortisol but not to the other two stress indicators: mmt_5, *sampEn*, *PE*, *DET*, and *maxSpec*.

Less channel-wise features were found to associate to sIgA. Only two ΔO_2Hb features (i.e., *HE* of channel 21 and SD_{ratio} of channel 26) and one ΔHHb feature (i.e., SD_{ratio} of channel 26) were supported by all six feature selection technique. A lower sIgA level was moderately associated to higher values of the Hurst exponent of the ΔO_2Hb signal at channel 21 as well as increased SD_{ratio} of both the ΔO_2Hb and ΔHHb signals at channel 26. The most frequently referred feature type for sIgA were *HE*, SD_{ratio}, τ_{opt}, and *ZC*. In addition, *ZC* was associated only to sIgA but not to the other two stress indicators.

Two time-domain features associated to perceived stress were supported by all six feature selection methods. These two features were also the only features that had non-zero coefficient when the Lasso method was applied. Higher perceived stress was strongly associated to reduced skewness of the ΔO_2Hb signal and increased crest factor of the ΔHHb signal at channel 16. The most frequently referred feature types were *HE*, *MLE*, and α_{OL}. In the meantime, α, α_{OL}, and X_{min} were the feature types specific to perceived stress.

Figures 6, **7** demonstrated that channel 3 (optode pair Tx3-Rx1), 16 (Tx6-Rx5), 20 (Tx9-Rx6), 21 (Tx5-Rx7), and 26 (Tx9-Rx8) were the most relevant channels. The features of both the

TABLE 3 | Channel-wise ΔO_2Hb features associated to salivary cortisol.

ChID[1]	Feature	F-test	MI	LR	Lasso	Ridge	RFE	$\bar{\zeta}$	*support*	*cor*	*p*
16	*mean*	1.00	1.00	1.00	1.00	1.00	0.91	0.99	6	0.99	0.001
26	*sampEn*	1.00	0.85	1.00	1.00	1.00	1.00	0.97	6	0.71	0.032
12	SD_{ratio}	1.00	0.63	0.78	1.00	1.00	1.00	0.90	6	−0.59	0.045
8	*PE*	1.00	0.54	0.46	1.00	0.93	0.97	0.82	5	−0.55	0.027
14	mmt_5	1.00	0.32	0.81	1.00	1.00	0.57	0.78	5	0.76	0.001
20	*SC*	1.00	0.21	0.52	1.00	1.00	0.89	0.77	5	0.52	0.038
3	*CF*	0.97	0.99	0.38	0.00	1.00	1.00	0.72	4	−0.56	0.045
10	*mean*	1.00	0.03	1.00	0.00	1.00	1.00	0.67	4	−0.81	0.004
5	*DET*	0.78	0.77	0.25	1.00	0.16	0.94	0.65	4	−0.69	0.003
21	*skew*	1.00	0.00	0.34	1.00	0.55	1.00	0.65	4	0.54	0.045
8	d_{opt}	0.86	1.00	0.35	0.00	0.67	0.94	0.64	4	−0.52	0.038
19	mmt_5	1.00	0.00	0.81	0.00	0.85	1.00	0.61	4	−0.53	0.034
11	*HE*	1.00	0.01	0.58	0.00	0.95	1.00	0.59	4	−0.67	0.016
11	SD_{ratio}	0.74	0.76	0.16	0.00	0.99	0.54	0.53	4	−0.62	0.033

[1]*Channel ID.*

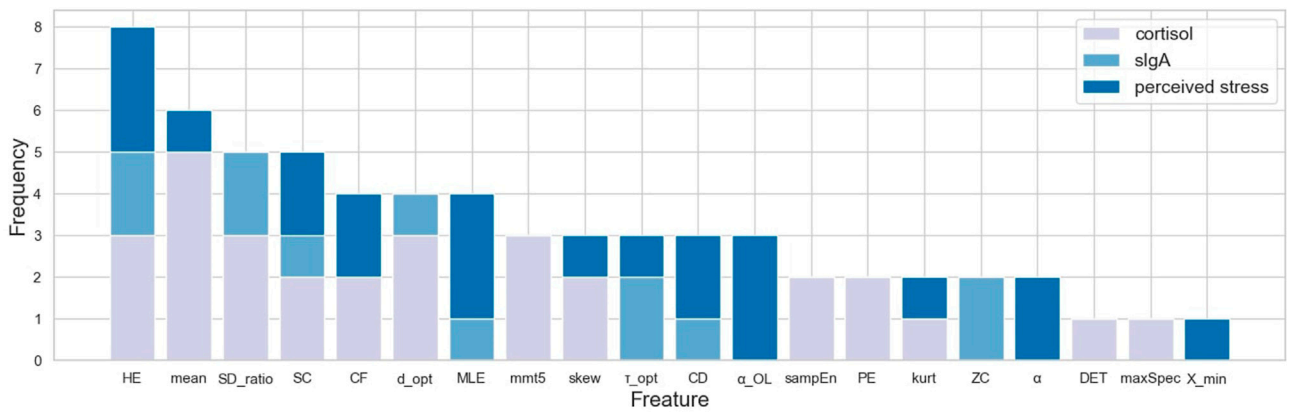


FIGURE 5 | Frequency of associated feature types summed across all channels.

ΔO_2Hb and the ΔHHb signals at channel 16 were strongly associated to all three stress indicators. Channel 20 had similar relevance but with only moderate associations. The features of the ΔO_2Hb signals at channels 3 and 21 were moderately associated to all three stress indicators. In addition, the features of both the ΔO_2Hb and the ΔHHb signals at channel 26 were moderately associated to both cortisol and sIgA.

5 DISCUSSION

This pilot study demonstrated the feasibility of investigating stressed brain during sleep by configuring a digital ecosystem with wearable/portable devices and mobile applications. The sleep data collected with Fitbit Sense provided information on sleep start time, and the heart and breath rate data facilitated personalized filtering of the fNIRS signals. The use of mobile applications such as the HealthLog app can also help reduce the burden of manual log, and thus is likely to improve subjects' adherence to the study protocol. We do not intend to draw conclusions due to the pilot nature of the study; instead, we discuss several observations that may inspire future studies in the same direction.

Stress can be measured along multiple dimensions using several indicators. In this study, we measured salivary cortisol, sIgA, and collected subjective ratings on perceived stress. The correlation analysis revealed that while some features derived from the cortical hemodynamic signals associated with all stress indicators, others may be specific to only one or two stress indicators. The analysis showed that the brain activity may be characterized using various features derived from the hemodynamic signals. Time-domain features, frequency-domain features, and nonlinear features all showed promise. Taken together, the top five most frequently referred feature types were Hurst exponent, mean, the ratio of the major/minor axis standard deviation of a Poincaré plot, statistical complexity, and crest factor. Breaking down into individual stress indicator, the mean of the cortical hemodynamic signals is the most frequently referred feature type for cortisol. This coincides with the fact that most studies that rely on cortisol as the

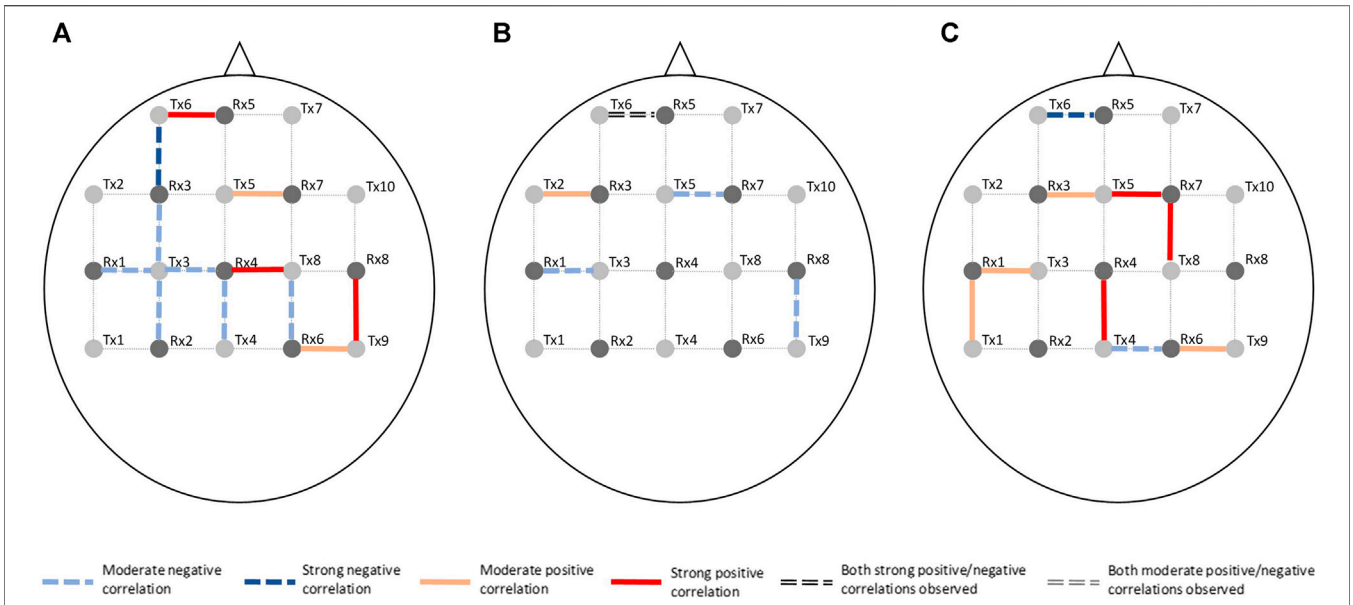


FIGURE 6 | Visualization of channels associated to ΔO_2Hb on the optode template. Stress indicator is **(A)** cortisol, **(B)** sIgA, and **(C)** perceived stress, respectively.

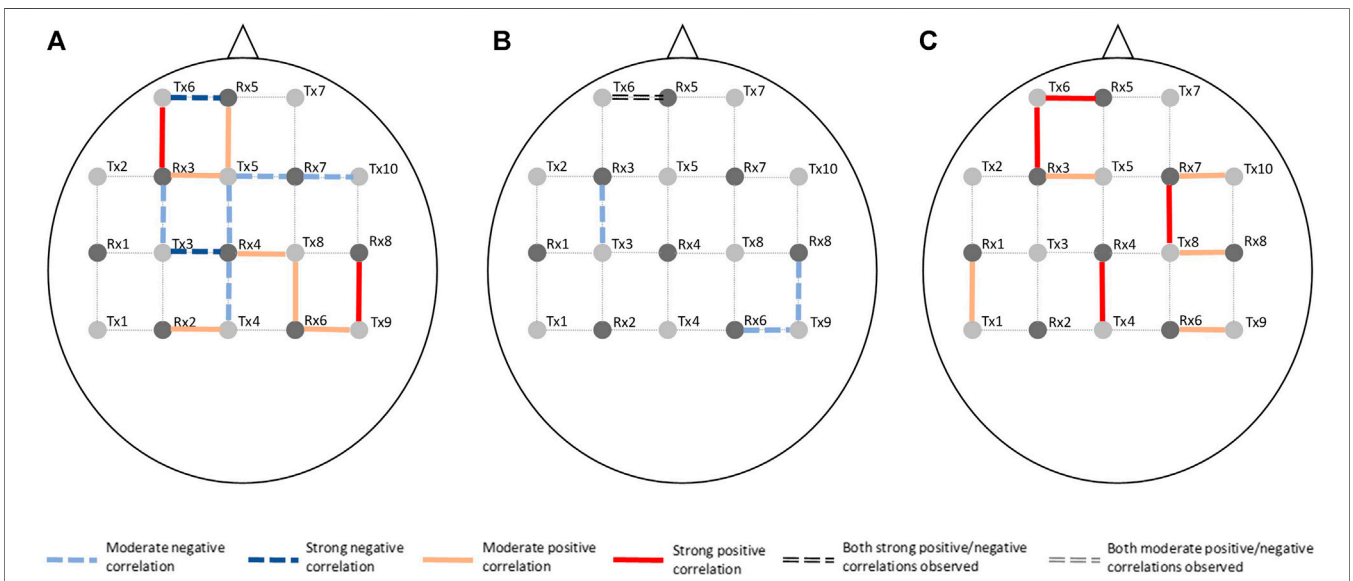


FIGURE 7 | Visualization of channels associated to ΔHHb on the optode template. Stress indicator is **(A)** cortisol, **(B)** sIgA, and **(C)** perceived stress, respectively.

stress indicator solely characterize the cortical hemodynamic patterns using the mean of the ΔO_2Hb and the ΔHHb signals. On the other hand, nonlinear features of the hemodynamic signals could be more useful when sIgA and perceived stress are used as stress indicators. Four feature types had the same highest frequency for sIgA: Hurst exponent, major/minor axis standard deviation ratio of the Poincaré plot, optimal delay, and zero crossing. For perceived stress, Hurst exponent, maximal Lyapunov exponent, and over lapped α in DFA were the most frequently referred feature types. It is also found that the time-domain features (e.g., mean) derived from the ΔO_2Hb signals and those from the ΔHHb signals demonstrated opposite correlation directions to stress indicators, whereas the nonlinear features (e.g., Hurst exponent, correlation dimension, and statistical complexity) derived from the two cortical hemodynamic signals demonstrate the same correlation direction. While the two hemodynamic signals share some common important features, each one also contributed unique features. This suggests the necessity of consider both signals when investigating braining activity using fNIRS, which provides support

TABLE 4 | Channel-wise ΔHHb features associated to salivary cortisol.

ChID	Feature	F-test	MI	LR	Lasso	Ridge	RFE	$\bar{\zeta}$	*support*	*cor*	*p*
26	*sampEn*	1.00	1.00	1.00	1.00	1.00	1.00	1.00	6	0.71	0.032
16	*mean*	1.00	0.80	1.00	1.00	1.00	1.00	0.97	6	–0.99	0.002
12	SD_{ratio}	1.00	0.69	0.60	1.00	1.00	0.83	0.85	6	–0.59	0.045
15	*HE*	1.00	0.91	0.46	1.00	0.88	0.37	0.77	4	0.54	0.030
8	*PE*	1.00	0.34	0.58	1.00	0.86	0.66	0.74	5	–0.55	0.027
10	*mean*	1.00	0.40	1.00	0.00	1.00	1.00	0.73	4	0.71	0.020
9	d_{opt}	1.00	0.20	0.58	1.00	0.47	1.00	0.71	4	0.57	0.022
13	*mean*	1.00	1.00	0.91	0.00	1.00	0.34	0.71	4	–0.63	0.009
11	*HE*	1.00	0.36	0.89	0.00	1.00	0.91	0.69	4	–0.76	0.004
20	*SC*	1.00	0.11	0.30	1.00	1.00	0.69	0.68	4	0.52	0.038
24	*kurt*	0.77	0.81	0.11	1.00	0.45	0.77	0.65	4	–0.54	0.031
21	*skew*	1.00	0.31	0.63	0.00	0.64	1.00	0.60	4	–0.59	0.026
6	mmt_5	1.00	0.65	0.37	0.00	0.83	0.69	0.59	4	0.68	0.004
19	*CF*	1.00	0.50	0.38	0.00	0.50	1.00	0.57	4	0.55	0.027
8	d_{opt}	0.86	0.74	0.25	0.00	0.62	0.86	0.56	4	–0.52	0.038
14	*maxSpec*	0.64	0.14	0.14	1.00	0.54	0.54	0.50	4	0.69	0.003

TABLE 5 | Channel-wise ΔO_2Hb features associated to salivary sIgA.

ChID	Feature	F-test	MI	LR	Lasso	Ridge	RFE	$\bar{\zeta}$	*support*	*cor*	*p*
21	*HE*	1.00	0.81	1.00	1.00	1.00	1.00	0.97	6	–0.58	0.029
26	SD_{ratio}	1.00	0.62	1.00	1.00	1.00	1.00	0.94	6	–0.69	0.041
7	*MLE*	1.00	0.00	0.95	1.00	1.00	1.00	0.82	5	0.55	0.029
16	*ZC*	0.17	0.75	1.00	1.00	1.00	1.00	0.82	5	–0.97	0.005
16	τ_{opt}	1.00	0.75	0.86	0.35	0.82	1.00	0.80	5	1.00	0.000
3	*HE*	1.00	1.00	0.22	0.03	1.00	0.83	0.68	4	–0.61	0.027

TABLE 6 | Channel-wise ΔHHb features associated to salivary sIgA.

ChID	Feature	F-test	MI	LR	Lasso	Ridge	RFE	$\bar{\zeta}$	*support*	*cor*	*p*
26	SD_{ratio}	1.00	0.55	1.00	1.00	1.00	1.00	0.93	6	–0.69	0.041
16	τ_{opt}	1.00	0.75	0.91	0.31	0.86	1.00	0.81	5	1.00	0.000
8	*CD*	1.00	0.85	0.66	0.47	1.00	0.74	0.79	5	–0.50	0.047
16	*ZC*	0.17	0.75	0.81	1.00	1.00	1.00	0.79	5	–0.97	0.005
20	d_{opt}	0.79	1.00	0.41	0.42	0.60	0.91	0.69	4	–0.59	0.017
20	*SC*	1.00	0.63	0.32	0.49	1.00	0.69	0.69	4	–0.63	0.009

TABLE 7 | Channel-wise ΔO_2Hb features associated to perceived stress.

ChID	Feature	F-test	MI	LR	Lasso	Ridge	RFE	$\bar{\zeta}$	*support*	*cor*	*p*
16	*skew*	1.00	0.58	1.00	1.00	1.00	1.00	0.93	6	–0.89	0.041
23	*MLE*	1.00	1.00	1.00	0.00	1.00	1.00	0.83	5	0.86	0.003
12	α	0.94	0.98	0.64	0.00	0.86	1.00	0.74	5	0.69	0.013
12	α_{OL}	1.00	1.00	0.61	0.00	1.00	0.80	0.73	5	0.74	0.006
1	*CD*	1.00	1.00	0.39	0.00	1.00	0.97	0.73	4	0.57	0.032
18	*mean*	1.00	0.20	1.00	0.00	1.00	1.00	0.70	4	–0.62	0.010
20	*HE*	1.00	0.92	0.48	0.00	1.00	0.74	0.69	4	0.58	0.018
3	*CF*	1.00	1.00	0.29	0.00	1.00	0.80	0.68	4	0.62	0.025
21	τ_{opt}	1.00	1.00	0.13	0.00	0.82	0.71	0.61	4	0.73	0.003
9	*SC*	1.00	0.00	0.78	0.00	0.99	0.69	0.58	4	0.51	0.042

to the argument made in previous studies (Tachtsidis and Scholkmann, 2016). It is also worth mentioning that the time-domain and frequency-domain features have the merit of their interpretability, whereas some nonlinear features such as the Hurst exponent and the optimal delay may hinder straightforward interpretation.

TABLE 8 | Channel-wise ΔHHb features associated to perceived stress.

ChID	Feature	F-test	MI	LR	Lasso	Ridge	RFE	$\bar{\zeta}$	*support*	*cor*	*p*
16	*CF*	1.00	0.80	0.98	1.00	1.00	1.00	0.96	6	0.89	0.041
12	α	1.00	0.91	1.00	0.00	1.00	1.00	0.82	5	0.69	0.013
23	*MLE*	1.00	1.00	1.00	0.00	1.00	0.94	0.82	5	0.79	0.011
1	*CD*	1.00	1.00	0.81	0.00	1.00	1.00	0.80	5	0.62	0.019
16	*kurt*	0.52	1.00	1.00	0.00	0.84	1.00	0.73	5	0.89	0.041
12	α_{OL}	1.00	0.94	0.52	0.00	0.96	0.86	0.71	5	0.72	0.008
9	*MLE*	0.92	1.00	0.36	0.00	0.92	0.94	0.69	4	0.50	0.047
10	X_{min}	1.00	1.00	0.55	0.00	0.58	0.89	0.67	5	0.70	0.024
24	α_{OL}	1.00	0.00	1.00	0.00	1.00	1.00	0.67	4	0.52	0.039
25	*HE*	1.00	0.72	0.47	0.00	1.00	0.69	0.65	4	0.58	0.018
20	*HE*	1.00	0.59	0.53	0.00	1.00	0.66	0.63	5	0.54	0.031
9	*SC*	1.00	0.00	0.76	0.00	1.00	1.00	0.63	4	0.51	0.042

TABLE 9 | Comparison of Experiment Protocols and Main Findings of Previous Daytime fNIRS Stress Studies and the Current In-sleep fNIRS Stress Study.

Study	Stress induction	Stress indicators	Main findings
Yang et al. (2007)	Negative pictures	None	Increased activity in the PFC among females of the experiment group
Yanagisawa et al. (2011)	Cyberball task	Subjective rating on social pain	Decreased activity in the VLPFC
Rosenbaum et al. (2018)	Trier Social Stress Test (TSST)	Cortisol[1], heart rate, subjective stress rating	Positive association between the activity in the right DLPFC and cortisol response; positive association between the activity in the bilateral DLPFC and subjective stress rating
Schaal et al. (2019)	Maastricht Acute Stress Test (MAST)	Cortisol, heart rate, subjective stress rating	Increased activity in the left DLPFC and the bilateral orbitofrontal cortex (OFC) during the mental arithmetic task; decreased activity in the left DLPFC during the hand immersion task
This study	Naturalistic daily life stressors	Cortisol, sIgA, subjective stress rating	Positive association between the activity in the right caudal-DLPFC, the left RLPFC and cortisol response; positive association between the activity in the caudal-DLPFC and subjective stress rating

[1]*Refers to salivary cortisol unless otherwise specified.*

This study also generated preliminary observations on the sub-regions in the PFC associated to stress. The left rostral prefrontal cortex (channel 16; optode pair Tx6-Rx5), or RLPFC for short, is undoubtedly the most relevant sub-region. Significantly strong correlations were found between the hemodynamic features derived at this sub-region and all three stress indicators. The specific features selected at channel 16 varied depending on the target stress indicator. Higher cortisol level (indicating stronger stress response) was associated to increased mean ΔO_2Hb and decreased mean ΔHHb. Lower sIgA level (indicating stronger stress response) was associated to higher levels of noisiness in the ΔO_2Hb and ΔHHb signals characterized by increased zero crossing. Higher perceived stress level was associated to increased symmetry in the ΔO_2Hb signal and higher peak in the ΔHHb signal. The dorsolateral prefrontal cortex (DLPFC) was also a relevant cortical area. The area of mid-DLPFC (channel 3; optode pair Tx3-Rx1) and caudal-DLPFC (channel 20; optode pair Tx9-Rx6) both demonstrated significant and moderate associations to all three stress indicators. In the case of the caudal-DLPFC (channel 20), consistent relationships were found between stress and a same feature: the statistical complexity of the ΔHHb signal. In contrast, a ΔO_2Hb feature at the mid-DLPFC (channel 3) correlates differently to different stress indicators. To be more specific, lower crest factor of the ΔO_2Hb signal at channel 3 was associated to higher cortisol level (indicating stronger hormonal stress response) but at the same time higher sIgA level (indicating weaker immunological stress response). Rather than viewing this as contradictory, an alternative interpretation could be that different stress indicators characterize different facets of the human stress response, suggesting the necessity of using multiple indicators in stress studies. The relevance of the DLPFC in stress response of healthy subjects during wake time has been documented in previous stress studies using near-infrared spectroscopy (NIRS) technique (Yang et al., 2007; Yanagisawa et al., 2011; Rosenbaum et al., 2018; Schaal et al., 2019). **Table 9** summarizes the experiment protocols and the main findings of these studies, which were all conducted when subjects were awake and were engaged in cognitive tasks. Our findings provide supplementary support to the within-person role of the DLPFC in processing stress during sleep. In addition, the preliminary result shed light on the possible role of the RLPFC, especially the left RLPFC, in processing stress during sleep. On the other hand, no relevance was found in the ventrolateral prefrontal cortex (VLPFC). While (Yanagisawa et al., 2011) found negative association between the activity in the VLPFC and the subjective ratings on social pain (which was induced by a feeling of social isolation), our finding suggests that this region may not be involved in processing daily life stress during sleep.

The preliminary findings should be interpreted with caution due to the strong limitations of the present study. First, while the

idiographic N-of-1 approach adopted in this study *represents a true scientific undertaking* (Barlow and Nock, 2009), the preliminary findings solely hold for this specific subject. Future studies may conduct the longitudinal measurement on more subjects to identify possible common patterns across subjects. Second, the data analysis protocol in this study did not count in the interplay among different channels nor the confounding effect of different sleep stages. Analyzing the orchestration of the cortical hemodynamic signals from a dynamic network perspective may leads to new insights into how the brain responds to stress during sleep. Furthermore, this study only focused on the activity in the PFC area in the first sleep cycle. The potential role of other cortical areas in stress response during a full course of sleep demands further studies.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of Kyoto University of Advanced Science. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

ZL conceived and designed the study; ZL performed the data collection experiments, retrieved the data, and performed data preprocessing and analysis; ZL drafted the manuscript and made revision.

FUNDING

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (Grant Number 16H07469, 19K20141, and 21K17670).

REFERENCES

Alonso, J. F., Romero, S., Ballester, M. R., Antonijoan, R. M., and Mañanas, M. A. (2015). Stress Assessment Based on Eeg Univariate Features and Functional Connectivity Measures. *Physiol. Meas.* 36, 1351–1365. doi:10.1088/0967-3334/36/7/1351

Arnsten, A. F. T., Raskind, M. A., Taylor, F. B., and Connor, D. F. (2015). The Effects of Stress Exposure on Prefrontal Cortex: Translating Basic Research into Successful Treatments for Post-Traumatic Stress Disorder. *Neurobiol. Stress* 1, 89–99. doi:10.1016/j.ynstr.2014.10.002

Barlow, D. H., and Nock, M. K. (2009). Why Can't We Be More Idiographic in Our Research? *Perspect. Psychol. Sci.* 4, 19–21. doi:10.1111/j.1745-6924.2009.01088.x

Burg, M. M., Schwartz, J. E., Kronish, I. M., Diaz, K. M., Alcantara, C., Duer-Hefele, J., et al. (2017). Does Stress Result in You Exercising Less? or Does Exercising Result in You Being Less Stressed? or Is it Both? Testing the Bi-Directional Stress-Exercise Association at the Group and Person (N of 1) Level. *Ann. Behav. Med.* 51, 799–809. doi:10.1007/s12160-017-9902-4

Buysse, D. J., Germain, A., Hall, M., Monk, T. H., and Nofzinger, E. A. (2011). A Neurobiological Model of Insomnia. *Drug Discov. Today Dis. Models* 8, 129–137. doi:10.1016/j.ddmod.2011.07.002

Cerqueira, J. J., Mailliet, F., Almeida, O. F. X., Jay, T. M., and Sousa, N. (2007). The Prefrontal Cortex as a Key Target of the Maladaptive Response to Stress. *J. Neurosci.* 27, 2781–2787. doi:10.1523/jneurosci.4372-06.2007

Chang, J., and Yu, R. (2018). Alternations in Functional Connectivity of Amygdalar Subregions under Acute Social Stress. *Neurobiol. Stress* 9, 264–270. doi:10.1016/j.ynstr.2018.06.001

Cheffer, A., Savi, M. A., Pereira, T. L., and de Paula, A. S. (2021). Heart Rhythm Analysis Using a Nonlinear Dynamics Perspective. *Appl. Math. Model.* 96, 152–176. doi:10.1016/j.apm.2021.03.014

Cui, X., Bray, S., and Reiss, A. L. (2010). Functional Near Infrared Spectroscopy (Nirs) Signal Improvement Based on Negative Correlation between Oxygenated and Deoxygenated Hemoglobin Dynamics. *NeuroImage* 49, 3039–3046. doi:10.1016/j.neuroimage.2009.11.050

Dedovic, K., D'Aguiar, C., and Pruessner, J. C. (2009). What Stress Does to Your Brain: a Review of Neuroimaging Studies. *Can. J. Psychiatry* 54, 6–15. doi:10.1177/070674370905400104

Delpy, D. T., Cope, M., Zee, P. v. d., Arridge, S., Wray, S., and Wyatt, J. (1988). Estimation of Optical Pathlength through Tissue from Direct Time of Flight Measurement. *Phys. Med. Biol.* 33, 1433–1442. doi:10.1088/0031-9155/33/12/008

Dunbar, J., Hazell, G., and Jehanli, A. (2015). "Evaluation of a New point of Care Quantitative Cube Reader for Salivary Analysis in Premier League Soccer Clubs," in International Sports Science and Sports Medicine Conference, Newcastle upon Tyne, United Kingdom, August 8–10, 2015.

Eckmann, J.-P., and Ruelle, D. (1985). Ergodic Theory of Chaos and Strange Attractors. *Rev. Mod. Phys.* 57, 617–656. doi:10.1103/revmodphys.57.617

Engeland, C. G., Hugo, F. N., Hilgert, J. B., Nascimento, G. G., Junges, R., Lim, H.-J., et al. (2016). Psychological Distress and Salivary Secretory Immunity. *Brain Behav. Immun.* 52, 11–17. doi:10.1016/j.bbi.2015.08.017

Estevez, P. A., Tesmer, M., Perez, C. A., and Zurada, J. M. (2009). Normalized Mutual Information Feature Selection. *IEEE Trans. Neural Netw.* 20, 189–201. doi:10.1109/TNN.2008.2005601

Feinberg, I., and Floyd, T. C. (1979). Systematic Trends Across the Night in Human Sleep Cycles. *Psychophysiology* 16, 283–291. doi:10.1111/j.1469-8986.1979.tb02991.x

Fishera, A. J., Medaglia, J. D., and Jeronimus, B. F. (2018). Lack of Group-To-Individual Generalizability Is a Threat to Human Subjects Research. *Proc. Natl. Acad. Sci. USA* 115, E6106–E6115. doi:10.1073/pnas.1711978115

Gianferante, D., Thoma, M. V., Hanlin, L., Chen, X., Breines, J. G., Zoccola, P. M., et al. (2014). Post-Stress Rumination Predicts Hpa Axis Responses to Repeated Acute Stress. *Psychoneuroendocrinology* 49, 244–252. doi:10.1016/j.psyneuen.2014.07.021

Goldberger, A. L., and West, B. J. (1992). Chaos and Order in the Human Body. *CHANCE* 5, 47–55. doi:10.1080/09332480.1992.11882463

Hains, A. B., and Arnsten, A. F. T. (2008). Molecular Mechanisms of Stress-Induced Prefrontal Cortical Impairment: Implications for Mental Illness. *Learn. Mem.* 15, 551–564. doi:10.1101/lm.921708

Hardstone, R., Poil, S.-S., Schiavone, G., Jansen, R., Nikulin, V., Mansvelder, H., et al. (2012). Detrended Fluctuation Analysis: A Scale-Free View on Neuronal Oscillations. *Front. Physiol.* 3, 450. doi:10.3389/fphys.2012.00450

Hoshi, R. A., Pastre, C. M., Vanderlei, L. C. M., and Godoy, M. F. (2013). Poincaré Plot Indexes of Heart Rate Variability: Relationships with Other Nonlinear Variables. *Auton. Neurosci.* 177, 271–274. doi:10.1016/j.autneu.2013.05.004

Iber, C., Ancoli-Israel, S., Chesson, A., and Quan, S. (2017). For the American Academy of Sleep Medicine. The Aasm Manual for the Scoring of Sleep and Associated Events Rules, Terminology and Technical Specifications. *Darien, IL: Am. Acad. Sleep* 538, 1–49. Medicine Version 2.4.

Joëls, M., and Baram, T. Z. (2009). The Neuro-Symphony of Stress. *Nat. Rev. Neurosci.* 10, 459–466. doi:10.1038/nrn2632

Jönsson, P., Wallergård, M., Österberg, K., Åse, M. H., Johansson, G., and Karlson, B. (2010). Cardiovascular and Cortisol Reactivity and Habituation to a Virtual Reality Version of the Trier Social Stress Test. *A pilot study* 35, 1397–1403. doi:10.1016/j.psyneuen.2010.04.003

Kantz, H., and Schreiber, T. (2003). *Nonlinear Time Series Analysis*. New York, NY: Cambridge University Press.

Krämer, B., Diekhof, E. K., and Gruber, O. (2017). Effects of City Living on the Mesolimbic Reward System-An Fmri Study. *Hum. Brain Mapp.* 38, 3444–3453. doi:10.1002/hbm.23600

Kudielka, B. M., von Känel, R., Preckel, D., Zgraggen, L., Mischler, K., and Fischer, J. E. (2006). Exhaustion Is Associated with Reduced Habituation of Free Cortisol Responses to Repeated Acute Psychosocial Stress. *Biol. Psychol.* 72, 147–153. doi:10.1016/j.biopsycho.2005.09.001

Levenstein, S., Prantera, C., Varvo, V., Scribano, M. L., Berto, E., Luzi, C., et al. (1993). Development of the Perceived Stress Questionnaire: A New Tool for Psychosomatic Research. *J. Psychosomatic Res.* 37, 19–32. doi:10.1016/0022-3999(93)90120-5

Liang, Z., and Chapa Martell, M. A. (2018). Validity of Consumer Activity Wristbands and Wearable Eeg for Measuring Overall Sleep Parameters and Sleep Structure in Free-Living Conditions. *J. Healthc. Inform. Res.* 2, 152–178. doi:10.1007/s41666-018-0013-1

Liang, Z., and Chapa-Martell, M. A. (2019). Accuracy of Fitbit Wristbands in Measuring Sleep Stage Transitions and the Effect of User-specific Factors. *JMIR Mhealth Uhealth* 7, e13384. doi:10.2196/13384

Liang, Z., Ploderer, B., Liu, W., Nagata, Y., Bailey, J., Kulik, L., et al. (2016). Sleepexplorer: A Visualization Tool to Make Sense of Correlations between Personal Sleep Data and Contextual Factors. *Personal. Ubiquitous Comput.* 20, 985–1000. doi:10.1007/s00779-016-0960-6

López-Ruiz, R., Mancini, H. L., and Calbet, X. (1995). A Statistical Measure of Complexity. *Phys. Lett.* A209, 321–326.

Luke, R., Larson, E., Shader, M. J., Innes-Brown, H., Yper, L. V., Lee, A. K., et al. (2021). Analysis Methods for Measuring Fnirs Responses Generated by a Block-Design Paradigm. *Neurophotonics* 8, 025008. doi:10.1117/1.NPh.8.2.025008

Ma, Y., Shi, W., Peng, C.-K., and Yang, A. C. (2018). Nonlinear Dynamical Analysis of Sleep Electroencephalography Using Fractal and Entropy Approaches. *Sleep Med. Rev.* 37, 85–93. doi:10.1016/j.smrv.2017.01.003

Menghini, L., Cellini, N., Goldstone, A., Baker, F. C., and Zambotti, M. D. (2020). A Standardized Framework for Testing the Performance of Sleep-Tracking Technology: Step-by-Step Guidelines and Open-Source Code. *Sleep* 44 (2), zsaa170. doi:10.1093/sleep/zsaa170

Mitsuishi, H., Okamura, H., Yusuke, M., and Yoshiko, A. (2019). "The Validity of Salivary Cortisol Analysis Using Cube Reader [in Japanese]," in The Proceedings of the Annual Convention of the Japanese Psychological Association, Osaka, Japan, September 11—13, 2019(The Japanese Psychological Association) 83. doi:10.4992/pacjpa.83.0_2c-028

Molenaar, P. C. M. (2004). A Manifesto on Psychology as Idiographic Science: Bringing the Person Back into Scientific Psychology, This Time Forever. *Meas. Interdiscip. Res. Perspect.* 2, 201–218. doi:10.1207/s15366359mea0204_1

Molenaar, P. C. M., and Campbell, C. G. (2009). The New Person-specific Paradigm in Psychology. *Curr. Dir. Psychol. Sci.* 18, 112–117. doi:10.1111/j.1467-8721.2009.01619.x

M. R. Mehl and T. S. Conner (Editors) (2012). *Handbook of Research Methods for Studying Daily Life* (New York, NY: The Guilford Press).

Nejati, V., Majdi, R., Salehinejad, M. A., and Nitsche, M. A. (2021). The Role of Dorsolateral and Ventromedial Prefrontal Cortex in the Processing of Emotional Dimensions. *Scientific Rep.* 11, 1971. doi:10.1038/s41598-021-81454-7

Oster, H., Challet, E., Ott, V., Arvat, E., de Kloet, E. R., Dijk, D.-J., et al. (2017). The Functional and Clinical Significance of the 24-hour Rhythm of Circulating Glucocorticoids. *Endocr. Rev.* 38, 3–45. doi:10.1210/er.2015-1080

Peng, C.-K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E., and Goldberger, A. (1994). Mosaic Organization of Dna Nucleotides. *Phys. Rev. A* 49, 1685–1689. doi:10.1103/PhysRevE.49.1685

Piccirillo, M. L., Beck, E. D., and Rodebaugh, T. L. (2019). A Clinician's Primer for Idiographic Research: Considerations and Recommendations. *Behav. Ther.* 50, 938–951. doi:10.1016/j.beth.2019.02.002

Pollonini, L., Olds, C., Abaya, H., Bortfeld, H., Beauchamp, M. S., and Oghalai, J. S. (2014). Auditory Cortex Activation to Natural Speech and Simulated Cochlear Implant Speech Measured with Functional Near-Infrared Spectroscopy. *Hear. Res.* 309, 84–93. doi:10.1016/j.heares.2013.11.007

Rampino, A., Torretta, S., Rizzo, G., Viscanti, G., Quarto, T., Gelao, B., et al. (2019). Emotional Stability Interacts with Cortisol Levels before Fmri on Brain Processing of Fearful Faces. *Neuroscience* 416, 190–197. doi:10.1016/j.neuroscience.2019.08.002

Richman, J., and Moorman, J. R. (2000). Physiological Time-Series Analysis Using Approximate Entropy and Sample Entropy. *Am. J. Physiology-Heart Circulatory Physiol.* 278, H2039–H2049. doi:10.1152/ajpheart.2000.278.6.h2039

Rosenbaum, D., Hilsendegen, P., Thomas, M., Haeussinger, F. B., Metzger, F. G., Nuerk, H.-C., et al. (2018). Cortical Hemodynamic Changes during the Trier Social Stress Test: An Fnirs Study. *Neuroimage* 1, 107–115. doi:10.1016/j.neuroimage.2017.12.061

Rosenbaum, D., Int-Veen, I., Laicher, H., Torka, F., Kroczek, A., Rubel, J., et al. (2021). Insights from a Laboratory and Naturalistic Investigation on Stress, Rumination and Frontal Brain Functioning in Mdd: An Fnirs Study. *Neurobiol. Stress* 15, 100344. doi:10.1016/j.ynstr.2021.100344

Ross, B. C. (2014). Mutual Information between Discrete and Continuous Data Sets. *PLoS ONE* 9, e87357. doi:10.1371/journal.pone.0087357

Schaal, N. K., Hepp, P., Schweda, A., Wolf, O. T., and Krampe, C. (2019). A Functional Near-Infrared Spectroscopy Study on the Cortical Haemodynamic Responses during the Maastricht Acute Stress Test. *Sci. Rep.* 9, 13459. doi:10.1038/s41598-019-49826-2

Schommer, N. C., Hellhammer, D. H., and Kirschbaum, C. (2003). Dissociation between Reactivity of the Hypothalamus-Pituitary-Adrenal Axis and the Sympathetic-Adrenal-Medullary System to Repeated Psychosocial Stress. *Psychosom Med.* 65, 450–460. doi:10.1097/01.psy.0000035721.12441.17

Steinbrink, J., Wabnitz, H., Obrig, H., Villringer, A., and Rinneberg, H. (2001). Determining Changes in Nir Absorption Using a Layered Model of the Human Head. *Phys. Med. Biol.* 46, 879. doi:10.1088/0031-9155/46/3/320

Strangman, G. E., Zhang, Q., and Li, Z. (2014). Scalp and Skull Influence on Near Infrared Photon Propagation in the Colin27 Brain Template. *NeuroImage* 85, 136–149. doi:10.1016/j.neuroimage.2013.04.090

Tachtsidis, I., and Scholkmann, F. (2016). False Positives and False Negatives in Functional Near-Infrared Spectroscopy: Issues, Challenges, and the Way Forward. *Neurophotonics* 3, 031405. doi:10.1117/1.nph.3.3.031405

Tak, S., and Ye, J. C. (2014). Statistical Analysis of Fnirs Data: A Comprehensive Review. *Neuroimage* 85, 72–91. doi:10.1016/j.neuroimage.2013.06.016

Toyoda, H., Kashikura, K., Okada, T., Nakashita, S., Honda, M., Yonekurae, Y., et al. (2008). Source of Nonlinearity of the Bold Response Revealed by Simultaneous Fmri and Nirs. *NeuroImage* 39, 997–1013. doi:10.1016/j.neuroimage.2007.09.053

van Ockenburg, S. L., Booij, S. H., Riese, H., Rosmalen, J. G. M., and Janssens, K. A. M. (2015). How to Assess Stress Biomarkers for Idiographic Research? *Psychoneuroendocrinolo* 62, 189–199. doi:10.1016/j.psyneuen.2015.08.002

Wolfram, M., Bellingrath, S., Feuerhahn, N., and Kudielka, B. M. (2013). Cortisol Responses to Naturalistic and Laboratory Stress in Student Teachers: Comparison with a Non-Stress Control Day. *Stress and Health* 29, 143–149. doi:10.1002/smi.2439

Yanagisawa, K., Masui, K., Furutani, K., Nomura, M., Yoshida, H., and Ura, M. (2011). Temporal Distance Insulates against Immediate Social Pain: an Nirs Study of Social Exclusion. *Soc. Neurosci.* 6, 377–387. doi:10.1080/17470919.2011.559127

Yang, H., Zhou, Z., Liu, Y., Ruan, Z., Gong, H., Luo, Q., et al. (2007). Gender Difference in Hemodynamic Responses of Prefrontal Area to Emotional Stress by Near-Infrared Spectroscopy. *Behav. Brain Res.* 178, 172–176. doi:10.1016/j.bbr.2006.11.039

Yuen, E. Y., Liu, W., Karatsoreos, I. N., Feng, J., McEwen, B. S., and Yan, Z. (2009). Acute Stress Enhances Glutamatergic Transmission in Prefrontal Cortex and Facilitates Working Memory. *PNAS* 106, 14075–14079. doi:10.1073/pnas.0906791106

Zeier, H., Brauchli, P., and Joller-Jemelka, H. I. (1996). Effects of Work Demands on Immunoglobulin a and Cortisol in Air Traffic Controllers. *Biol. Psychol.* 42, 413–423. doi:10.1016/0301-0511(95)05170-8

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Computer Science

ORIGINAL RESEARCH
published: 07 January 2022
doi: 10.3389/fcomp.2021.770492

Predicting Activation Liking of People With Dementia

Lars Steinert, Felix Putze, Dennis Küster and Tanja Schultz*

Cognitive Systems Lab, Department of Mathematics and Computer Science, University of Bremen, Bremen, Germany

Physical, social and cognitive activation is an important cornerstone in non-pharmacological therapy for People with Dementia (PwD). To support long-term motivation and well-being, activation contents first need to be perceived positively. Prompting for explicit feedback, however, is intrusive and interrupts the activation flow. Automated analyses of verbal and non-verbal signals could provide an unobtrusive means of recommending suitable contents based on implicit feedback. In this study, we investigate the correlation between engagement responses and self-reported activation ratings. Subsequently, we predict ratings of PwD based on verbal and non-verbal signals in an unconstrained care setting. Applying Long-Short-Term-Memory (LSTM) networks, we can show that our classifier outperforms chance level. We further investigate which features are the most promising indicators for the prediction of activation ratings of PwD.

Keywords: dementia, activation, rating prediction, engagement, LSTM

OPEN ACCESS

Edited by:
Youngjun Cho,
University College London,
United Kingdom

Reviewed by:
Saturnino Luz,
University of Edinburgh,
United Kingdom
Emilie Brotherhood,
University College London,
United Kingdom

****Correspondence:***
Lars Steinert
lars.steinert@uni-bremen.de

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *03 September 2021*
Accepted: *13 December 2021*
Published: *07 January 2022*

Citation:
Steinert L, Putze F, Küster D and Schultz T (2022) Predicting Activation Liking of People With Dementia. Front. Comput. Sci. 3:770492. doi: 10.3389/fcomp.2021.770492

1. INTRODUCTION

Dementia describes a syndrome that is characterized by the loss of cognitive function and behavioral changes. This includes memory, language skills, and the ability to focus and pay attention (WHO, 2017). It has been shown that the physical, social, and cognitive stimulation of People with Dementia (PwD) has significant positive effects on their cognitive functioning (Spector et al., 2003; Woods et al., 2012) and can lead to a higher quality of life (Schreiner et al., 2005; Cohen-Mansfield et al., 2011). It is furthermore often (implicitly) assumed, that activation contents need to be perceived positively to help maintain long-term motivation and well-being. This can be supported by a recommender system that suggests appropriate activation contents. Here, an activation content is defined as a stimulus of a certain type (image gallery, video, audio, quiz, game, phrase or text) on a certain topic, e.g. gardening, sports, or animals to cognitively, socially, or physically activate PwD and which aims for the general maintenance or enhancement of the according functions (Clare and Woods, 2004). However, prompting for explicit user feedback is intrusive as it disturbs the activation flow. Studies have shown that verbal and non-verbal signals can be promising indicators for the internal states of healthy individuals (Masip et al., 2014; Tkalčič et al., 2019). Even PwD who might suffer from blunted affect or aphasia, might remain able to provide verbal and non-verbal signals throughout all stages of the disease (Steinert et al., 2021). For this study, we use the I-CARE dataset (Schultz et al., 2018, 2021) which consists of verbal and non-verbal signals of PwD who used a tablet-based activation system over multiple sessions in an unconstrained care setting. Previous studies have already investigated the recognition of engagement of PwD (Steinert et al., 2020, 2021), which is defined as "the act of being occupied or involved with an external stimulus" (Cohen-Mansfield et al., 2009). Here, we explicitly consider the argument that activation contents should not only be engaging but also need to be perceived positively to maintain long-term motivation and well-being. In this study, we thus first investigate the correlation between engagement responses and self-reported activation ratings.

Second, we analyze if self-reported activation ratings of PwD can be predicted based on verbal and non-verbal signals. Third, we explore the permutation-based feature importance of our classifier to generate hypotheses about possible underlying mechanisms. Last, we discuss the unique challenges involved with predicting activation ratings of elderly PwD. To the best of our knowledge, there are no prior studies that have investigated the prediction of activation ratings of PwD based on verbal and non-verbal signals.

2. RELATED WORKS

Research into the preservation of cognitive resources of PwD has a long history. A number of studies have investigated the effects of activation on perceived well-being, affect, engagement, and other affective states. However, detecting and interpreting the verbal and non-verbal signals of PwD can be particularly challenging due to the broad range of deleterious effects of aphasia or blunted affect on communication (Jones et al., 2015; WHO, 2017). In this section, we will (1) provide an overview of different non-pharmacological interventions that target the activation of PwD and (2) highlight relevant research into the production of (interpretable) verbal and non-verbal signals of PwD.

Over 20 years ago, Olsen et al. (2000) introduced "Media Memory Lane," a system that provides nostalgic music and videos to elicit long term memory stimulation for people with Alzheimer's Disease (AD). An evaluation of this system with 15 day care clients showed positive effects on engagement, affect, activity-related talking, and reduced fidgeting. Astell et al. (2010) evaluated the Computer Interactive Reminiscence and Conversation Aid (CIRCA) system, a touch screen system that presents photographs, music and video clips to enhance the interaction between PwD and caregivers. Their study demonstrated significant differences in verbal and non-verbal behavior when comparing the system with traditional reminiscence therapy sessions. Smith et al. (2009) produced audiovisual biographies based on photographs and personally meaningful music in cooperation with families of PwD. They further used a television set and a DVD player as a familiar interface for their participants. Several studies have also proposed music as a promising factor in non-pharmacological approaches (Spiro, 2010). Accordingly, Riley et al. (2009) introduced a touch screen system that allows PwD to create music regardless of any prior musical knowledge. Evaluating the system in three pilot studies, the authors reported engagement in the activity for all participants. Manera et al. (2015) developed a tablet-based kitchen and cooking simulation for elderly people with mild cognitive impairment. After four weeks of training, most participants rated the experience to be interesting, highly satisfying, and as eliciting more positive than negative emotions. Together, these findings underline the positive effects of non-pharmacological interventions for PwD, as well as for their (in)formal caregivers.

Asplund et al. (1995) investigated affect in the facial expressions of four severe demented participants during activities such as morning care or playing music. The authors compared unstructured judgements of facial expressions with assessments using the Facial Action Coding System [FACS, Ekman et al. (2002)] and showed that while facial cues become sparse and unclear, they are still interpretable to a certain degree. Mograbi et al. (2012) conducted a study with 22 participants with mild to moderate dementia who watched films for emotion elicitation. The authors manually annotated facial expressions, namely happiness, surprise, fear, sadness, disgust, anger, and contempt of the PwD and the controls. While they reported little difference in their production, PwD showed a narrower range of expressions which were less intense. This is in line with other studies that report that PwD may suffer from emotional blunting (Kumfor and Piguet, 2012; Perugia et al., 2020). To examine the quality and the decrease of emotional responses of PwD, Magai et al. (1996) conducted a study with 82 PwD with moderate or severe dementia and their families. Two research assistants were trained to manually code the participants' affective behavior, namely interest, joy, sadness, anger, contempt, fear, disgust, and knit brow expressions. Their results suggest that emotional expressivity, however, may not vary much depending on the stage of the disease.

Another important modality for the recognition of affective states is speech (Schuller, 2018). Nazareth (2019) demonstrated that lexical and acoustic features can be used to predict emotional valence in spontaneous speech of elderly. However, research has shown that speech also undergoes disease-related changes in dementia, e.g. impairments in the production of prosody (Roberts et al., 1996; Horley et al., 2010). This is particularly pertinent in frontotemporal dementia (Budson and Kowall, 2011).

Overall, there seems to be no strong direct link between the ability to produce (interpretable) verbal and non-verbal signals of emotions and the stage of the disease. It rather appears to be a combination of multiple factors such as the dementia type, co-morbidities, medication, and personality. Also, the context seems to play a role. Lee et al. (2017) showed that social and verbal interactions increase positive emotional responses. Notably even the merely implicit presence of a friend has been shown to be sufficient for eliciting this effect in healthy adults (Fridlund, 1991). Thus, emotional expressiveness appears to be extremely sensitive to contextual factors, and PwD might stand to benefit from such factors.

3. DATA COLLECTION

3.1. I-CARE System

The dataset used in this study was collected with the I-CARE system. I-CARE is a tablet-based activation system that is designed to be jointly used by PwD and (in)formal caregivers. The system is mobile and can be used at any location with and internet connection. It provides 346 user-specific activation contents (image galleries, videos, audios, quizzes, games, phrases and texts) on various topics such as gardening, sports, baking, or animals. The system also allows for the uploading of one's own contents to put more emphasis on biographical work (Schultz et al., 2018, 2021). At the same time, it allows for a multimodal

FIGURE 1 | The left figure shows two participants and one instructor from a project partner, who explains the procedure. The right figure shows two participants during an activation session (©AWO Karlsruhe).

data collection using the tablet's camera and microphone to capture video (30 FPS) and audio signals (16 kHz), respectively. The tablet used in the present work was a Google Pixel C (10.2-inch display) or Huawei MediaPad M5 (10.8-inch display). **Figure 1** shows exemplary how an activation session could look like.

3.2. Experimental Setting

The data collection for this study was conducted in different care facilities in Southern Germany as a part of the I-CARE project (Schultz et al., 2018, 2021). Participants of the study were PwD who fulfilled the clinical criteria for dementia according to the ICD-10 system (Alzheimer dementia, vascular dementia, frontotemporal dementia, Korsakoff's syndrome, or Dementia Not Otherwise Specified) ranging from mild to severe, and their (in)formal caregivers. All participants provided written consent and there was no financial compensation. For this study, a setup with minimal supervision and setup requirements was selected with activation sessions taking place in private rooms or in commonly used spaces in the care facilities. The tablet was placed on a stand in front of the participant with dementia so that their face was well-aligned with the field of view of the tablet camera.

At the beginning of each session, the system enquired about the daily well-being ("How are you today?") of the PwD using a smiley rating scale (positive, neutral, negative). Next, the system's recommender system suggested four different activation items, based on interests, personal information of the PwD, and previous ratings. The system also provided the opportunity to search for specific contents and view an activation history. Next, the PwD chose the activation content, e.g. an image gallery on baking, a video on gardening and so on. After each activation, the system asked the PwD for a rating of how well they liked the activation ("Did you enjoy the content?"), again, on a smiley rating scale (positive, neutral, negative). **Figure 2** shows the thumbnail images of four activation recommendations (left) and the rating options after the activation (right). Following the smiley rating, the system went directly back to the overview with recommended activation contents. Here, the PwD could decide whether or not to continue with another activation. Usually, activation sessions consisted of multiple individual activations.

The dataset used in this study consists of 187 activation sessions comprising 804 individual activations and, correspondingly, 804 activation ratings. These sessions cover 25 PwD (gender: 15 f, 10 m; age: 58–95 years, *M*: 82.4 years, *SD*: 9.0 years; dementia stage: 8 mild-moderate, 5 severe, 12 unspecified). Individual participants contributed with different number of sessions ($M = 7.48$, $SD = 2.42$, $Min = 2$, $Max = 12$).

4. METHODS

4.1. Rating Measurement

Self-reported activation ratings of the PwD were collected using an smiley rating scale (positive, neutral, negative) at the end of each activation. **Figure 3** shows the distribution of activation ratings for the participants individually and in total. The colors correspond to the rating (positive = green, neutral = yellow, negative = red). It is evident that activation contents were more frequently perceived as positive than neutral or negative by most participants. A Kruskal-Wallis test shows that these differences are statistically significant ($H = 54.571$, $p < 0.001$). Accordingly, investigating the class distribution across all participants provides a similar picture (positive = 68.23 %, neutral = 25.46 %, negative = 6.3 %). This demonstrates that the activation contents were mostly perceived positively.

4.2. Engagement Analysis

While effective activation contents are typically perceived as positive, not all positive contents are likely to be highly engaging. Furthermore, activation contents will only be effective in the long run if they succeed in engaging PwD. Thus, predicting engagement from verbal and non-verbal signals can be regarded as a separate challenge. As shown by previous work (Steinert et al., 2020, 2021), engagement can indeed be automatically recognized from verbal and non-verbal signals. Engagement in I-CARE was annotated retrospectively based on audio-visual data using the "Video Coding-Incorporating Observed Emotion" (VC-IOE) protocol (Jones et al., 2015) by two independent raters. We computed Cohen's Kappa (κ) between both raters after intensive training on six random test sessions to evaluate inter-rater reliability. The VC-IOE defines different engagement dimensions which were evaluated separately. These are emotional ($\kappa = 0.824$), verbal ($\kappa = 0.783$), visual ($\kappa = 0.887$), behavioral ($\kappa = 0.745$), and agitation ($\kappa = 0.941$) [1]. To obtain the level of engagement for each activation content, we calculated an engagement score by summing up the number of positive engagement outcomes per dimension over all frames of an activation content, divided by the total number of frames covering that activation.

Figure 4 shows the distribution of engagement scores with regards to the self-reported activation ratings of the participants. A Kruskal-Wallis test demonstrated a statistically significant difference ($H = 7.199$, $p < 0.05$) in the group means between the negative ($M = 0.75$, $SD = 0.56$), the neutral ($M = 0.78$, *SD*

[1]The VC-IOE further suggests collective engagement as a dimension which is defined as "Encouraging others to interact with STIMULUS. Introducing STIMULUS to others." (Jones et al., 2015). We interpreted "others" as third persons who did not originally take part in the session. As collective engagement was not apparent in this dataset, we dismissed this dimension.

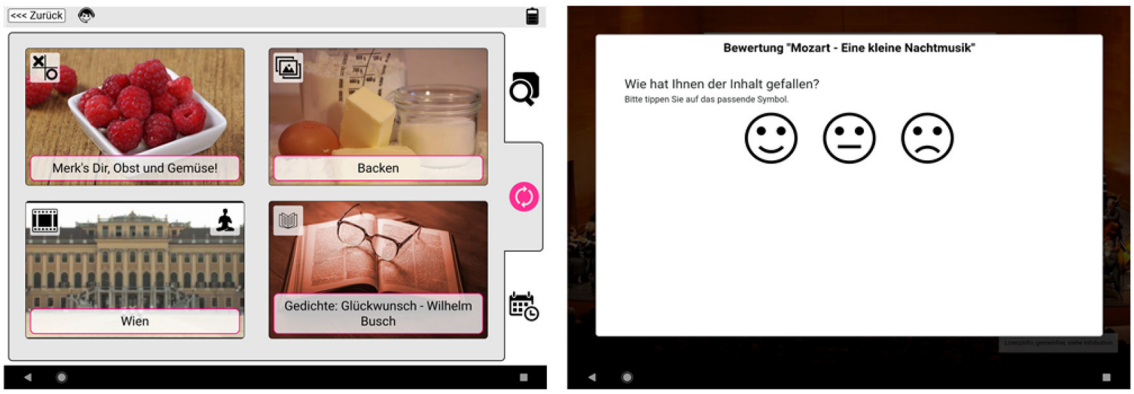


FIGURE 2 | User interface of the I-CARE system. The left figure shows the activation recommendations (top left: memory game, top right: image gallery, bottom left: video, bottom right: phrase). The right figure illustrates the rating options after the activation.

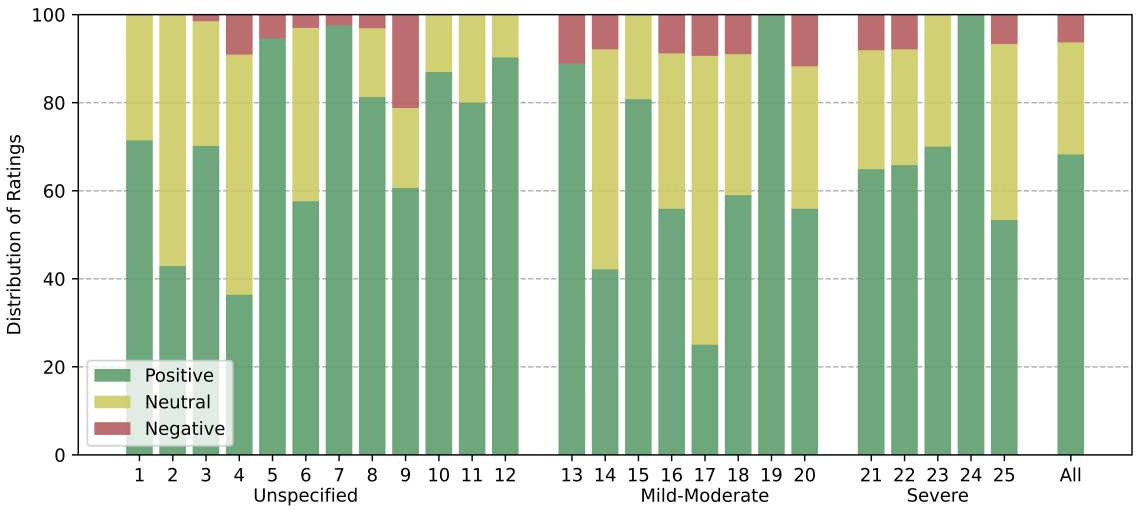


FIGURE 3 | Rating distribution for individual participants **(left)** and in total **(right)** grouped based on their dementia stage. Bar colors correspond to the ratings (positive = green; neutral = yellow; negative = red). Positive ratings significantly outweigh neutral and negative ratings ($H = 54.571$, $p < 0.001$).

$= 0.51$) and the positive class ($M = 0.89$, $SD = 0.47$), indicating a small effect of slightly more evidence for engagement toward positively evaluated activations compared to more negatively perceived contents. Similarly, a Spearman rank correlation analysis ($\rho = 0.094$, $p < 0.001$) showed a significant but small correlation between the engagement score and the rating of individual activation contents.

4.3. Multimodal Features

Human affective behavior and signaling is multimodal by nature. Thus, it can only be fully interpreted by jointly considering information from different modalities (Pantic et al., 2005). We argue that this is especially valid for PwD in an unconstrained care setting because PwD might suffer from aphasia or blunted affect (Kumfor and Piguet, 2012; Perugia et al., 2020). As individual channels begin to degrade, compensation by other channels is well-known to become more important. However, PwD may not only face greater challenges when decoding signals from by their interaction partners (receiver role) - but also with respect to clearly encoding their own socio-emotional signals in any individual channel (sender role). The Signal-to-Noise Ratio (SNR) can also be low for some modalities due to (multiple) background speakers, room reverberation or adverse lighting conditions. Accordingly, we use video-based features (OpenFace, OpenPose, and VGG-FACE) and audio-based (ComParE, DeepSpectrum) features, for the prediction of activation liking of PwD.

4.3.1. Video

The face is arguably the most important non-verbal source for information about another person's affective states (Kappas et al., 2013) and can provide information about affective states throughout all stages of dementia (see section 2). Here, we use the video signal captured with the tablet's camera to detect,

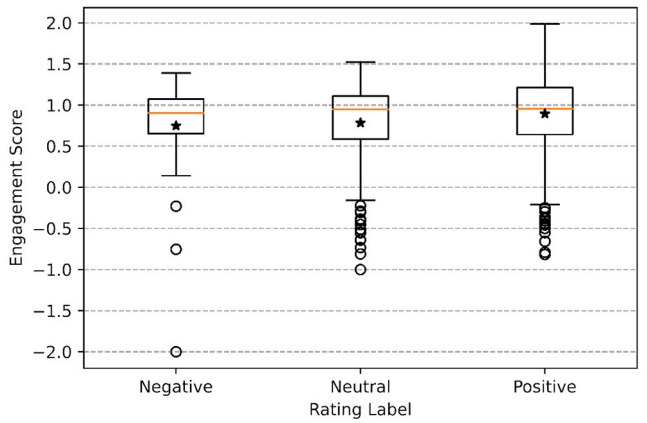


FIGURE 4 | Distribution of engagement scores with regards to the self-reported activation ratings (negative, neutral, positive). There are statistically significant difference ($H = 7.199$, $p < 0.05$) in the group means between the negative ($M = 0.75$, $SD = 0.56$), the neutral ($M = 0.78$, $SD = 0.51$) and the positive class ($M = 0.89$, $SD = 0.47$). Outliers were not removed from further analyses. The * symbol indicates the arithmetic mean.

align, and crop faces from the participants with dementia. From these pre-processed video frames, we extract facial features, namely the (binary scaled) presence of 18 and the (continuously scaled) intensity of 17 Action Units (AUs)[2] ranging from 0 to 5, the location and rotation of the head (head pose), and the direction of eye gaze in world coordinates using OpenFace 2.0 (Baltrusaitis et al., 2018). In the same vein, we extract skeleton features using OpenPose (Cao et al., 2019) to calculate relevant features, namely the distance between shoulders, eyes, ears, hands to nose, and the visibility of the hands. Last, we apply transfer learning using the pre-trained VGG-Face network (Parkhi et al., 2015). We retrained the network for five epochs using the FER2013 dataset with stochastic gradient descent, a learning rate of 0.0001, and a momentum of 0.9. Next, all video frames are rescaled to 224x224 pixels to match the input size of the Convolutional Neural Network (CNN), and normalized by subtracting the mean. The feature vectors for each video frame is the extracted from the *fc6* layer of the network. Overall, concatenating the feature vectors from all feature extractors leads to a 4138-dimensional feature vector for each video frame.

4.3.2. Audio

The recognition of affective states from speech is also a highly active research area (Akçay and Oğuz, 2020). While previous research has shown that speech undergoes disease-related changes in dementia, e.g. impairments in the production of prosody (Roberts et al., 1996; Horley et al., 2010), recent studies suggest that speech of PwD may still help to improve the automatic recognition of engagement (Steinert et al., 2021). We first apply denoising on all raw audio files recorded with the tablet's microphone to remove stationary and non-stationary background sounds, and to enhance participant's speech (Defossez et al., 2020). From the denoised audios, we extract the 2013 Interspeech Computational Paralinguistics Challenge features set (ComParE) using OpenSMILE (Eyben et al., 2010, 2013). We extract audio frame-wise (60 ms frame size; 10 ms steps) frequency, energy, and spectral related Low-Level Descriptors (LLD) which leads to a 130-dimensional feature vector (65 LLDs + deltas) for each step of 10 ms. Next, we create mel spectrograms using Hanning windows (512 samples size, 256 samples steps). We forward spectrograms (227x227 pixels, viridis colormap) to the pre-trained CNN AlexNet to receive bottleneck features from the *fc7* layer which results in a 4096-dimensional feature vector (Amiriparian et al., 2017).

4.4. Data Pre-processing

To take interpersonal and intrapersonal variations into account, we scale each feature to a range between zero and one. We assume that the verbal and non-verbal signals from the time interval shortly before the rating are likely to be most diagnostic for the subsequent activation rating. Correspondingly, we consider the 30 s of verbal and non-verbal signals before the rating was provided. Next, we slice features into 1 s segments with 25 % overlap and assign each segment to the corresponding rating label. Due to the class imbalance (see **Figure 3**), we combine the neutral and negative classes to formulate a two-class prediction problem. This seems reasonable as especially the prediction of positively perceived activation contents is relevant for an individual's well-being and motivation (Cohen-Mansfield, 2018). These pre-processed and labeled feature sequences are then forwarded to the classifier.

4.5. Prediction and Evaluation

The applied prediction approach is based on Long-Short-Term-Memory (LSTM) networks which allow for the preservation of temporal dependencies. This is especially important as verbal and non-verbal signals such as speech or facial expressions are subject to continuous change, especially in interactive activation sessions. Due to the different sampling rates of the feature sets of video and audio features (ComParE and DeepSpectrum), the classifier consists of three different input branches. Each input branch consists of a CNN layer (filter size = 256, 64, 256) followed by a MaxPooling layer (pool size = 3, 5, 3). Next, outputs are forwarded to an LSTM layer (units = 512, 64, 512). The three resulting context vectors are concatenated and passed to a Dense layer (units = 256) followed by the output layer (units = 2) with a Softmax activation function which outputs the class prediction. **Figure 5** shows the proposed system architecture. For regularization, we use a dropout rate of 0.3 in the LSTM layers and after the concatenation layer. We train the model for 50 epochs with a batch size of 16. We use a cross-entropy loss function and Adam optimizer with a learning rate of 0.001. To retrieve the overall rating prediction from individual segments, we apply majority voting. We apply a session-independent model evaluation through 10-fold cross-validation on session level where individual folds contain multiple sessions (18–19) and, thus, multiple activation ratings (67–87) ranging from negative

[2]AU01, AU02, AU04, AU05, AU06, AU07, AU09, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, AU45. For AU28, OpenFace only provides information about whether the AU is present.

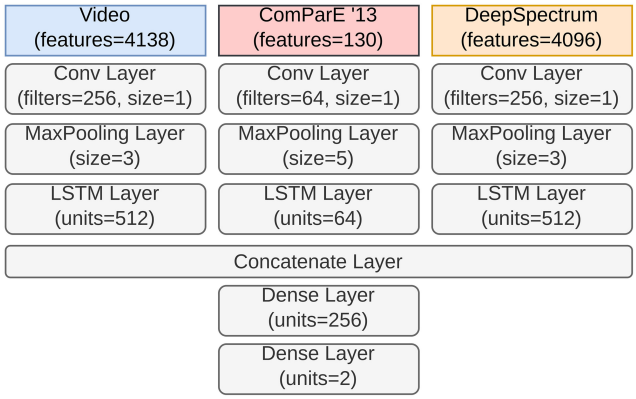


FIGURE 5 | Overview of our proposed prediction model.

to positive. Based on this approach, the proposed system learns behavioral characteristics elicited through subjective activation likings of multiple participants for inference on unseen sessions. The performance of our approach is compare to chance level. We select Unweighted Average Precision, Recall and F1-Score as the evaluation metrics as they are particularly suitable for unevenly distributed classes. To test for statistical significance between our model and the baseline, i.e. chance level, we apply a McNemar Test.

4.6. Permutation-Based Feature Importance

Explainable artificial intelligence has become an important research field in recent years (Linardatos et al., 2021). Knowing about the underlying mechanisms behind the predictions of black-box classifiers such as neural networks helps to understand and interpret their output. Accordingly, we compute permutation-based feature importances to investigate the importance of individual features for the prediction results (Molnar, 2020). For this, we break the association between individual features and labels by shuffling each feature sequence and adding random noise. For particularly relevant features, this should increase the model's prediction error, i.e. the cross-entropy loss (Kuhn and Johnson, 2013; Molnar, 2020). This is especially useful because it (1) provides insights into which verbal and non-verbal signals are relevant for the prediction of activation rating/ liking of PwD and allows for comparison with healthy individuals, and (2) it can help reveal irrelevant features, which can then be removed to decrease model complexity and computational costs.

5. RESULTS AND DISCUSSION

Table 1 shows the prediction results as the *M* and *SD*, Precision, Recall and F1-Score for each class individually and as an unweighted average over all folds. It is apparent that the model is especially capable of correctly predicting the positive class. A possible explanation for this may be the imbalance toward this

TABLE 1 | Prediction results based on the session-independent 10-fold cross-validation on session level.

Class	Precision	Recall	F1-Score
Pos.	0.726 (0.096)	0.754 (0.209)	0.729 (0.127)
Neu./ Neg.	0.308 (0.224)	0.364 (0.277)	0.328 (0.238)
Unweighted avg.	0.517 (0.272)	0.559 (0.312)	0.528 (0.277)
Chance	0.342 (0.354)	0.500 (0.513)	0.405 (0.417)

Results are reported as the M and SD Precision, Recall and F1 Score for each class individually and as the unweighted average over all folds.

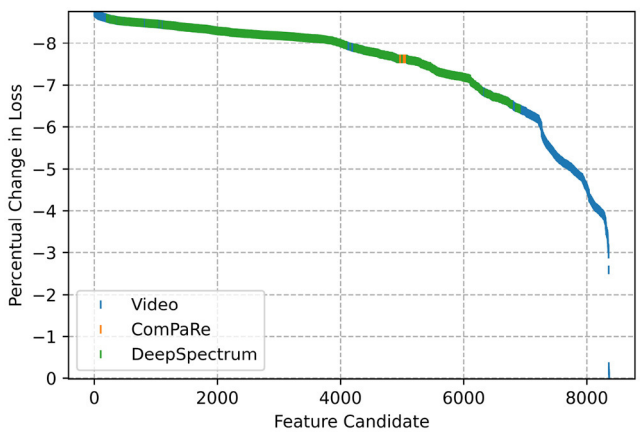


FIGURE 6 | Permutation-based feature importance averaged over all folds. The y-axis represents the perceptual change when comparing the cross-entropy loss before and after permutation, the x-axis shows the feature candidates. The colors indicates the set the feature belongs to (Video = blue, ComParE = Orange, DeepSpectrum = green).

class (see **Figure 3**). The model might not have seen a sufficient variation of data to accurately predict neutral and negative activation ratings. We also assume that participants showed only rather subtle negative expressions due to the highly supportive social context (Lee et al., 2017).

What stands out is that overall the prediction model significantly ($\chi^2 = 4.91$, $p < 0.05$) outperforms the baseline. Accordingly, verbal and non-verbal signals of PwD in different stages of the disease contain sufficient information for the prediction of activation ratings - despite the challenging recording conditions. The standard deviation indicates performance fluctuations throughout the folds. There are several possible explanations for this result. Participants in our study contributed substantially different numbers of sessions and, thus, different numbers of training samples (see section 3.2). As individual folds do not necessarily represent the overall data distribution, predictions can be based on a variable number of training samples of the same participant. The unstable recording conditions (background speakers, room reverberation, or lighting) throughout individual sessions might further increase the heterogeneity within folds. At the same time, this seems inevitable as the I-CARE system is designed for mobile usage.

Thus, these results are not comparable to clean and unambiguous data obtained in laboratory studies with healthy individuals.

Figure 6 provides an overview of the permutation-based feature importance averaged over all folds. The y-axis indicates the percentage change when comparing the cross-entropy loss before and after permutation. The bigger the negative change, the more important we consider the feature to be. This x-axis represents all 8364 feature candidates (see section 4.3). It is apparent that video-based and DeepSpectrum features seem to be important for the prediction. Especially video-based have been found as an import predictor in other tasks, namely the investigation of music (Tkalčič et al., 2019) or image (Masip et al., 2014) preferences. The curve progression further suggests that there are no individual features that stand out. Instead, it is rather the combination of different features on which the model relies. This finding could also be due to colinearity in the features, i.e. if one feature is permuted, the model relies on a highly correlated neighbor.

6. CONCLUSION

The main goal of the current study was to determine if activation ratings of PwD can be predicted in a real-life environment. We investigated a dataset collected with the I-CARE system of 25 PwD throughout all stages of the disease, and showed that contents provided by the system are mainly perceived positively, which can lead to more engagement and positive mood (Cohen-Mansfield, 2018). Moreover, participants' verbal and non-verbal signals contain sufficient information to successfully predict their activation ratings. Also, we could show that, in line with studies on healthy individuals (Masip et al., 2014; Tkalčič et al., 2019), the face remains an important source of information for inferring preferences. Interestingly, in our sample, there seems to be only a weak link between observed engagement and subjective activation liking. In general, this finding is indeed more consistent with prior reviews and meta-analyses focused on healthy adults, which have demonstrated only weak to moderate associations between subjective experience and different types of physiological or behavioral responses to emotion-eliciting stimuli in healthy adults (Mauss and Robinson, 2009; Hollenstein and Lanteigne, 2014). However, it is remarkable that (1) this relationship appears to be even further degraded among PwD and (2) that machine learning approaches based on multimodal data may still succeed in successfully predicting subjective ratings of PwD. At the same time, our approach still faces a number of limitations. A session-independent model evaluation implies the existence of annotated samples of the participants. While user-independent modeling would be preferable for the real-world application, this seems too ambitious with a small and heterogeneous dataset. As the presented results are not easily comparable to other studies, future work could also consider the assessments of the present caregivers. This could provide further information about the validity of our results. Despite these limitations, the present results make an important contribution to a, thus far, sparsely populated part of the field with regards to predicting activation liking of PwD.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available as the used dataset consists of data of People with Dementia. Requests to access the datasets should be directed to lars.steinert@uni-bremen.de.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University Of Bremen. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

LS conceived and designed the analyses, performed the analyses, and wrote the paper. FP conceived and designed the analyses, collected the data, and wrote the paper. DK conceived and designed the analyses and wrote the paper. TS conceived and designed the analyses, collected the data, and supervision of project. All authors contributed to the article and approved the submitted version.

FUNDING

This work was partially funded by the Klaus-Tschira-Stiftung. Data collection and development of the I-CARE system was funded by the BMBF under reference BMBF-number V4PIDO62. We also gratefully acknowledge the support of the Leibniz ScienceCampus Bremen Digital Public Health (lsc-diph.de), which is jointly funded by the Leibniz Association (W4/2018), the Federal State of Bremen and the Leibniz Institute for Prevention Research and Epidemiology—BIPS.

REFERENCES

Akçay, M. B., and Oğuz, K. (2020). Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun.* 116, 56–76. doi: 10.1016/j.specom.2019.12.001

Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Freitag, M., Pugachevskiy, S., et al. (2017). "Snore sound classification using image-based deep spectrum features," in *Interspeech 2017* (Stockholm), 3512–3516. doi: 10.21437/Interspeech.2017-434

Asplund, K., Jansson, L., and Norberg, A. (1995). Facial expressions of patients with dementia: A comparison of two methods of interpretation. *Int. Psychogeriatr.* 7, 527–534. doi: 10.1017/S1041610295002262

Astell, A. J., Ellis, M. P., Bernardi, L., Alm, N., Dye, R., Gowans, G., et al. (2010). Using a touch screen computer to support relationships between people with dementia and caregivers. *Interact. Comput.* 22, 267–275. doi: 10.1016/j.intcom.2010.03.003

Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L.-P. (2018). "Openface 2.0: facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (Xi'an: IEEE), 59–66. doi: 10.1109/FG.2018.00019

Budson, A. E., and Kowall, N. W. (2011). *The Handbook of Alzheimer's Disease and Other Dementias*, Vol. 7. Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9781444344110

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., and Sheikh, Y. A. (2019). Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 172–186. doi: 10.1109/TPAMI.2019.2929257

Clare, L., and Woods, R. T. (2004). Cognitive training and cognitive rehabilitation for people with early-stage Alzheimer's disease: a review. *Neuropsychol. Rehabil.* 14, 385–401. doi: 10.1080/09602010443000074

Cohen-Mansfield, J. (2018). Do reports on personal preferences of persons with dementia predict their responses to group activities? *Dement. Geriatr. Cogn. Disord.* 46, 100–108. doi: 10.1159/000491746

Cohen-Mansfield, J., Dakheel-Ali, M., and Marx, M. S. (2009). Engagement in persons with dementia: the concept and its measurement. *Am. J. Geriatr. Psychiatry* 17, 299–307. doi: 10.1097/JGP.0b013e31818f3a52

Cohen-Mansfield, J., Marx, M. S., Thein, K., and Dakheel-Ali, M. (2011). The impact of stimuli on affect in persons with dementia. *J. Clin. Psychiatry* 72:480. doi: 10.4088/JCP.09m05694oli

Defossez, A., Synnaeve, G., and Adi, Y. (2020). "Real time speech enhancement in the waveform domain," in *Interspeech* (Shanghai). doi: 10.21437/Interspeech.2020-2409

Ekman, P., Friesen, W. V., and Hager, J. C. (2002). *Facial Action Coding System (FACS), 2nd Edn*. Salt Lake City, UT: Research Nexus Division of Network Information Research Corporation.

Eyben, F., Weninger, F., Groß, F., and Schuller, B. (2013). "Recent developments in opensmile, the Munich open-source multimedia feature extractor," in *MM '13: Proceedings of the 21st ACM International Conference on Multimedia* (Barcelona). doi: 10.1145/2502081.2502224

Eyben, F., Wöllmer, M., and Schuller, B. (2010). "Opensmile: the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia, MM '10* (New York, NY: Association for Computing Machinery), 1459–1462. doi: 10.1145/1873951.1874246

Fridlund, A. J. (1991). Sociality of solitary smiling: potentiation by an implicit audience. *J. Pers. Soc. Psychol.* 60, 229–240. doi: 10.1037/0022-3514.60.2.229

Hollenstein, T., and Lanteigne, D. (2014). Models and methods of emotional concordance. *Biol. Psychol.* 98, 1–5. doi: 10.1016/j.biopsycho.2013.12.012

Horley, K., Reid, A., and Burnham, D. (2010). Emotional prosody perception and production in dementia of the Alzheimer's type. *J. Speech Lang. Hear. Res.* 53, 1132–1146. doi: 10.1044/1092-4388(2010/09-0030)

Jones, C., Sung, B., and Moyle, W. (2015). Assessing engagement in people with dementia: a new approach to assessment using video analysis. *Arch. Psychiatr. Nurs.* 29, 377–382. doi: 10.1016/j.apnu.2015.06.019

Kappas, A., Krumhuber, E., and Küster, D. (2013). "Facial behavior," in *Nonverbal Communication*, eds J. A. Hall and M. L. Knapp (Berlin: Mouton de Gruyter), 131–166. doi: 10.1515/9783110238150.131

Kuhn, M., and Johnson, K. (2013). *Applied Predictive Modeling*, Vol. 26. New York, NY: Springer. doi: 10.1007/978-1-4614-6849-3

Kumfor, F., and Piguet, O. (2012). Disturbance of emotion processing in frontotemporal dementia: a synthesis of cognitive and neuroimaging findings. *Neuropsychol. Rev.* 22, 280–297. doi: 10.1007/s11065-012-9201-6

Lee, K. H., Boltz, M., Lee, H., and Algase, D. L. (2017). Does social interaction matter psychological well-being in persons with dementia? *Am. J. Alzheimers Dis. Other Dement.* 32, 207–212. doi: 10.1177/1533317517704301

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021). Explainable AI: a review of machine learning interpretability methods. *Entropy* 23:18. doi: 10.3390/e23010018

Magai, C., Cohen, C., Gomberg, D., Malatesta, C., and Culver, C. (1996). Emotional expression during mid- to late-stage dementia. *Int. Psychogeriatr.* 8, 383–395. doi: 10.1017/S104161029600275X

Manera, V., Petit, P.-D., Derreumaux, A., Orvieto, I., Romagnoli, M., Lyttle, G., et al. (2015). "Kitchen and cooking," a serious game for mild cognitive impairment and Alzheimer's disease: a pilot study. *Front. Aging Neurosci.* 7:24. doi: 10.3389/fnagi.2015.00024

Masip, D., North, M. S., Todorov, A., and Osherson, D. N. (2014). Automated prediction of preferences using facial expressions. *PLoS ONE* 9:e87434. doi: 10.1371/journal.pone.0087434

Mauss, I. B., and Robinson, M. D. (2009). Measures of emotion: a review. *Cogn. Emot.* 23, 209–237. doi: 10.1080/02699930802204677

Mograbi, D. C., Brown, R. G., and Morris, R. G. (2012). Emotional reactivity to film material in Alzheimer's disease. *Dement. Geriatr. Cogn. Disord.* 34, 351–359. doi: 10.1159/000343930

Molnar, C. (2020). *Interpretable Machine Learning*. Morrisville: lulu.com.

Nazareth, D. S. (2019). "Emotion recognition in dementia: advancing technology for multimodal analysis of emotion expression in everyday life," in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* (Cambridge), 45–49. doi: 10.1109/ACIIW.2019.8925059

Olsen, R. V., Hutchings, B. L., and Ehrenkrantz, E. (2000). "Media memory lane" interventions in an Alzheimer's day care center. *Am. J. Alzheimers Dis.* 15, 163–175. doi: 10.1177/153331750001500307

Pantic, M., Sebe, N., Cohn, J. F., and Huang, T. (2005). "Affective multimodal human-computer interaction," in *Proceedings of the 13th Annual ACM International Conference on Multimedia, MULTIMEDIA '05* (Singapore: Association for Computing Machinery), 669–676. doi: 10.1145/1101149.1101299

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). "Deep face recognition," in *British Machine Vision Conference* (Swansea). doi: 10.5244/C.29.41

Perugia, G., Diaz-Boladeras, M., Catala, A., Barakova, E. I., and Rauterberg, M. (2020). ENGAGE-DEM: a model of engagement of people with dementia. *IEEE Trans. Affect. Comput.* 1. doi: 10.1109/TAFFC.2020.2980275

Riley, P., Alm, N., and Newell, A. (2009). An interactive tool to promote musical creativity in people with dementia. *Comput. Hum. Behav.* 25, 599–608. doi: 10.1016/j.chb.2008.08.014

Roberts, V. J., Ingram, S. M., Lamar, M., and Green, R. C. (1996). Prosody impairment and associated affective and behavioral disturbances in Alzheimer's disease. *Neurology* 47, 1482–1488. doi: 10.1212/WNL.47.6.1482

Schreiner, A. S., Yamamoto, E., and Shiotani, H. (2005). Positive affect among nursing home residents with Alzheimer's dementia: the effect of recreational activity. *Aging Mental Health* 9, 129–134. doi: 10.1080/13607860412331336841

Schuller, B. W. (2018). Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* 61, 90–99. doi: 10.1145/3129340

Schultz, T., Putze, F., Schulze, T., Steinert, L., Mikut, R., Doneit, W., et al. (2018). *I-CARE - Ein Mensch-Technik Interaktionssystem zur Individuellen Aktivierung von Menschen mit Demenz* (Oldenburg).

Schultz, T., Putze, F., Steinert, L., Mikut, R., Depner, A., Kruse, A., et al. (2021). I-CARE-an interaction system for the individual activation of people with dementia. *Geriatrics* 6:51. doi: 10.3390/geriatrics6020051

Smith, K. L., Crete-Nishihata, M., Damianakis, T., Baecker, R. M., and Marziali, E. (2009). Multimedia biographies: a reminiscence and social stimulus tool for persons with cognitive impairment. *J. Technol. Hum. Serv.* 27, 287–306. doi: 10.1080/15228830903329831

Spector, A., Thorgrimsen, L., Woods, B., Royan, L., Davies, S., Butterworth, M., et al. (2003). Efficacy of an evidence-based cognitive stimulation therapy programme for people with dementia: randomised controlled trial. *Brit. J. Psychiatry* 183, 248–254. doi: 10.1192/bjp.183.3.248

Spiro, N. (2010). Music and dementia: observing effects and searching for underlying theories. *Aging Ment. Health* 14, :891–899. doi: 10.1080/13607863.2010.519328

Steinert, L., Putze, F., Küster, D., and Schultz, T. (2020). "Towards engagement recognition of people with dementia in care settings," in *Proceedings of the 2020 International Conference on Multimodal Interaction* (Virtual Event), 558–565. doi: 10.1145/3382507.3418856

Steinert, L., Putze, F., Kuster, D., and Schultz, T. (2021). "Audio-visual recognition of emotional engagement of people with dementia," in *Proc. Interspeech 2021* (Brno), 1024–1028. doi: 10.21437/Interspeech.2021-567

Tkalčič, M., Maleki, N., Pesek, M., Elahi, M., Ricci, F., and Marolt, M. (2019). "Prediction of music pairwise preferences from facial expressions," in *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI '19* (New York, NY: Association for Computing Machinery), 150–159. doi: 10.1145/3301275.3302266

WHO (2017). *Dementia*. Available online at: https://www.who.int/news-room/fact-sheets/detail/dementia (accessed August 5, 2021).

Woods, B., Aguirre, E., Spector, A. E., and Orrell, M. (2012). Cognitive stimulation to improve cognitive functioning in people with dementia. *Cochrane Database Syst. Rev*. 2:CD005562. doi: 10.1002/14651858.CD005562.pub2

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computer Science

ORIGINAL RESEARCH
published: 07 January 2022
doi: 10.3389/fcomp.2021.808507

Implicit Estimation of Paragraph Relevance From Eye Movements

Michael Barz[1,2]*, *Omair Shahzad Bhatti*[1] *and Daniel Sonntag*[1,2]

[1] German Research Center for Artificial Intelligence, Interactive Machine Learning Department, Saarbrücken, Germany,
[2] Applied Artificial Intelligence, Oldenburg University, Oldenburg, Germany

Eye movements were shown to be an effective source of implicit relevance feedback in constrained search and decision-making tasks. Recent research suggests that gaze-based features, extracted from scanpaths over short news articles (g-REL), can reveal the perceived relevance of read text with respect to a previously shown trigger question. In this work, we aim to confirm this finding and we investigate whether it generalizes to multi-paragraph documents from Wikipedia (Google Natural Questions) that require readers to scroll down to read the whole text. We conduct a user study ($n = 24$) in which participants read single- and multi-paragraph articles and rate their relevance at the paragraph level with respect to a trigger question. We model the perceived document relevance using machine learning and features from the literature as input. Our results confirm that eye movements can be used to effectively model the relevance of short news articles, in particular if we exclude difficult cases: documents which are on topic of the trigger questions but irrelevant. However, our results do not clearly show that the modeling approach generalizes to multi-paragraph document settings. We publish our dataset and our code for feature extraction under an open source license to enable future research in the field of gaze-based implicit relevance feedback.

Keywords: implicit relevance feedback, reading analysis, machine learning, eye tracking, perceived paragraph relevance, eye movements and reading

OPEN ACCESS

Edited by:
Siyuan Chen,
University of New South Wales, Australia

Reviewed by:
Nora Castner,
University of Tübingen, Germany
Xi Wang,
ETH Zürich, Switzerland

***Correspondence:**
Michael Barz
michael.barz@dfki.de

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *03 November 2021*
Accepted: *12 December 2021*
Published: *07 January 2022*

Citation:
Barz M, Bhatti OS and Sonntag D (2022) Implicit Estimation of Paragraph Relevance From Eye Movements. Front. Comput. Sci. 3:808507. doi: 10.3389/fcomp.2021.808507

1. INTRODUCTION

Searching for information on the web or in a knowledge base is pervasive. However, search queries to information retrieval systems seldom represent a user's information need precisely (Carpineto and Romano, 2012). At the same time, a growing number of available documents, sources, and media types further increase the required effort to satisfy an information need. Implicit relevance feedback, obtained from users' interaction signals, was proposed to improve information retrieval systems as an alternative to more accurate, but costly explicit feedback (Agichtein et al., 2006). Behavioral signals that were investigated in this regard include clickthrough data (Agichtein et al., 2006; Joachims et al., 2017), dwell time of (partial) documents (Buscher et al., 2009), mouse movements (Eickhoff et al., 2015; Akuma et al., 2016), and eye movements (Buscher et al., 2012). This data may originate from search logs, which can be used to tune the ranking model of a search engine offline, or from real-time interaction data to extend search queries during a search session or to identify relevant text passages. In this work, we aim at identifying relevant paragraphs using real-time eye tracking data as input.

Eye movements play an important role in information acquisition (Gwizdka and Dillon, 2020) and were shown to be an effective source of implicit relevance feedback in search (Buscher et al., 2008a) and decision-making (Feit et al., 2020). However, eye movements highly depend on the user characteristics, the task at hand, and the content visualization (Buchanan et al., 2017). Related approaches use eye tracking to infer the perceived relevance of text documents with respect to previously shown trigger questions (Salojarvi et al., 2003, 2004, 2005a; Buscher et al., 2008a; Loboda et al., 2011; Gwizdka, 2014a; Bhattacharya et al., 2020a,b), and to extend (Buscher et al., 2008b; Chen et al., 2015) or generate search queries (Hardoon et al., 2007; Ajanki et al., 2009). A common disadvantage of approaches for gaze-based relevance estimation is that they are tested using documents with constrained layouts and topics such as single sentences (Salojarvi et al., 2003, 2004, 2005a) or short news articles that fit on the screen at once (Buscher et al., 2008a; Loboda et al., 2011; Gwizdka, 2014a; Bhattacharya et al., 2020a,b). Hence, it is unclear whether related findings generalize to more realistic settings such as those that include Wikipedia-like web documents.

We investigate whether eye tracking can be used to infer the perceived relevance of read documents with respect to previously shown trigger questions in a less constrained setting. We include multi-paragraph documents that exceed the display size and require scrolling to read the whole text. For this, we conduct a user study with $n = 24$ participants in which participants read single- and multi-paragraph articles and rate their relevance at the paragraph level while their eye movements are recorded. Pairs of single paragraph documents and questions are taken from the g-REL corpus (Gwizdka, 2014a). Multi-paragraph documents with corresponding questions are selected from the Google Natural Questions (GoogleNQ) corpus (Kwiatkowski et al., 2019). We assemble a corresponding dataset, the *gazeRE* dataset, and make it available to the research community under an open source license via Github (see section 3.5). Using the gazeRE dataset, we aim for confirming the findings from the literature on short news articles and investigate whether they generalize to the multi-paragraph documents from Wikipedia. We model the perceived relevance using machine learning and the features from Bhattacharya et al. (2020a) as input.

2. RELATED WORK

Prior research addressed the question whether eye movements can be linked to the relevance of a read text and how this implicit feedback can be leveraged in information retrieval settings.

2.1. Relevance Estimation From Reading Behavior

One group of work addressed the question whether the relevance of a text with respect to a task or trigger questions can be modeled using the user's gaze. For instance, Salojarvi et al. (2003, 2004, 2005a) investigated whether eye tracking can be used to estimate the user's perceived relevance of a document. They used machine learning to predict the relevance using the eye movements from reading the document titles as input. The authors organized a related research challenge, which is described in Salojarvi et al. (2005b). Loboda et al. (2011) presented an approach for gaze-based estimation of sentence relevance using fixations to sentence-terminal words, i.e., words at the end of a sentence, as there is empirical evidence that these words are fixated longer on average. This is known as the sentence wrap-up effect, which is a manifestation of the integrative process in reading. Buscher et al. (2008a) investigated the relation between reading behavior and document relevance using eye tracking technology. They found that the ratio of skimming is higher in irrelevant documents and the ratio of continuous reading behavior is higher for relevant documents. Further, they introduced the concept of attentive documents that keep track of the perceived relevance based on eye movements (Buscher et al., 2012). Gwizdka (2014a,b) modeled the relation between eye movements and perceived document relevance and investigated the cognitive effort involved in the relevance judgement. They introduced the g-REL corpus, a collection of short news stories and corresponding questions, which they used for collecting ground-truth and eye tracking data. The authors could confirm the findings from Buscher et al. (2012) that relevant documents tend to be read continuously, while irrelevant documents are rather skimmed (Gwizdka, 2014a). Akuma et al. (2016) compared gaze-based relevance feedback with implicit relevance feedback from more common sensors such as mouse movements. They found a high correlation between both feedback options and a relationship between gaze-based features and the perceived document relevance. Li et al. (2018) investigated the reading behavior for relevant and irrelevant documents for factual and intellectual tasks. Based on data from a user study, they suggested a two-staged reading model for explaining the cognitive processes inherent in relevance judgements. Jacob et al. (2018) investigated whether eye movements can be used to infer the interest of a reader in a currently read article. Bhattacharya et al. (2020b) encoded fixations from participants' scanpaths over documents from the g-REL corpus and trained a convolutional neural network (CNN) with the perceived relevance as prediction target. This approach is limited to small texts of similar lengths. Further, they suggested novel features based on the convex hull of scanpath fixations to model the participants' perceived relevance (Bhattacharya et al., 2020a). In addition, they simulated the user interaction to investigate whether their approach can be used in real-time scenarios by cumulatively adding fixations of the scanpath and normalizing the convex hull features with the elapsed time of interaction. Other related approaches include, for instance, a generic approach to map gaze-signals to HTML documents at the word level (Hienert et al., 2019). Davari et al. (2020) use this tool to investigate the role of word fixations in query term prediction. Feit et al. (2020) modeled the user-perceived relevance of information views in a graphical user interface for decision-making. They showed room advertisements in a web-based interface via multiple viewports and asked users what information was perceived as relevant for their decision to book a room or not. In this paper, we investigate whether the perceived relevance can be estimated for paragraphs of long Wikipedia-like documents in contrast to

sentences or short articles. This requires to compensate for the scrolling activity, which may distort the gaze signal and fixation extraction, and to develop a method for effectively extracting consecutive gaze sequences to individual paragraphs.

2.2. Query Expansion Methods

Other work focused on generating or expanding search queries based on the user's gaze behavior. Miller and Agne (2005) presented a system that extracts relevant search keywords from short texts based on eye movements. Hardoon et al. (2007) and Ajanki et al. (2009) proposed methods for implicitly generating search queries from eye movements during an information retrieval task. The generated query is used to proactively retrieve relevant documents using content-based ranking algorithms. Buscher et al. (2008b) proposed a technique for automatic query expansion and re-ranking for document retrieval. They use relevance estimates to identify recently read paragraphs that are relevant to the user and, eventually, to reformulate the search query. Chen et al. (2015) presented a query expansion method based on eye tracking and topic modeling. They identified fixated terms and modeled the user's latent intent using the Latent Dirichlet Allocation (LDA) for topic modeling.

2.3. Factors That Influence Eye Movements

Buchanan et al. (2017) surveyed works in the field of gaze-based implicit relevance feedback. They identified several factors that might influence gaze patterns and, hence, should be considered when building gaze-enhanced information retrieval systems. Key factors include the task type, the task complexity, individual differences such as expertise, and the presentation of the search results. For instance, Cole et al. (2013) showed that "the user's level of domain knowledge can be inferred from their interactive search behaviors." Bhattacharya and Gwizdka (2018) modeled the knowledge-change while reading using gaze-based features: a high change in knowledge coincides with significant differences in the scan length and duration of reading sequences, and in the number of reading fixations. Gwizdka (2017) investigated the task-related differences in reading strategies between word search and relevance decisions during information search. Eickhoff et al. (2015) studied the relationship between the user's visual attention to tokens in a search engine result page (SERP) or document and the corresponding search query: users fixate terms, which are part of their current query more often and longer than others. Further, they found that the semantic proximity of the search query to the user's attention increases for different reformulation strategies such as specialization, generalization, and reformulation.

3. USER STUDY

We conduct a user study ($n = 24$) with the goal to collect eye movement data during relevance estimation tasks. The participants are asked to read documents of different lengths and to judge, per paragraph, whether it provides an answer to a previously shown trigger question. We use this data to model the relation between the recorded eye movement data and the perceived relevance using machine learning (see section 4).

3.1. Participants

For our study, we invited 26 students (15 female) with an average age of 27.19 years ($SD = 5.74$). Data from two participants had to be discarded, because they withdrew their participation. The remaining participants reported to have normal (11) or corrected to normal (13) vision of which 11 wore eyeglasses and 2 wore contact lenses. Ten of them participated in an eye tracking study before. The participants rated their language proficiency in English for reading texts as native (1), fluent (18), or worse (5). Each participant received 15 EUR as compensation.

3.2. Stimuli

The stimuli data used in our study are pairs of trigger questions and documents with one or multiple paragraphs (see **Figure 1**). We use a subset from the g-REL corpus (Gwizdka, 2014a) with single-paragraph documents that fit on one page and selected pairs from the Google Natural Questions (NQ) corpus, which includes multi-paragraph documents that require scrolling (Kwiatkowski et al., 2019). Both corpora include relevance annotations per paragraph to which we refer as system relevance.

3.2.1. g-REL Corpus

The g-REL corpus includes a set of 57 trigger questions and 19 short English news texts that fit on one page. Questions include, for instance, "Where is the headquarters of OPEC located?" and "What was Camp David originally named?". The news texts are either irrelevant, topically relevant, or relevant with respect to these questions: the corpus includes three questions per document. If a document is irrelevant, it is off-topic and does not contain an answer to the question. Topically relevant and relevant documents are on topic, but only the relevant texts contain an answer to the question. The original news texts were selected from the AQUAINT Corpus of English News Texts (Graff, 2002) as used in the TREC 2005 Question Answering track.[1] The questions and judgements (system relevance) from TREC data were further revised and tested by Michael Cole and Jacek Gwizdka. Prior results for this corpus have been published in, e.g., Gwizdka (2014a,b, 2017), Bhattacharya et al. (2020a,b). Like Bhattacharya et al. (2020a,b), we consider a binary relevance classification. Hence, the topically relevant document-question pairs are counted as irrelevant ones.

For our user study, we select a balanced subset of 12 distinct documents of which four are relevant, four are topical, and four are irrelevant with respect to the accompanying trigger question. We select two additional documents for the training phase of which one is relevant and one is topical. We select the news texts such that the length distribution is similar to the whole corpus. The mean number of tokens of the selected news texts is 170.5 ($SD = 14.211$). The mean number of tokens, if all documents were included, is 176.404 ($SD = 12.346$). We used a simple whitespace tokenizer, which segments each document into a list words, to determine the number of tokens in each document.

[1] https://trec.nist.gov/data/qa.html

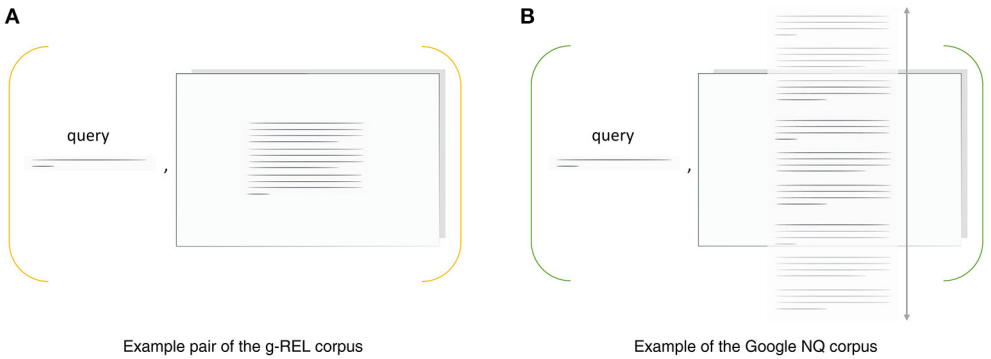


FIGURE 1 | We sample stimuli from **(A)** the g-REL corpus, which includes pairs of questions and short English news articles, and **(B)** from the Google NQ corpus, which includes pairs of questions and English Wikipedia articles.

3.2.2. Google Natural Questions Corpus

The Natural Questions (NQ) corpus[2] by Google includes 307k pairs of questions and related English Wikipedia documents (Kwiatkowski et al., 2019). Example questions include "What is the temperature at bottom of ocean?" and "What sonar device let morse code messages be sent underwater from a submarine in 1915?". Each document includes multiple HTML containers such as paragraphs, lists, and tables. Each container that provides an answer to the accompanying question is listed as a *long answer*. We consider this container to be relevant (system relevance). In addition, the corpus provides a *short answer* annotation, if a short phrase exists within a container that fully answers the question. The Google NQ questions are longer and more natural compared to other question answering corpora including TREC 2005 and, hence, g-REL.

For our user study, we select a subset of 12 pairs of documents and questions (plus one for training) from the NQ training data using a set of filters followed by a manual selection. Our filter removes all documents that include at least one container different than a paragraph, because we focus on continuous texts in this work. Further, it selects documents that have exactly one long and one short answer. This means that all but one paragraph per document can be considered to be irrelevant. Also, it removes all documents that have very short (less than 20 tokens) or very long (greater than 200 tokens) documents. Finally, our filter selects all documents with five to seven paragraphs, which leaves 355 of the 307k pairs for manual selection. The manual selection is guided by two factors: the average number of tokens and the position of the relevant paragraph. The remaining documents have an average length of 420.083 ($SD = 54.468$) tokens, which approximately corresponds to two times the height of the display, i.e., participants need to scroll through the document to read all paragraphs. The position of relevant paragraphs is balanced: we select two documents with an answer at position i with i ranging from 0 to 5. On average, each paragraph contains 72.55 tokens.

[2]https://ai.google.com/research/NaturalQuestions

3.3. Tasks and Procedure

In the beginning of the study, each participant is asked to sign an informed consent form and to fill in a demography questionnaire. The remainder of the study is divided in two blocks, which follow the same pattern (see **Figure 2**). In each block, stimuli from one of the two corpora are presented (within-subjects design). The starting order is alternating to avoid ordering effects. In the beginning of each block, the experimenter provides block-specific *instructions* and asks the participant to calibrate the eye tracking device. Next, the participant completes a training phase to get familiar with the task, the user interface, and with characteristics of the stimuli from the current corpus. We include two training examples for g-REL and one for Google NQ. The participant is encouraged to ask questions about the system and the task in this phase. Subsequently, the participant completes the main phase of the block, which includes 12 stimuli of the respective corpus. After both blocks are finished, participants receive the compensation payment. The task of participants is to mark all paragraphs of a document as relevant that contain an answer to the previously shown trigger question (query). First, participants read the query and, then, navigate to the corresponding document, which is either a *news article* or a *wiki article*. There is no time constraint for reading the article. Next, participants move to the rating view which enables to enter a binary relevance estimate (perceived relevance) per paragraph. At this stage, the query and the text of the paragraph are available to the participant. For stimuli from the g-REL corpus, participants have to provide one relevance estimate (there is one paragraph). For stimuli from the Google NQ corpus, participants have to provide five to seven relevance estimates (depending on the number of paragraphs).

3.4. Apparatus

The study is conducted in a separate room of our lab. We use the Tobii 4C eye tracker[3], a non-intrusive remote eye tracker, which is attached to the lower bezel of a 27-inch screen. This monitor has a resolution of 2560 × 1440 pixels and the attached

[3]https://help.tobii.com/hc/en-us/articles/213414285-Specifications-for-the-Tobii-Eye-Tracker-4C

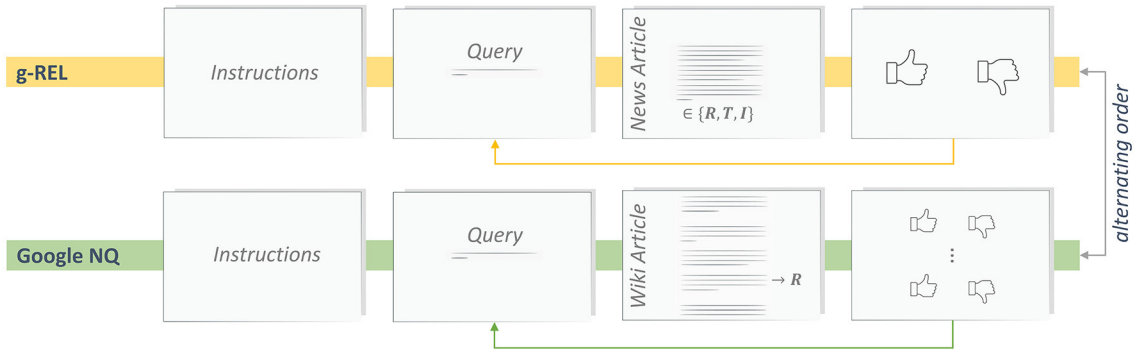


FIGURE 2 | Procedure of our user study with one block of tasks per corpus of stimuli: g-REL and Google NQ.

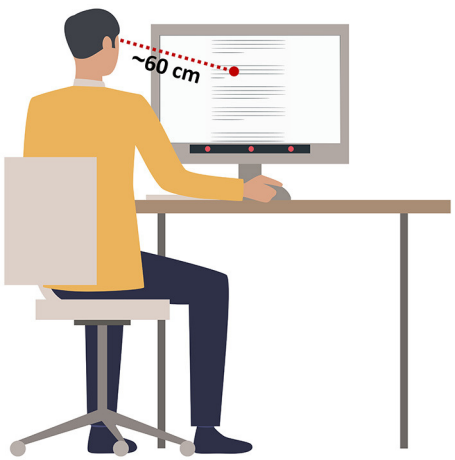


FIGURE 3 | Setup of our user study: A user is seated approximately 60 cm from a 27 inch display with the remote eye tracker mounted at its lower bezel.

eye tracker collects the gaze data with a sampling rate of 90 Hz. The monitor and eye tracker are connected to an experimenter laptop running the study software and a monitoring tool. The participants are seated approximately 60 cm in front of the connected display (see **Figure 3**). A mouse is provided to scroll through documents, to navigate between views, and to rate each paragraph for its relevancy. The text-based stimuli are displayed in black, 38-points Roboto font[4] on a white background. Before the user starts executing the tasks, we perform a calibration using the built-in 9-point calibration of the eye tracker. During the calibration process, the user is asked to look at calibration dots on the connected display until they vanish. We use the multisensor-pipeline (Barz et al., 2021), our Python-based framework for building stream processing pipelines, to implement the study software that is responsible to show the stimuli and record the interaction signals according to our experiment procedure.

3.5. gazeRE Dataset

We assembled the stimuli and the recorded interaction signals into the *gazeRE* dataset, a dataset for **gaze**-based **R**elevance **E**stimation. It includes relevance ratings (perceived relevance) from 24 participants for 12 stimuli from the g-REL corpus and 12 stimuli from the Google NQ corpus. Also, it includes participants' eye movements per document in terms of 2D gaze coordinates on the connected display. We use the gazeRE dataset for modeling the perceived relevance based on eye tracking in this work and make it publicly available under an open source license on GitHub.[5]

3.5.1. Processing of Eye Tracking Data

The gaze data included in the gazeRE dataset is preprocessed and cleaned. We correct irregular timestamps caused by transferring the gaze signal to our study software by resampling the signal with a fixed sampling rate of 83 Hz. Further, we use the gap_fill algorithm, similar to Olsen (2012), which linearly interpolates the gaze signal to close small gaps between valid gaze points, which may occur due to a loss of tracking. In addition, we use the Dispersion-Threshold Identification (I-DT) algorithm to detect fixation events (Salvucci and Goldberg, 2000).

3.5.2. Dataset Format

The gazeRE dataset includes synchronized time-series data per document and user. Each record includes a column for timestamps, gaze coordinates (x and y), a fixation ID, if the gaze point belongs to a fixations, the scroll position, and the ID of the paragraph that is hit by the current point of gaze. The origin of the gaze and fixation coordinates is the lower-left corner of the display $(0, 0)$ while $(2560, 1440)$ denotes the upper-right corner. The scroll position reflects the status of the scrollbar and lies between 0 and 1. The position is 1, if the document head is visible, or the document is not scrollable. It is 0, if the tail of the document is visible. We provide the perceived relevance per document and user: *True* is used for positive ratings, i.e., if a paragraph was perceived as relevant, *False* represents irrelevant ratings.

[4]https://fonts.google.com/specimen/Roboto (accessed February 16, 2021).

[5]https://github.com/DFKI-Interactive-Machine-Learning/gazeRE-dataset

3.5.3. Descriptive Statistics

We report descriptive statistics and agreement statistics of the relevance ratings in our dataset. We use Fleiss' κ to determine, if there was an agreement in our participants' judgement on whether paragraphs are relevant with respect to a trigger question. If the agreement among participants is low, the rating task might have been too difficult or participants might have given inadequate ratings. Further, we compute Cohen's κ to determine the level of agreement between each participant's relevance rating (perceived relevance) and the ground-truth relevance (system relevance). We report the mean agreement over all participants. We expect that the ratings of our participants moderately differ from the system relevance, similar to the findings in Bhattacharya et al. (2020a). For the g-REL corpus, we include a total of 288 trials, i.e., eye movements and a corresponding relevance estimate per paragraph (see **Figure 1**). The 12 different documents include 4 relevant paragraphs (system relevance), while one document corresponds to one paragraph. On average, the participants rated 4.46 ($SD = 1.04$) paragraphs as relevant: they perceived 107 (37%) as relevant and 181 (63%) as irrelevant. Fleiss' κ reveals a good agreement for perceived relevance ratings with $\kappa = 0.641$. The mean of Cohen's κ of 0.769 ($SD = 0.197$) indicates a substantial agreement between participant and ground-truth relevance ratings. We obtained a total of 1,680 trials using the Google NQ corpus. The 12 stimuli include 12 relevant paragraphs out of 70. On average, the participants rated 18.75 ($SD = 4.361$) paragraphs as relevant: they perceived 450 (27%) as relevant and 1,230 (73%) as irrelevant. Fleiss' κ reveals a moderate agreement for perceived relevance ratings with $\kappa = 0.576$. Also, the mean of Cohen's κ of 0.594 ($SD = 0.126$) indicates a moderate agreement between the perceived and the system relevance.

4. GAZE-BASED RELEVANCE ESTIMATION

We investigate different methods for predicting the perceived relevance of a read paragraph based on a user's eye movements. We consider the relevance prediction as a binary classification problem because each paragraph could be marked as either relevant or irrelevant in our user study. Each classification model takes a user's eye movements from reading a paragraph as input to predict the perceived relevance for this paragraph. The explicit user ratings are used as ground truth. In the following, we describe our method for extracting gaze-based features at the paragraph level, we depict our procedure for model training and evaluation, and we report the results based on the gazeRE dataset.

4.1. Extraction of Gaze-Based Features

To encode the eye movements of a user for a certain paragraph p, we have to extract coherent gaze sequences that lie within the paragraph area. A user might visit a paragraph multiple times during the relevance judgement process. We refer to these gaze sequences as visits $v_p^i \in V_p$ where i indicates the order of visits. We implement an algorithm that extracts all visits to a paragraph with a minimum length while ignoring short gaps. It identifies consecutive gaze samples that lie within the area of the given paragraph and groups them into a visit instance each. As long as there is a pair of two subsequent visits with a gap shorter than 0.2 s, these are merged. Afterwards, all visits that satisfy a minimum length of 3 s are returned as a list. We found that this duration ensures that at least 3 fixations are contained in each visit, which is required to compute the convex hull features.

We use the longest visit per paragraph v_p^* for encoding the eye movements.

To encode eye movements, we implement a set of 17 features that was successfully used to model the perceived relevance of short news articles in Bhattacharya et al. (2020a). This requires to select one visit or to merge them. We decided to use the longest visit under the assumption that the largest consecutive sequence of gaze points has the highest likelihood to capture indicative eye movements. Our feature extraction function f returns a vector of size 17 per visit: $f(v) \rightarrow \mathbb{R}^{17}$. Four of these features are based on fixation events, eight are based on saccadic movements, and five are based on the area spanned by all fixations. **Table 1** provides an overview of all features and describes how they are computed. Some features are normalized by a width factor `w` or a height factor `h`. In Bhattacharya et al. (2020a), these correspond to the display width and height, respectively. We set `w` and `h` to the width and height of the current paragraph, because the display size does not respect the different paragraph sizes and the scrolling behavior.

The absolute reading time of a visit (`scan_time`) is used to compute velocity-based or time-normalized features. The `hull_area`, i.e., the area of the convex hull around all fixations, is used to compute two area-based features.

4.2. Model Training and Evaluation

We build and compare several machine learning models that take an encoded paragraph visit v_p^* as input and yield a binary relevance estimate as output. The models are implemented using the scikit-learn machine learning framework (Pedregosa et al., 2011). Model training and testing is done using our gazeRE dataset, which includes eye movements and relevance estimates for documents from the g-REL corpus and from the Google NQ corpus. We refer to these partitions as g-REL data and Google NQ data.

4.2.1. Model Training Conditions

We largely replicate the conditions for model training and evaluation from Bhattacharya et al. (2020a) because we aim for confirming their findings: we group all visits $v \in V^*$ by their relevance rating into three subsets, train each model on 80% of the data of each subset, and evaluate it on the remaining 20% of the data. The grouping yields an *agree* subset, a *topical* subset, and the complete data denoted as *all*. **Table 2** depicts how many relevant and irrelevant samples are included in our dataset per subset. The *agree* subset includes all visits for which the perceived relevance rating agrees with the system relevance. All visits to topical articles, i.e., visits to on-topic articles that are irrelevant, are excluded as well. The *topical* subset includes visits to topical articles only, which are expected to be more difficult to classify. This subset is empty for the Google NQ corpus, because its paragraphs are marked as either relevant or irrelevant. We report

TABLE 1 | Overview of the 17 features adapted from Bhattacharya et al. (2020a) based on fixation events, saccadic eye movements, and the scanned area, which we use to encode paragraph visits.

	Feature	Description
fixation-based	`fixn_n`	Number of fixations
	`fixn_dur_sum`	Sum of fixation durations
	`fixn_dur_avg`	Mean of fixation durations
	`fixn_dur_sd`	Standard deviation of fixation durations
saccade-based	`scan_dist_h`	Sum of horizontal amplitudes of all saccades, normalized by a factor `w`
	`scan_dist_v`	Sum of vertical amplitudes of all saccades, normalized by a factor `h`
	`scan_dist_euclid`	Sum of Euclidean distances of normalized amplitudes of all saccades
	`scan_hv_ratio`	Ratio of horizontal to vertical amplitudes: `scan_dist_h`/`scan_dist_v`
	`avg_sacc_length`	Average saccade amplitude: `scan_dist_euclid`/(`fixn_n` − 1)
	`scan_speed_h`	Horizontal saccade velocity: `scan_dist_h`/`scan_time`
	`scan_speed_v`	Vertical saccade velocity: `scan_dist_v`/`scan_time`
	`scan_speed`	Saccade velocity: `scan_dist_euclid`/`scan_time`
area-based	`box_area`	Area spanned by summed saccade amplitudes: `scan_dist_h` ∗ `scan_dist_v`
	`box_area_per_time`	The `box_area` normalized by the scan time: `box_area`/`scan_time`
	`fixns_per_box_area`	Number of fixations per scanned area: `fixn_n`/`box_area`
	`hull_area_per_time`	The `hull_area` normalized by the scan time: `hull_area`/`scan_time`
	`fixns_per_hull_area`	Number of fixations per convex hull area: `fixn_n`/`hull_area`

TABLE 2 | Number of samples in our dataset per corpus and subset.

Corpus	Subset	Relevant	Irrelevant	Total
g-REL	*agree*	86 (48%)	95 (52%)	181 (63%)
	topical	20 (20%)	76 (80%)	96 (33%)
	all	**107 (37%)**	**181 (63%)**	**288 (100%)**
Google NQ	*agree*	248 (17%)	1190 (83%)	1438 (86%)
	all	**450 (27%)**	**1,230 (73%)**	**1,680 (100%)**

The topical subset includes samples for irrelevant paragraphs that are on topic of the trigger questions. The agree subset includes samples for which the participant's relevance rating matches with the system relevance and which is not in topical. Each trial corresponds to one paragraph that was either perceived as relevant or irrelevant.

the model performance metrics averaged over 10 random train-test splits to estimate the generalization performance. We use the `train_test_split()` function of scikit-learn to split the visits in a stratified fashion with prior shuffling.

4.2.2. Metrics

We include the same metrics than Bhattacharya et al. (2020a): the F1 score, i.e., the harmonic mean of precision and recall, the area under curve of the receiver operator characteristic (ROC AUC), and the balanced accuracy. In addition, we report the true positive rate (TPR) and the false positive rate (FPR), which allow us to estimate the suitability of our models for building adaptive user interfaces similar to Feit et al. (2020).

4.2.3. Model Configurations

We consider the random forest classifier of scikit-learn with default parameters (`n_estimators` = 100) as our baseline model (RF), which turned out to work well in Bhattacharya et al. (2020a). In addition, we investigate the effect of using two pre-processing steps with either a random forest classifier (RF*) or a support vector classifier (SVC*) with default parameters (`kernel` = `"rbf"`, `C` = 1) in an estimator pipeline. First, we apply the oversampling technique SMOTE Chawla et al. (2002) from the imbalanced-learn package Lemaitre et al. (2017) because visits to relevant paragraphs are underrepresented in our dataset (see **Table 2**). Second, we apply a standard feature scaling method that removes the mean and scales features to unit variance. We train separate models for g-REL data and Google NQ data.

4.2.4. Hypotheses

We hypothesize that our models can effectively estimate the perceived relevance of short news articles as shown in Bhattacharya et al. (2020a), but using our newly assembled gazeRE dataset (H1). Confirming this hypothesis would also serve as a validation of our dataset. Further, we assume that the visit-based scanpath encoding enables the prediction of a participants' perceived relevance for individual paragraphs of long Wikipedia articles. In particular, if the participant must scroll through the document to read all contents (H2).

4.3. Results

We compare the performance of three models in predicting a user's perceived relevance using our gazeRE dataset, which is based on documents of the g-REL and the Google NQ corpus. The performance scores for each model and subset are shown in **Table 3** (g-REL) and **Table 4** (Google NQ). For the g-REL data, we observe the best performance for the *agree* subset. Models trained on the *topical* subset achieve the worst results. Models for the *all* subset, which includes both other subsets, rank second. Across all subsets, the SVC* model performs best, or close to best, for most metrics. For the *topical* subset, the RF model without over-sampling and feature scaling achieves better ROC AUC and

TABLE 3 | Scores for all relevance prediction models trained and evaluated with data collected based on the g-REL corpus.

	Model	F1 Score	ROC AUC	Balanced accuracy	TPR	FPR
agree	RF	0.674	0.748	0.680	0.694	0.333
	RF*	0.677	0.747	**0.689**	0.688	**0.317**
	SVC*	**0.702**	**0.787**	0.683	**0.782**	0.417
topical	RF	0.119	**0.546**	**0.527**	0.100	**0.047**
	RF*	0.247	0.528	0.518	0.250	0.213
	SVC*	**0.270**	0.460	0.509	**0.325**	0.307
all	RF	0.458	0.650	0.594	0.405	**0.217**
	RF*	0.495	0.652	0.594	0.505	0.317
	SVC*	**0.506**	**0.652**	**0.605**	**0.510**	0.300

TABLE 4 | Scores for all relevance prediction models trained and evaluated with data collected based on the Google NQ corpus.

	Model	F1 Score	ROC AUC	Balanced accuracy	TPR	FPR
agree	RF	0.052	0.54	0.502	0.03	**0.027**
	RF*	0.246	0.543	**0.543**	0.278	0.229
	SVC*	**0.297**	**0.563**	0.54	**0.467**	0.388
all	RF	0.189	0.552	0.517	0.129	**0.095**
	RF*	0.331	0.552	0.527	0.343	0.289
	SVC*	**0.428**	**0.596**	**0.57**	**0.552**	0.412

FPR scores. However, we observe a very low TPR and F1 score in this case. For the Google NQ data, models trained on the *all* subset rank best compared to their counterpart trained on the *agree* subset. Similar to our experiment on the g-REL data, the SVC* model performs best, or close to best, for both subsets. Also, the RF model achieves the best FPR score, but the worst TPR and F1 scores.

5. DISCUSSION

The results of our machine learning experiment for short news articles (g-REL data) are similar to those in Bhattacharya et al. (2020a) (see **Table 3**). Our results indicate that we can effectively predict the perceived relevance for the *agree* subset, i.e., if the user's relevance rating agrees with the actual relevance of a paragraph and if irrelevant articles are not on topic. The *topical* trials are most difficult to classify: our models fail in differentiating between relevant and irrelevant paragraphs if they are on topic. Including *all* samples for training, our models perform better than chance with an F1 score greater than 0.5. The best-performing model pipeline, on average, is SVC*, a support vector classifier with over-sampling and feature scaling. Bhattacharya et al. (2020a) reported results for the RF model based on the original g-REL corpus using the same features for training, but with data from other participants. For the *agree* subset, their best model achieved an F1 score of 0.82, an ROC AUC of 0.92, and a balanced accuracy of 0.84. For the *topical* subset, they observed an F1 score of 0.3, an ROC AUC of 0.77, and a balanced accuracy of 0.59. Using *all* data samples results in an F1 score of 0.65, an ROC AUC of 0.85, and a balanced accuracy of 0.73. Even though we observed worse results per subset, we found the same overall pattern: the best performance is observed for models trained on the *agree* subset, followed by models for the *all* subset, and model for the *topical* subset rank last. This similarity is a good indicator for the validity of our gazeRE dataset and, eventually, it suggests that we may confirm our hypothesis H1. The differences in model performance may have several reasons. For instance, it is likely that the higher amount of training data in Bhattacharya et al. (2020a) yields better models. They used 3355 trials from 48 participants compared to 288 trials from 24 participants in our experiment. Further, our user study was conducted at a University in Germany with participants being, besides one, non-native English speakers, while the studies reported in Bhattacharya et al. (2020a) were conducted at two universities in the Unites States and predominantly included native English speakers. This may lead to a higher degree of variance in eye movements from our study. Another aspect may be that we used another eye tracking device and, hence, the data quality and pre-processing steps likely differ.

Using the Google NQ data in our machine learning experiment, we observe better scores when training on *all* data than when training on the *agree* subset only (see **Table 4**). However, the best-performing model, which is also the SVC* model, achieves F1 scores less than 0.5 in both cases although we have access to a higher number of training samples (see **Table 2**). The area under the ROC curve indicates classification performances better than chance, but we do not see enough evidence to confirm our hypothesis H2. A potential reason for the low performance might be that irrelevant paragraphs in fact belong to the same Wikipedia article than the relevant ones: the *agree* subset is rather a *topical* subset for which all user ratings agree with the system relevance. This would explain why models for the *agree* subset perform worse than models trained on *all* data. Also, the individual paragraphs in the Google NQ corpus are smaller than the ones in the g-REL corpus. This means that we aggregate less information per scanpath, which may deteriorate the model performance. Further, having multiple paragraphs allows the participants to revisit paragraphs. As we decided to encode the longest visit to a paragraph, we may miss indicative gaze patterns from another visit, which would have a negative impact on model training. In addition, the gaze estimation error inherent in eye tracking (Cerrolaza et al., 2012) may lead to a higher number of incorrect gaze-to-paragraph mappings: gaze-based interfaces should be aware of this error and incorporate it in the interaction design (Feit et al., 2017; Barz et al., 2018).

5.1. Feature Importance

We use 17 features as input to model the perceived paragraph relevance. In the following, we assess the importance of individual features to our best-performing model, the SVC* model. We use the permutation feature importance[6] method of the scikit-learn package (Pedregosa et al., 2011) to estimate feature importance, because SVCs with an rbf kernel do not

[6]https://scikit-learn.org/stable/modules/permutation_importance.html (accessed on Dec 2nd, 2021).

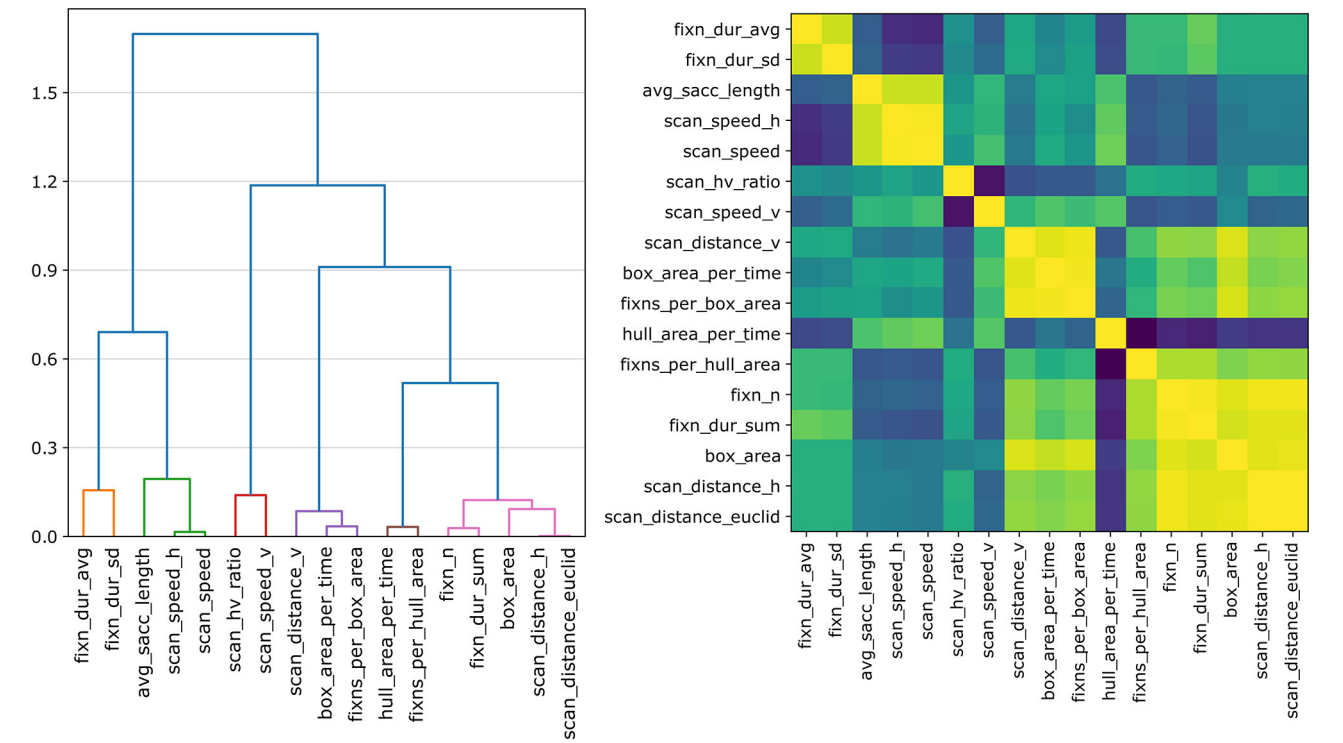


FIGURE 4 | The heatmap **(right)** shows the pairwise Spearman rank-order correlations for our features using all samples of the g-REL data. The dendrogram **(left)** shows feature groupings based on their correlation-based distances. Setting the distance threshold to $t = 0.3$ yields six feature clusters (see colored leafs).

allow direct feature analysis. This method randomly shuffles the values of one feature at a time and investigates the impact on the model performance. The loss in model performance reflects the dependency of the model on this feature. We report the mean loss in the f1 score from 30 repetitions per feature as importance measure. We analyze the feature importance for the *all* and *agree* subsets of the g-REL corpus only, because we observed f1 scores lower than 0.5 for all other conditions. The importance is reported on the training and test set of a single train-test split (80/20 split). We include both because features that are important on the training data but not on the test data might cause the model to overfit. The f1 test scores are 0.714 for the *agree* subset and 0.682 for *all* samples. However, this method might return misleading values if two features correlate. A model would still have access to nearly the same amount of information, if one feature was permuted but could be represented by another one. Hence, we perform a hierarchical clustering on the feature's Spearman rank-order correlations and use one feature per cluster to asses its importance.[7] The pairwise correlations and a grouping of our features based on correlation-based distances are visualized in **Figure 4** (*all* samples of the g-REL data). We set the distance threshold to $t = 0.3$ for the feature importance analysis for which we obtain six feature clusters as indicated by the colored leafs of the dendrogram. We obtain the same feature clusters for the *agree* subset and for both subsets of the Google NQ data. Using one feature per cluster to train and evaluate the SVC* model, we observe a drop in f1 scores of 0.015 for the *all* subset and no decline for the *agree* subset. These representative features include `fixn_dur_avg`, `scan_speed_h`, `scan_speed_v`, `scan_distance_v`, `scan_distance_h`, and `hull_area_per_time`. We remain at $t = 0.3$ because higher thresholds lead to substantially lower f1 scores and to differences in the resulting feature clusters between subsets and corpora.

The importance of feature clusters is visualized in **Figure 5**. For the *all* subset, we observe f1 losses ranging from 0.065 for `scan_speed_v` and 0.142 for `scan_distance_h` for the test set. For the train set, we observe slightly lower losses but the same importance ranking. Eventually, the features `scan_distance_h` and `hull_area_per_time` are most important when using *all* samples. For the *agree* subset, `hull_area_per_time` is by far the most important feature with an f1 loss of 0.162 on the train set and 0.137 of the test set. The features `scan_distance_v` and `fixn_dur_avg` are somewhat important with losses of 0.036 and 0.032. For `scan_speed_h`, we observe a higher importance on the train set (0.057) than on the test set (0.01), which may indicate that this feature causes the model to overfit to the training data. Overall, the `hull_area_per_time` feature introduced by

[7]We follow the scikit-learn manual for handling multicollinearity: https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance_multicollinear.html (accessed on Dec 2nd, 2021).

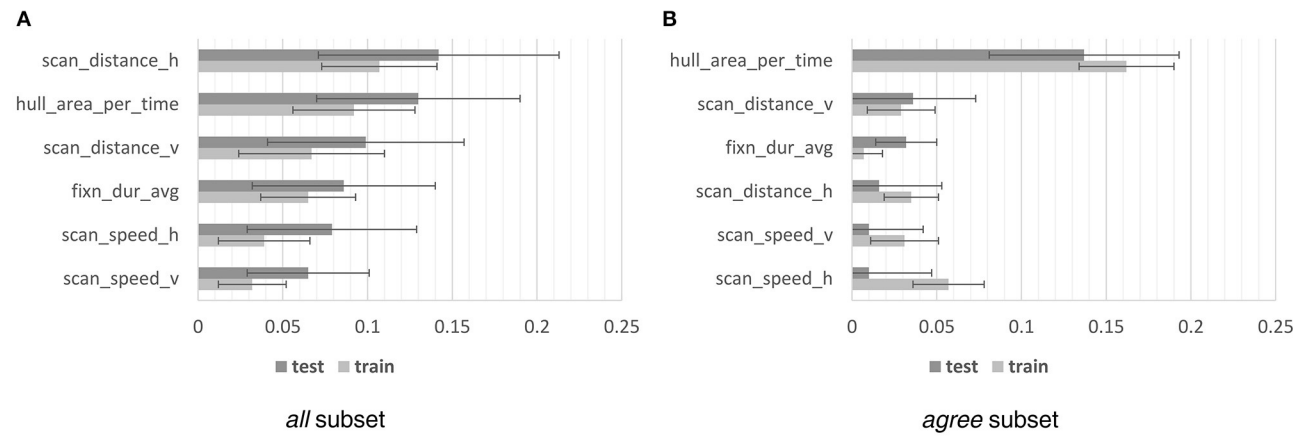


FIGURE 5 | Permutation feature importance in terms of the f1 loss for the *all* subset **(A)** and the *agree* subset **(B)** of the g-REL data. We show the mean loss in f1 scores from 30 permutation iterations ± the standard deviation. We include the feature importance estimates for the train and test split.

Bhattacharya et al. (2020a) is of high importance for modeling the perceived paragraph relevance and stable when including *topical* samples and samples for which the user rating disagrees with the ground truth. The remaining five features are important when including all samples, in particular the `scan_distance_h`. This result suggests that, in a first stage, these five features could be used to identify *topical* (irrelevant) samples and, in a second stage, the `hull_area_per_time` can predict paragraphs perceived as relevant among the remaining, non-topical samples.

5.2. Application to Adaptive User Interfaces

Our relevance estimation method can enable the development of adaptive user interfaces (UIs) that emphasize relevant contents or suppress irrelevant ones similar to Feit et al. (2020). Over time, their system detects relevant and irrelevant elements of a UI that shows different records of flat advertisements: a certain UI element always shows the same type of information, which depends on the currently viewed flat record. Our use case differs in that we want to highlight relevant text passages of a document or hide irrelevant ones. Adaptations may be based on perceived relevance estimates from recent eye movements and could, e.g., ease revisiting of relevant paragraphs in a document by immediately highlighting them or by hiding irrelevant passages. Alternatively, collecting relevant and irrelevant text passages in the pass of a search session may allow an adaptive UI to properly format text passages of documents hitherto unseen by the user. An adaptation method requires a precise recognition of relevant (true positive) or irrelevant (true negative) paragraphs to emphasize or suppress them, respectively. Misclassifications would lead to incorrect adjustments and subsequently to usability problems. Emphasizing irrelevant content (false positive) or suppressing relevant content (false negative) is likely to have a stronger negative impact on the user interaction than failing to suppress irrelevant content or to highlight a relevant one (Feit et al., 2020). To avoid strong negative impacts, adjustments by accentuation require a relevance model with a low false positive rate (FPR) and adjustments by suppression require a model with a high true positive rate (TPR), i.e., with a low number of false negatives. Depending on the type of adjustment, the TPR and FPR could be traded off against each other by using different decision thresholds. We show possible trade offs for our SVC* models using ROC curves. One model is trained on *all* g-REL data and one on *all* Google NQ data (see **Figure 6**). We do not consider other subsets for realistic application scenarios, because we would not be able to determine whether a user agreed with the actual (system) relevance of a paragraph or whether a text passage was on topic but irrelevant (*topical*). This differentiation, which is aligned to the work in Bhattacharya et al. (2020a), requires prior knowledge about the paragraphs and was meant to identify *topical* samples as being the most challenging cases for classification algorithms. Analogous to Feit et al. (2020), the shaded areas in our ROC plots in **Figure 6** indicate acceptable true and false positive rates for emphasizing or suppressing contents. For g-REL data, the ROC curve of the SVC* model hits the *emphasize* area, which indicates that it could be used to emphasize short news articles that were perceived as relevant, if the decision threshold is tuned accordingly. However, many relevant contents would be missed, as indicated by the low true positive rate (recall). Also, the shaded areas reveal that our models are not suitable for other kinds of UI adjustments.

6. CONCLUSION

In this work, we investigated whether we can confirm the findings from Bhattacharya et al. (2020a) that gaze-based features can be used to estimate the perceived relevance of short news articles read by a user. Further, we investigated whether the approach can be applied to multi-paragraph documents that require the user to scroll down to see all text passages. For this, we conducted a user study with $n = 24$ participants who read documents from two corpora, one including short news articles and one including longer Wikipedia articles in English, and rated their relevance at the paragraph-level with respect to a previously shown trigger question. We used

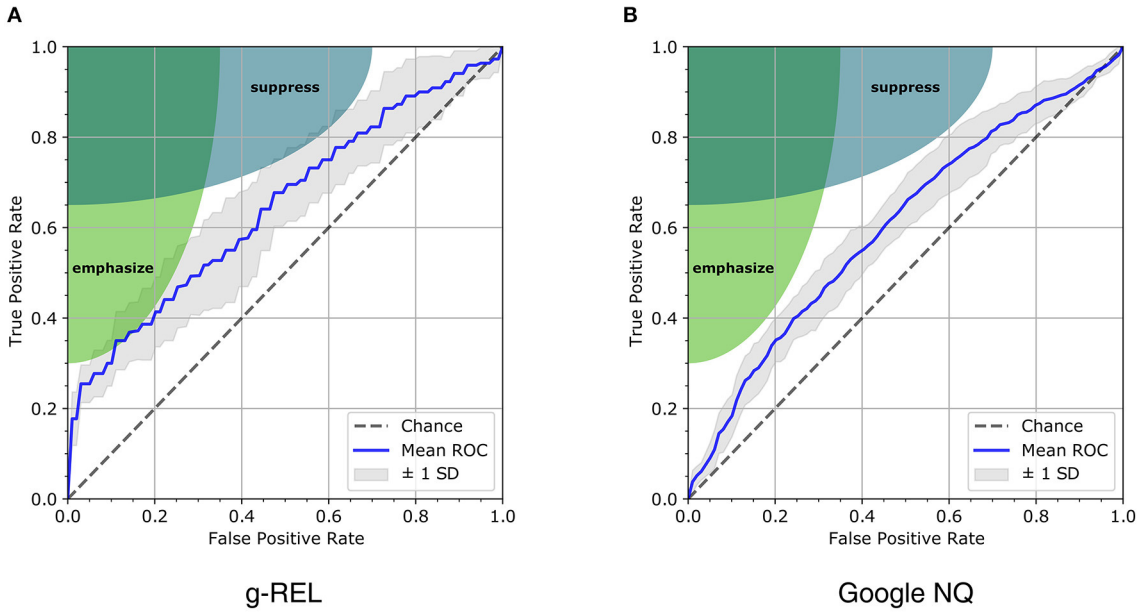


FIGURE 6 | Receiver Operator Characteristic curves for the SVC* relevance prediction models trained and tested using *all* samples from the g-REL data **(A)** and the Google NQ data **(B)**. The curves are averaged over the 10 training and test cycles; the gray area indicates the standard deviation (± 1 SD). The elliptic areas indicate acceptable true and false positive for emphasizing or suppressing contents based on Feit et al. (2020).

this data to train and evaluate machine learning models that predict the perceived relevance at the paragraph-level using the user's eye movements as input. Our results showed that, even though we achieved lower model performance scores than Bhattacharya et al. (2020a), we could replicate their findings under the same experiment conditions: eye movements are an effective source for estimating the perceived relevance of short news articles, if we leave out articles that are on topic but irrelevant. However, we could not clearly show that the approach generalizes to multi-paragraph documents. In both cases, the best model performance was observed when using over-sampling and feature scaling on the training data and a support vector classifier with an RBF kernel for classification. Future investigations should aim to overcome the limited estimation performances. A potential solution could be to use higher-level features such as the *thorough reading ratio*, i.e., the ratio of read and skimmed text lengths (Buscher et al., 2012), or the *refixation count*, i.e., the number of re-visits to a certain paragraph (Feit et al., 2020). Another solution could be found in using scanpath encodings based deep learning Castner et al. (2020); Bhattacharya et al. (2020b). We envision the gaze-based relevance detection to be a part of future adaptive UIs that leverage multiple sensors for behavioral signal processing and analysis Oviatt et al. (2018); Barz et al. (2020a,b). We published our new gazeRE dataset and our code for feature extraction under an open source license on Github to enable other researchers to replicate our approach and to implement and evaluate novel methods in the domain of gaze-based implicit relevance feedback.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/DFKI-Interactive-Machine-Learning/gazeRE-dataset.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MB, OB, and DS contributed to conception and design of the study. MB performed the statistical analysis and the machine learning experiment and wrote the first draft of the manuscript. OB conducted the study and processed the dataset and wrote sections of the manuscript. MB and DS acquired the funding for this research. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was funded by the German Federal Ministry of Education and Research (BMBF) under grant number 01JD1811C (GeAR) and in the Software Campus project SciBot.

REFERENCES

Agichtein, E., Brill, E., and Dumais, S. (2006). "Improving web search ranking by incorporating user behavior information," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06* (New York, NY: Association for Computing Machinery), 19–26. doi: 10.1145/1148170.1148177

Ajanki, A., Hardoon, D. R., Kaski, S., Puolamaki, K., and Shawe-Taylor, J. (2009). Can eyes reveal interest? Implicit queries from gaze patterns. *User Model. User Adapt. Interact.* 19, 307–339. doi: 10.1007/s11257-009-9066-4

Akuma, S., Iqbal, R., Jayne, C., and Doctor, F. (2016). Comparative analysis of relevance feedback methods based on two user studies. *Comput. Hum. Behav.* 60, 138–146. doi: 10.1016/j.chb.2016.02.064

Barz, M., Altmeyer, K., Malone, S., Lauer, L., and Sonntag, D. (2020a). "Digital pen features predict task difficulty and user performance of cognitive tests," in *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2020* (Genoa: ACM), 23–32. doi: 10.1145/3340631.3394839

Barz, M., Bhatti, O. S., Laers, B., Prange, A., and Sonntag, D. (2021). "Multisensor-pipeline: a lightweight, flexible, and extensible framework for building multimodal-multisensor interfaces," in *Companion Publication of the 2021 International Conference on Multimodal Interaction, ICMI '21 Companion* (Montreal, QC: ACM).

Barz, M., Daiber, F., Sonntag, D., and Bulling, A. (2018). "Error-aware gaze-based interfaces for robust mobile gaze interaction," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, ETRA 2018*, eds B. Sharif and K. Krejtz (Warsaw: ACM), 24:1–24:10. doi: 10.1145/3204493.3204536

Barz, M., Stauden, S., and Sonntag, D. (2020b). "Visual search target inference in natural interaction settings with machine learning," in *ACM Symposium on Eye Tracking Research and Applications, ETRA '20*, eds A. Bulling, A. Huckauf, E. Jain, R. Radach, and D. Weiskopf (Stuttgart: Association for Computing Machinery), 1–8. doi: 10.1145/3379155.3391314

Bhattacharya, N., and Gwizdka, J. (2018). "Relating eye-tracking measures with changes in knowledge on search tasks," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, ETRA '18* (New York, NY: Association for Computing Machinery). doi: 10.1145/3204493.3204579

Bhattacharya, N., Rakshit, S., and Gwizdka, J. (2020a). "Towards real-time webpage relevance prediction using convex hull based eye-tracking features," in *ACM Symposium on Eye Tracking Research and Applications, ETRA '20 Adjunct* (New York, NY: Association for Computing Machinery). doi: 10.1145/3379157.3391302

Bhattacharya, N., Rakshit, S., Gwizdka, J., and Kogut, P. (2020b). "Relevance prediction from eye-movements using semi-interpretable convolutional neural networks," in *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval, CHIIR '20* (New York, NY: Association for Computing Machinery), 223–233. doi: 10.1145/3343413.3377960

Buchanan, G., McKay, D., Velloso, E., Moffat, A., Turpin, A., and Scholer, F. (2017). "Only forward? Toward understanding human visual behaviour when examining search results," in *Proceedings of the 29th Australian Conference on Computer-Human Interaction, OZCHI '17* (New York, NY: Association for Computing Machinery), 497–502. doi: 10.1145/3152771.3156165

Buscher, G., Dengel, A., Biedert, R., and Elst, L. V. (2012). "Attentive documents: eye tracking as implicit feedback for information retrieval and beyond," in *ACM Transactions on Interactive Intelligent Systems* (New York, NY: Association for Computing Machinery). doi: 10.1145/2070719.2070722

Buscher, G., Dengel, A., and van Elst, L. (2008a). "Eye movements as implicit relevance feedback," in *CHI '08 Extended Abstracts on Human Factors in Computing Systems, CHI EA '08* (New York, NY: Association for Computing Machinery), 2991–2996. doi: 10.1145/1358628.1358796

Buscher, G., Dengel, A., and van Elst, L. (2008b). "Query expansion using gaze-based feedback on the subdocument level," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08* (New York, NY: Association for Computing Machinery), 387–394. doi: 10.1145/1390334.1390401

Buscher, G., van Elst, L., and Dengel, A. (2009). "Segment-level display time as implicit feedback: a comparison to eye tracking," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09* (New York, NY: Association for Computing Machinery), 67–74. doi: 10.1145/1571941.1571955

Carpineto, C., and Romano, G. (2012). "A survey of automatic query expansion in information retrieval," in *ACM Computing Surveys* (New York, NY: Association for Computing Machinery). doi: 10.1145/2071389.2071390

Castner, N., Kuebler, T. C., Scheiter, K., Richter, J., Eder, T., Huettig, F., et al. (2020). "Deep semantic gaze embedding and scanpath comparison for expertise classification during OPT viewing," in *ACM Symposium on Eye Tracking Research and Applications, ETRA '20* (New York, NY: Association for Computing Machinery). doi: 10.1145/3379155.3391320

Cerrolaza, J. J., Villanueva, A., Villanueva, M., and Cabeza, R. (2012). "Error characterization and compensation in eye tracking systems," in *Proceedings of the Symposium on Eye Tracking Research and Applications, ETRA '12* (New York, NY: Association for Computing Machinery), 205–208. doi: 10.1145/2168556.2168595

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Chen, Y., Zhang, P., Song, D., and Wang, B. (2015). "A real-time eye tracking based query expansion approach via latent topic modeling," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15* (New York, NY: Association for Computing Machinery), 1719–1722. doi: 10.1145/2806416.2806602

Cole, M. J., Gwizdka, J., Liu, C., Belkin, N. J., and Zhang, X. (2013). Inferring user knowledge level from eye movement patterns. *Inform. Process. Manage.* 49, 1075–1091. doi: 10.1016/j.ipm.2012.08.004

Davari, M., Hienert, D., Kern, D., and Dietze, S. (2020). "The role of word-eye-fixations for query term prediction," in *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval, CHIIR '20*, (New York, NY: Association for Computing Machinery), 422–426. doi: 10.1145/3343413.3378010

Eickhoff, C., Dungs, S., and Tran, V. (2015). "An eye-tracking study of query reformulation," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '15* (New York, NY: Association for Computing Machinery), 13–22. doi: 10.1145/2766462.2767703

Feit, A. M., Vordemann, L., Park, S., Berube, C., and Hilliges, O. (2020). "Detecting relevance during decision-making from eye movements for UI adaptation," in *ACM Symposium on Eye Tracking Research and Applications, ETRA '20* (New York, NY: Association for Computing Machinery), 13–22. doi: 10.1145/3379155.3391321

Feit, A. M., Williams, S., Toledo, A., Paradiso, A., Kulkarni, H., Kane, S., et al. (2017). "Toward everyday gaze input: accuracy and precision of eye tracking and implications for design," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery), 1118–1130. doi: 10.1145/3025453.3025599

Graff, D. (2002). *The AQUAINT Corpus of English News Text LDC2002T31*. Philadelphia, PA.

Gwizdka, J. (2014a). "Characterizing relevance with eye-tracking measures," in *Proceedings of the 5th Information Interaction in Context Symposium* (New York, NY: Association for Computing Machinery), 58–67. doi: 10.1145/2637002.2637011

Gwizdka, J. (2014b). "News stories relevance effects on eye-movements," in *Proceedings of the Symposium on Eye Tracking Research and Applications, ETRA '14* (New York, NY: Association for Computing Machinery), 283–286. doi: 10.1145/2578153.2578198

Gwizdka, J. (2017). "Differences in reading between word search and information relevance decisions: evidence from eye-tracking," in *Information Systems and Neuroscience*, eds F. D. Davis, R. Riedl, J. vom Brocke, P. M. Lager, and A. B. Randolph (Cham. Springer International Publishing), 141–147. doi: 10.1007/978-3-319-41402-7_18

Gwizdka, J., and Dillon, A. (2020). "Eye-tracking as a method for enhancing research on information search," in *Understanding and Improving Information Search: A Cognitive Approach*, eds W. T. Fu and H. van Oostendorp (Cham: Springer International Publishing), 161–181. doi: 10.1007/978-3-030-38825-6_9

Hardoon, D. R., Shawe-Taylor, J., Ajanki, A., Puolamaki, K., and Kaski, S. (2007). "Information retrieval by inferring implicit queries from eye movements," in

Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, eds M. Meila and X. Shen (San Juan; Puerto Rico: PMLR), 179–186.

Hienert, D., Kern, D., Mitsui, M., Shah, C., and Belkin, N. J. (2019). "Reading protocol: understanding what has been read in interactive information retrieval tasks," in *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, CHIIR '19* (New York, NY: Association for Computing Machinery), 73–81. doi: 10.1145/3295750.3298921

Jacob, S., Ishimaru, S., Bukhari, S. S., and Dengel, A. (2018). "Gaze-based interest detection on newspaper articles," in *Proceedings of the 7th Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction, PETMEI '18* (New York, NY: Association for Computing Machinery). doi: 10.1145/3208031.3208034

Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. (2017). *Accurately Interpreting Clickthrough Data as Implicit Feedback*. New York, NY: Association for Computing Machinery. doi: 10.1145/3130332.3130334

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., et al. (2019). Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguist.* 7, 453–466. doi: 10.1162/tacl_a_00276

Lemaitre, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* 18, 1–5. Available online at: http://jmlr.org/papers/v18/16-365.html

Li, X., Liu, Y., Mao, J., He, Z., Zhang, M., and Ma, S. (2018). "Understanding reading attention distribution during relevance judgement," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18* (New York, NY: Association for Computing Machinery), 733–742. doi: 10.1145/3269206.3271764

Loboda, T. D., Brusilovsky, P., and Brunstein, J. (2011). "Inferring word relevance from eye-movements of readers," in *Proceedings of the 16th International Conference on Intelligent User Interfaces, IUI '11* (New York, NY: Association for Computing Machinery), 175–184. doi: 10.1145/1943403.1943431

Miller, T., and Agne, S. (2005). "Attention-based information retrieval using eye tracker data," in *Proceedings of the 3rd International Conference on Knowledge Capture, K-CAP '05* (New York, NY: Association for Computing Machinery), 209–210. doi: 10.1145/1088622.1088672

Olsen, A. (2012). *The Tobii I-VT Fixation Filter: Algorithm Description*, Danderyd: Tobii Technology.

Oviatt, S., Schller, B., Cohen, P. R., Sonntag, D., Potamianos, G., and Kruger, A. (eds.). (2018). *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition*, volume 2. New York, NY: Association for Computing Machinery.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Salojarvi, J., Kojo, I., Simola, J., and Kaski, S. (2003). "Can relevance be inferred from eye movements in information retrieval?" in *Workshop on Self-Organizing Maps (WSOM'03)* (Hibikino), 261–266.

Salojarvi, J., Puolamaki, K., and Kaski, S. (2004). "Relevance feedback from eye movements for proactive information retrieval," in *Workshop on Processing Sensory Information for Proactive Systems (PSIPS 2004)* (Oulu), 14–15.

Salojarvi, J., Puolamaki, K., and Kaski, S. (2005a). "Implicit relevance feedback from eye movements," in *Artificial Neural Networks: Biological Inspirations – ICANN 2005* (Berlin; Heidelberg), 513–518. doi: 10.1007/11550822_80

Salojarvi, J., Puolamaki, K., Simola, J., Kovanen, L., Kojo, I., and Kaski, S. (2005b). *Inferring Relevance from Eye Movements: Feature Extraction*. Helsinki University of Technology.

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

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Computer Science

ORIGINAL RESEARCH
published: 14 January 2022
doi: 10.3389/fcomp.2021.775282

Relevant Physiological Indicators for Assessing Workload in Conditionally Automated Driving, Through Three-Class Classification and Regression

Quentin Meteier[1]*, *Emmanuel De Salis*[2], *Marine Capallera*[1], *Marino Widmer*[3], *Leonardo Angelini*[1], *Omar Abou Khaled*[1], *Andreas Sonderegger*[4] and *Elena Mugellini*[1]

[1] HumanTech Institute, University of Applied Sciences of Western Switzerland, HES-SO, Fribourg, Switzerland, [2] He-Arc, University of Applied Sciences of Western Switzerland, HES-SO, Saint-Imier, Switzerland, [3] Department of Informatics, University of Fribourg, Fribourg, Switzerland, [4] Business School, Institute for New Work, Bern University of Applied Sciences, Bern, Switzerland

OPEN ACCESS

Edited by:
Youngjun Cho,
University College London,
United Kingdom

Reviewed by:
Jun Shen,
University of Wollongong, Australia
Jerry Chu-Wei Lin,
Western Norway University of Applied Sciences, Norway

***Correspondence:**
Quentin Meteier
quentin.meteier@hes-so.ch

Specialty section:
This article was submitted to Mobile and Ubiquitous Computing, a section of the journal Frontiers in Computer Science

Received: 13 September 2021
Accepted: 13 December 2021
Published: 14 January 2022

Citation:
Meteier Q, De Salis E, Capallera M, Widmer M, Angelini L, Abou Khaled O, Sonderegger A and Mugellini E (2022) Relevant Physiological Indicators for Assessing Workload in Conditionally Automated Driving, Through Three-Class Classification and Regression. Front. Comput. Sci. 3:775282. doi: 10.3389/fcomp.2021.775282

In future conditionally automated driving, drivers may be asked to take over control of the car while it is driving autonomously. Performing a non-driving-related task could degrade their takeover performance, which could be detected by continuous assessment of drivers' mental load. In this regard, three physiological signals from 80 subjects were collected during 1 h of conditionally automated driving in a simulator. Participants were asked to perform a non-driving cognitive task (N-back) for 90 s, 15 times during driving. The modality and difficulty of the task were experimentally manipulated. The experiment yielded a dataset of drivers' physiological indicators during the task sequences, which was used to predict drivers' workload. This was done by classifying task difficulty (three classes) and regressing participants' reported level of subjective workload after each task (on a 0–20 scale). Classification of task modality was also studied. For each task, the effect of sensor fusion and task performance were studied. The implemented pipeline consisted of a repeated cross validation approach with grid search applied to three machine learning algorithms. The results showed that three different levels of mental load could be classified with a f1-score of 0.713 using the skin conductance and respiration signals as inputs of a random forest classifier. The best regression model predicted the subjective level of workload with a mean absolute error of 3.195 using the three signals. The accuracy of the model increased with participants' task performance. However, classification of task modality (visual or auditory) was not successful. Some physiological indicators such as estimates of respiratory sinus arrhythmia, respiratory amplitude, and temporal indices of heart rate variability were found to be relevant measures of mental workload. Their use should be preferred for ongoing assessment of driver workload in automated driving.

Keywords: automated driving, classification, driver, indicators, physiology, regression, workload, non-driving related task

1. INTRODUCTION

A recent study of critical reasons for traffic crashes found that the driver was at fault in 94% of the cases (Singh, 2015). It includes recognition errors (including driver inattention and distractions), decision errors (driving too fast, misjudging the gap), performance errors, and non-performance errors (such as sleeping). To address this issue, car manufacturers are automating several functions of the driving task to assist the driver. In 2021, the last cars sold on the market are defined as partially automated vehicles and classified as Level 2 in the Society of Automotive Engineers (SAE) taxonomy (Society of Automotive Engineers, 2018). These vehicles automate certain functions such as maintaining speed, keeping distance from the car in front, or keeping the vehicle in the lane laterally. However, automotive manufacturers are already preparing for the next step by developing conditionally automated cars (Level 3), but also highly and fully automated cars (Levels 4 and 5) (Society of Automotive Engineers, 2018). At higher levels of automation, the car will be responsible for performing the dynamic driving task and monitoring the driving environment. It frees drivers from the primary task of driving and allows them to engage in a non-driving related task (NDRT). However, performing a NDRT may distract them and increase their mental workload (MWL; Mehler et al., 2009). Previous research has shown that an underloaded or overloaded state impacts the performance of a user interacting with automation (Wickens et al., 2014). The increase in automation in cars should therefore prompt solutions to intelligently and non-intrusively measure the mental load of drivers. The use of machine learning techniques coupled with the increasing amount of available data allows the development of intelligent models that can accurately predict the level of workload (Mehler et al., 2009). Depending on the level of driver workload, the driver-vehicle interaction must be continuously adapted to ensure safe use of the automation and improve the user experience.

2. RELATED WORK

2.1. Definition of Mental Workload

The tasks performed by drivers will change as cars increase in automation. Some secondary tasks may lead to an increase in MWL, which needs to be evaluated in this context. MWL is defined as a balance between the exigencies of a situation and the resources available to the operator to deal with that situation. (Wickens, 2008). Multiple dimensions play a role in this complex construct such as operator characteristics (skills and attentional resources), task characteristics difficulty and modality) and environmental context (Young et al., 2015).

In the driving context, MWL is of great importance because a suboptimal level of MWL (mental underload or overload) can lead the driver to errors in attention, which can result in accidents (Brookhuis and De Waard, 2001). Three categories of measures are effective for assessing MWL: task performance measures (primary and secondary task), subjective questionnaire-based assessments and psychophysiological measures (Paxion et al., 2014; Gawron, 2019).

The primary-secondary task paradigm has proven to be a good indicator of MWL in experimental research, specifically in the context of driving (Engstrm et al., 2005; Mehler et al., 2009). In general, the assessment of task performance is done on the primary task (dynamic driving task) and the secondary task (NDRT). An acceptable level of performance can be maintained in the primary task under high workload conditions. It is typically measured by longitudinal (speed and distance from the car in front) and lateral (direction and position in the lane) parameters computed from driving data collected in simulators or road experiments (Engstrm et al., 2005; Mehler et al., 2009). The secondary task performance is highly correlated with MWL since it is associated with a spare capacity not used for completion of the primary task (Young et al., 2015). Thus, secondary task performance (e.g., NDRT) is an indicator of MWL in the context of driving (Engstrm et al., 2005; Mehler et al., 2009). However, measuring MWL by task performance presents some downsides, including control of the task scenarios, monitoring of task performance and artificial configuration of the test environment (Fisk et al., 1986).

Operators' can also report the perceived MWL with subjective ratings. There are several standardized questionnaires for subjectively measuring MWL such as the NASA Task Load Index (NASA-TLX; Hart and Staveland, 1988), the Subjective Workload Assessment Technique (SWAT; Reid and Nygren, 1988) or the Workload Profile (WP; Tsang and Velazquez, 1996). Two other questionnaires can evaluate, respectively, the mental effort and the mental workload generated by the dynamic driving task : the Rating Scale Mental Effort (RSME; Zijlstra and Doorn, 1985) and the Driving Activity Load Index (DALI; Pauzié, 2008). These questionnaires are easy to apply and implement (Rubio et al., 2004) but present some methodological drawbacks. The subjective nature of the measure, as well as the recall bias due to post-task assessment can lead to a discrepancy between the subjective report and the actual level of MWL (Bulmer et al., 2004; Paxion et al., 2014). In addition, a subjective post-task assessment of the MWL does not capture the MWL variation during the task, which could be of great interest (Paxion et al., 2014).

Another approach to measure MWL is the use of psychophysiological indicators. It includes indicators of the central and autonomic nervous system s, such as measures of cardiac activity (heart rate and heart rate variability), electrodermal activity (tonic and phasic skin conductivity), and brain activity through electroencephalography (EEG). Previous research showed that they are reliable indicators of MWL (De Waard, 1997; Dornhege et al., 2007; Haapalainen et al., 2010; Ferreira et al., 2014; Hogervorst et al., 2014; Paxion et al., 2014). Recently, near-infrared spectroscopy (NIRS) has shown great potential as source of data for evaluating driver's MWL (Le et al., 2018). However, EEG and NIRS might not be used in real-world driving conditions, as many drivers may be reluctant to wear a headset while driving. There are some disadvantages to assessing MWL using physiological indicators, such as tedious and delicate placement of electrodes on the user's body, noise in the signal and the spurious influence of physical activity (Huigen et al., 2002). Recent advances in smart wearable devices and clothing (Baek

et al., 2009) may help democratize the use of physiological signals to measure MWL in real-world driving conditions. Physiological signals could thus be collected in a continuous, non-intrusive manner to provide a robust assessment of driver's MWL.

2.2. Assessment of MWL Through Physiological Indicators

2.2.1. Relevant Physiological Indicators of MWL

Similarly, as indicators of Electrodermal activity (EDA) (Boucsein, 2012), indices of cardiac activity computed from an electrocardiogram (ECG), such as heart rate (HR) and heart rate variability (HRV), are widely used to assess changes in the autonomic nervous system. Previous research has shown that EDA and HRV indicators are sensitive to increases in MWL (Brookhuis et al., 2004; Engstrm et al., 2005; Collet et al., 2009; Mehler et al., 2009, 2012; Brookhuis and de Waard, 2010). Indicators can be temporal measures (SDNN, RMSSD..), or frequency measures such as the ratio of power in the low and high frequency bands of the HRV (Malik and Terrace, 1996). Recent studies have shown that 10–60 s may be sufficient to obtain reliable time-based measurements of HRV, whereas 20–90 s may be sufficient to capture changes in the autonomic nervous system using frequency-based measures (Salahuddin et al., 2007; Baek et al., 2015). Besides, the respiratory system can influence both EDA and cardiac activity. The close coupling of ECG and respiration (RESP) signals is no longer in question (Cacioppo et al., 2007). This phenomenon is referred to as Respiratory Sinus Arrhythmia (RSA) and describes how the respiratory pattern modulates the heart rate (Hirsch and Bishop, 1981). Several methods can be used to quantify this phenomenon, but its assessment by the Porges-Bohrer method may be the most appropriate measure of RSA according to Lewis et al. (2012).

2.2.2. Effect of Task Difficulty and Modality

Task difficulty has been shown to have an effect on mental workload measured by physiological indicators. Whether in a simulation environment or a real-world driving environment, MWL has been shown to increase with task difficulty (Engstrm et al., 2005; Mehler et al., 2009, 2012). Physiological indicators that were found to be sensitive to increased workload were mean skin conductance level (Engstrm et al., 2005; Mehler et al., 2009, 2012), heart rate (Collet et al., 2009; Mehler et al., 2009), some HRV indicators such as beat-to-beat intervals (Engstrm et al., 2005) or frequency-based measures (Brookhuis et al., 2004; Brookhuis and de Waard, 2010), and respiratory rate (Mehler et al., 2009). An increase in MWL is accompanied by an increase in heart rate, skin conductance, and respiratory rate (Mehler et al., 2009, 2012). Among these previous studies, only a non-significant effect was found for the task difficulty on skin conductance during an auditory task in the work of Engstrm et al. (2005). This could be due to low driver engagement in the non-driving task, as suggested later by Mehler et al. (2012). Therefore, task performance should be carefully recorded if the workload is measured using physiological indicators. This ensures that the participants are engaged in the non-driving-related task, and possibly uses performance as a control variable in statistical analysis. The effect of task modality on workload was not analyzed. Yet, results of increased workload due to task difficulty have been shown using different tasks involving various modalities such as visual (Engstrm et al., 2005), auditory (Engstrm et al., 2005; Collet et al., 2009; Mehler et al., 2009, 2012) or verbal (Engstrm et al., 2005; Collet et al., 2009; Mehler et al., 2009, 2012) tasks. In other words, regardless of task modality, the same increase in workload is observed as task difficulty increases, based on different physiological measures. This suggests that it might be more difficult to predict task modality with this source of data. This hypothesis will be tested in this work.

2.3. Workload Evaluation Using Physiological Signals and Machine Learning

One of the objectives of this paper is to predict drivers' MWL using physiological indicators and artificial intelligence (AI) techniques. Previous studies that predicted subjects' MWL using physiological signals and machine learning were reviewed. Only studies that used at least 2 signals among ECG, EDA, and RESP as inputs of machine learning models were reviewed. The studies considered are presented in **Table 1**. They are compared and discussed on several parameters that can affect the accuracy of a model trained with machine learning techniques, including the environmental settings, the task used to induce MWL, the time intervals used for calculating physiological indicators, the number of classes, and the evaluation approach. Previous studies were conducted in different environments, such as laboratories (Haapalainen et al., 2010; Ferreira et al., 2014; Hogervorst et al., 2014), driving simulators (Son et al., 2013; Darzi et al., 2018; Meteier et al., 2021) or on roads (Solovey et al., 2014). For the driving studies, participants were required to drive manually and perform an additional NDRT to manipulate the level of MWL, except for Meteier et al. (2021) study in which the car drove in conditional automation, and participants were required to count backward orally. Different cognitive tasks were used to manipulate MWL, such as the Pursuit test, the Scattered X (Ferreira et al., 2014), or the N-back task. The latter can involve visual resources with letters displayed on a screen (Hogervorst et al., 2014) or auditory and verbal when the letters are auditory stimuli and participants have to respond verbally (Son et al., 2013; Solovey et al., 2014). Also, the difficulty of the task has an impact on the workload, suggesting that the task used to manipulate the MWL experimentally should be chosen carefully (Mehler et al., 2009, 2012).

The time window used to calculate features can also influence the models' performance in time-series classification tasks. The length of time windows differed between studies, ranging from 30 to 240 s. Solovey et al. (2014) and Meteier et al. (2021) investigated the influence of time window length on model accuracy. For windows shorter than 30 s, Solovey et al. (2014) showed that model accuracy increases with time window size. For longer time windows (30 s–20 min), Meteier et al. (2021) showed that model accuracy increases up to a size of 4 min but decreases if it is longer.

As shown in **Table 1**, previous studies only classified the user's MWL at two levels. Model performance were evaluated

TABLE 1 | State of the art of previous similar studies.

Reference	Only physio	Study	Task	Time window	Classes	Evaluation	Perf. (%)
Haapalainen et al. (2010)	Yes, with EEG	In lab, on a computer	6 tasks, testing speed of closure, flexibility of closure and perceptual speed	43 s (easy task), 106 s (hard task)	2	Within-subject	83.7
Son et al. (2013)	Yes	Driving simulator : Manual driving on a highway	Auditory N-Back task	30 s	2	Between-subject	82.9
Ferreira et al. (2014)	Yes, with EEG	In lab, on a computer	2 tasks: testing perceptual speed (Pursuit Test) and visio-spatial capacities (Scattered X)	60 s	2	Within-subject	86.0
Hogervorst et al. (2014)	Yes	In lab, on a computer	Visual N-Back task	120 s	2	Within-subject	75.0
Solovey et al. (2014)	Yes	Manual driving on a highway	Auditory stimuli verbal prompt N-back	30 s (sliding)	2	Within-subject	75.7
	Yes					Between-subject	90.0
Darzi et al. (2018)	Yes	Moving-base driving simulator : manual driving	Cell phone use	240 s	2	Between-subject	82.3
Meteier et al. (2021)	Yes	Driving simulator : Conditionally automated driving	Oral backwards counting	240 s	2	Between-subject	95.0

Perf. column is the best score achieved in the study, using mean accuracy as metric.

using the mean accuracy as a metric. Accuracy scores range from 75 to 95%, either using between-subject or within-subject evaluation. A three-level workload classification was done with EEG signals (Plechawska-Wojcik et al., 2019), but not using only physiological signals.

Complex and recent approaches of time series classification can be used in order to classify continuously the user's state (Bagnall et al., 2016). The recent emergence of deep learning offers new possibilities to build even more efficient models for time series classification (Ismail Fawaz et al., 2019). The ResNet model (He et al., 2016) showed to outperform other models on different categories of datasets, but not on ECG datasets (Ismail Fawaz et al., 2019). A fully convolutional network (FCN) might be a best option for classification with physiological signals (Wang et al., 2017; Ismail Fawaz et al., 2019). However, these types of deep architectures require to have a large dataset to achieve good accuracy. Other recent models such as XGBoost are also efficient for predicting cognitive workload with physiological signals (Momeni et al., 2019).

3. PRESENT STUDY

The present study aims to classify drivers' MWL at three different levels (low vs. medium vs. high) based on physiological indicators. These different levels of MWL are induced by NDRTs performed by the drivers during conditionally automated driving. To obtain a more refined assessment of MWL, post-task subjective reports are used to regress drivers' MWL (on a 0–20 scale). Task modality is also classified at two levels (visual vs. auditory task). For these classification and regression tasks, the effect of sensor fusion and task performance are investigated, because some drivers might disengage from the tasks (mental fatigue or task too difficult) and thus result in lower physiological activation (Mehler et al., 2012).

The main novelty of this work is to perform a finer evaluation of drivers' MWL than in previous studies, by doing three-class classification and regression tasks only with physiological signals. This work uses ECG, EDA, and RESP for assessing drivers' workload as EEG or NIRS may be considered less suitable for real-world condition. Also, the effect of drivers' task performance on models' accuracy has not been done in previous research. Finally, using a data-driven approach with an explainable AI (xAI) technique to find the most relevant indicators of MWL has not been done so far. To summarize, the following are the contributions made in this manuscript:

- Statistical analysis of the effect of task difficulty, modality, measurement time and interaction of them on three physiological measures (one for each signal).
- Analysis of task performance and sensor fusion on the performance of classification and regression models to predict MWL.
- Use of an xAI approach to find the most relevant indicators of MWL in the context of conditionally automated driving.

Drivers' MWL prediction is done in the specific context of automated driving, while most of previous studies focused on assessing MWL in manual driving scenarios. Only one recent study focused on the evaluation of MWL in conditionally automated driving (Meteier et al., 2021), but authors used a verbal task to induce MWL and suggested that it might have induced a bias in the classification of the driver's state. For this reason, the manipulation of drivers' MWL was done at three different levels, with participants performing a succession of short non-verbal tasks (90 s each). Previous research showed that indicators of skin conductance and heart rate variability are reliable measures of MWL (Engstrm et al., 2005; Collet et al., 2009; Mehler et al., 2009, 2012), so we expect to see higher performance when EDA and ECG signals are used to train the models.

4. MATERIALS AND METHODS

4.1. Experimental Method

4.1.1. Participants and Experimental Design

For this study, 80 participants were recruited. 67.5% consider themselves as female ($N = 54$) and 32,5% as male ($N = 26$). The sample of drivers was rather young ($M = 23,9$ years old, SD $= 8.2$), ranging from 19 to 66 years old. They reported holding their driving license for 5.42 years (SD $= 8.08$ years) and driving 6312 kilometers per year on average (SD $= 14$ 415 km). 76.3% of participants did not have an accident in the last 3 years and 36% indicated that they have already used an automated car. 25% of them reported that they drove in a simulator before. Most of the participants were students at the university. They were recruited by e-mail and advertising flyers. The participants needed a driving license and adequate knowledge of German, French, or Italian to participate in the study. Thirty-eight were German native speakers, 18 were French native speakers, 21 were Italian native speakers, and 2 had another mother tongue. As compensation for participating in the experiment, the participants received 2 experimental hours counting for their study program. Before taking part in the study, all participants were informed in detail about the automated driving systems, the purpose of the study and the procedure. They agreed to our consent form based on the ethics committee of the university and the federal law on data protection. Participants were randomly assigned to the experimental groups.

The study consisted of an experimental mixed design with four independent variables. Two of were within-subject variables: the task difficulty (low vs. medium vs. high cognitive task) and the task modality (no task vs. auditory vs. visual task). To manipulate these two factors, the N-back task was chosen (Kirchner, 1958). It is a continuous performance task that has been widely used in research as a tool to induce various levels of MWL to participants, through different modalities (either visual or auditory). "N" is the factor that can be varied to make the task more or less difficult. The participant has to press a button if the current letter is the same as the one presented N-steps before, as shown in **Figure 1**. In this study, the 1-back and 3-back tasks, respectively, correspond to the condition of the medium and high cognitive tasks. For the task modality, the sequence was either presented visually on a screen or played through audio files. Both modalities were done on the same tablet. Audio files were recorded before the experiment and played in the participant's native language. Additionally, a control variable was used and common to both variables. It is a condition in which participants did not perform the N-back task. During these periods, they were only asked to monitor the driving environment while the car was driving in conditional automation. The order of the non-driving related task sequences was randomized throughout the experiment but controlled before the takeover situations by following a Latin Square design (Kirk, 2013).

There were two other between-subject factors in the experimental design: the information on automated cars limitations before the experiment (information vs. no information) and the presence of a mobile application giving context-related information of the driving situation on the tablet (application vs. no application). Also, participants had to react to five different takeover situations. The effect of these two between-subject factors and takeover situations are not presented in this work, see the work of Meteier et al. (2020) for more details.

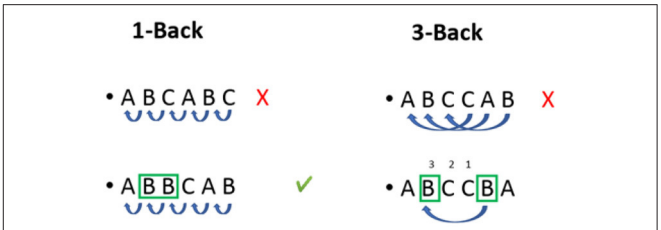


FIGURE 1 | Illustration of the N-back task operation according to the difficulty modality (1-back vs. 3-back).

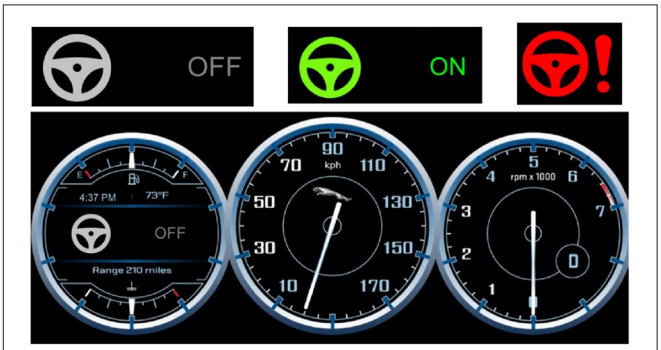


FIGURE 2 | Top: The icons showing the state of the automation. Gray icon: Autopilot OFF, Green icon: Autopilot ON, Red icon: Takeover Request (TOR). Bottom: The display of the dashboard, showing the state of automation mode, the speed and the number of engine's revolutions per minute of the car.

4.1.2. Material and Instruments

The experiment was carried out in a fixed-base driving simulator. It was a semi-enclosed cabin with low luminosity, with two car seats, a steering wheel (Logitech G27), and the pedals (throttle and brake). The orientation and position of the seats were adjustable. The scenario was a 2-lane road passing through a national park (Yosemite National Park, USA) without traffic. The car used conditionally automated driving features. The driving simulation was projected on a large screen (62 x 83 inch) using a beamer (Epsilon EH-TW3200). Two speakers behind the seats played sounds of the driving environment to immerse the driver in the simulation. The drivers could steer the wheel (more than 26 degrees), brake, or press a button on the steering to turn off the autopilot and regain full control of the vehicle. The dashboard (speed, engine rotations per minute, and autopilot mode) was run on a laptop and was displayed to the participant on a screen behind the steering wheel (cf. **Figure 2**).

Besides, a data acquisition unit (Biopac MP36) recorded the physiological signals of drivers at a sample rate of 1,000

Hz. A digital low pass filter (cut-off frequency: 66.5Hz, Q-factor: 0.5) removed the noise from the signals. The filters had a respective gain of 2,000 and 1,000 gain for EDA and RESP signals. Disposable Ag/AgCl pre-gelled electrodes (EL507 and EL503, Biopac) plugged on lead sets (SS57LA and SS2LB, Biopac) collected the EDA and ECG signals. Three electrodes were attached to record the ECG, two above both ankles and one at the right wrist. Two electrodes for recording EDA were attached to the non-dominant hand (one on the ring finger and one on the little finger) to ensure easy use of the tablet and the steering wheel during the experiment. The SS5LB respiratory effort transducer (Biopac) was attached to the participants' chest to collect the respiration signal. The Biopac Student Lab 3.7.7 software recorded the signals on a computer with a 17-inch display for a visual check of signals before starting the experiment.

Participants performed the successive sequences of non-driving-related tasks and answered midterm questionnaires on a tablet (10). An Android mobile application was developed to administer the N-back task and collect data on task performance. The N-back task was constructed using the design from Jaeggi et al. (2007). They used the letters "C," "G," "H," "K," "P," "Q," "T," and "W." In this study, the letters "G" and "W" were replaced by "N" and "F" due to the translations into French, German and Italian letters, to ensure that all letters were pronounced as differently as possible from the other letters in all three languages. It was important for the correct comprehension and recall of letters during sequences of auditory n-back. Each sequence lasted 90 s and contained 28 letters, with four letters considered as correct answers (targets) on which the participant had to press a button located on the middle of the screen. Each letter was displayed/played for 2.5 s, with an inter-stimulus of 500 ms. In the visual condition, the letter was displayed in the middle of the screen, above the red button, while in the auditory condition, the letter was only announced orally through the audio file and no letter was displayed.

4.1.3. Measures

Physiological signals (EDA, ECG, RESP) of participants were recorded continuously during the experiment. Based on these raw signals, physiological indicators could be calculated during the baseline phase (rest) and during each N-back task sequence. The tonic level of skin conductance, heart rate, and respiration rate during task epochs (with baseline correction) were used to evaluate the effect of task difficulty and modality on drivers' MWL (Mehler et al., 2009).

After each N-back task sequence, the participants reported their level of MWL through the mental demand item of the NASA-TLX questionnaire (Hart and Staveland, 1988). Participants rated it on a Likert scale from 0 (low) to 20 (high). Also, the performance on the N-back task was recorded by the mobile application. For each participant and each task sequence, the number of correct, wrong, and missed answers as well as the mean reaction time was saved. Each task sequence contained 28 items, but the participants could achieve a maximum of 27 correct answers for the 1-back task and 25 for the 3-back task. To take that into account, an indicator of performance was computed according to this formula:

$$TaskScore = (TotalAnswers - WrongAnswers - MissedTargets) * 100 / TotalAnswers \quad (1)$$

with *WrongAnswers* the number of wrong answers, *MissedTargets* the number of missed targets, and *TotalAnswers* the total number of letters that could be a target in a sequence. This aggregated score was computed to allow a fair comparison of performance between 1- and 3-back tasks. Each measure was computed 15 times because every five types of tasks (medium/high and visual/auditory + no task) was performed three times. Other dependent variables such as trust in automation, situation awareness, takeover quality, and user experience about the mobile application and the driving simulator were measured but the results are not presented in this work.

4.1.4. Procedure

Figure 3 shows the experimental procedure of the study. After initial instructions about the experiment, participants answered a questionnaire containing socio-demographic questions. Electrodes and respiration belt were then attached on the participant's body.

The experiment consisted of three main periods, which took place in the same environment: baseline, training and main driving session. During the baseline (5 min), participants were only asked to monitor the environment of the car while it was driving in conditional automation for 5 min. No takeover could be requested by the car during this period. Indicators computed during this period corresponded to the physiological baseline of each participant.

During the training period, (5 min) participants had to familiarize themselves with the driving functions (steering wheel and pedals) and the takeover process. The experimenter reminded that the car was a conditionally automated vehicle and explained the meaning of icons on the dashboard (cf. **Figure 3**). When a takeover was requested, the car displayed a red icon on the dashboard and played an audio chime in the speakers. Participants also received instructions on different ways for taking over control. In this practice session, three false alarms (e.g., no stimuli on the road) were triggered. The experimenter made sure that participants understood the takeover process and then they could drive manually until the end of the 5 min. The classification and regression tasks did not consider data from that training phase.

The main driving session lasted about an hour. The participants were given a tablet. The mobile application led them through the whole driving session and presented sequentially the instructions, the N-back tasks, and the questionnaires. Participants were asked to focus on completing the N-back task while the car was driving. No specific instruction regarding visual attention was provided for the auditory task. Participants were instructed to react accordingly to takeover requests and drive the car manually until the critical situation was handled. They were instructed to activate the automation again when they estimated

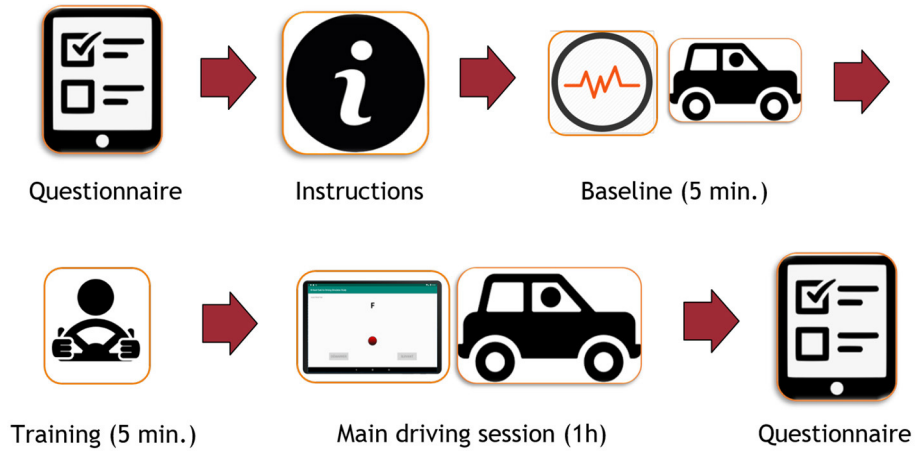


FIGURE 3 | Global experimental procedure of the study.

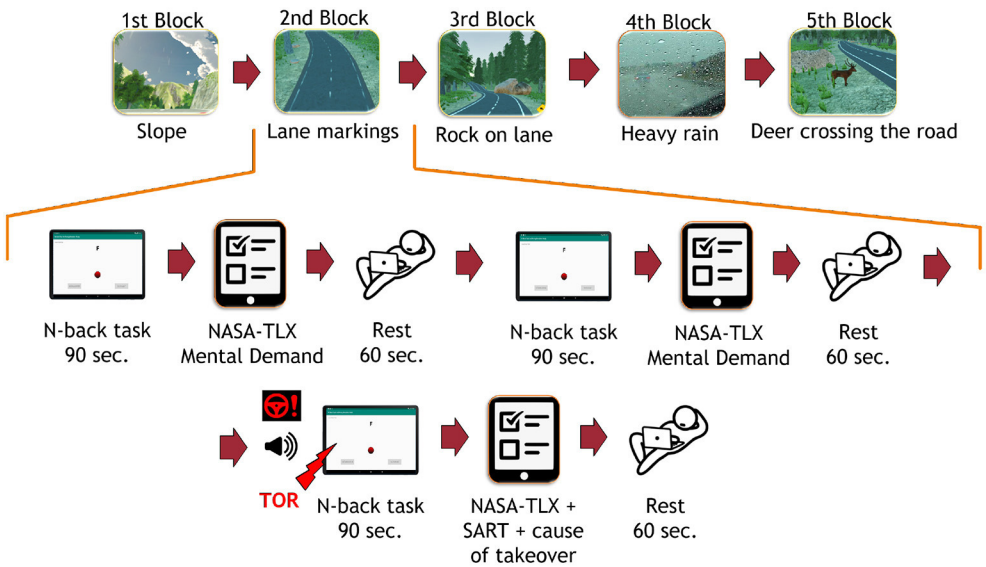


FIGURE 4 | Top: The experimental procedure during the whole driving session. Captions below images correspond to the cause of takeover request sent by the car in each block. Bottom: The experimental procedure in one block of the main driving session. TOR, Take-Over Request; SART, Situation Awareness Rating Technique. The TOR did not appear in the same position in each block.

that the situation was safe after a takeover situation. **Figure 4** shows an overview of the procedure during the main session. It consisted of five blocks, each composed of a takeover situation. During each block, the participant had to perform three N-back task sequences. The same **Figure 4** shows the procedure in one block. Each N-back task sequence was followed by a questionnaire and 60 s of rest. After the NDRT sequence in which the takeover occurred, participants had to answer the questionnaire on the tablet. At the end of the session, participants were asked to stop the car and leave the simulator to fill in the last part of the questionnaire. Electrodes were removed and participants were thanked and discharged.

4.1.5. Statistical Analysis

To check for the success of MWL manipulation, repeated measures analyses of variances (ANOVAs) were calculated using mental demand ratings and task performance for each task sequence. For both dependant variables, instructions before driving and mobile application while driving were included as between-subject factors, while task difficulty, task modality, and measurement time (2 measures) were included as within-subject factors in the statistical analysis. For the task performance, two levels were used for the task difficulty as a between-subject factor (1- vs. 3-back). For the mental demand and physiological indicators (corrected with baseline), three levels were used for

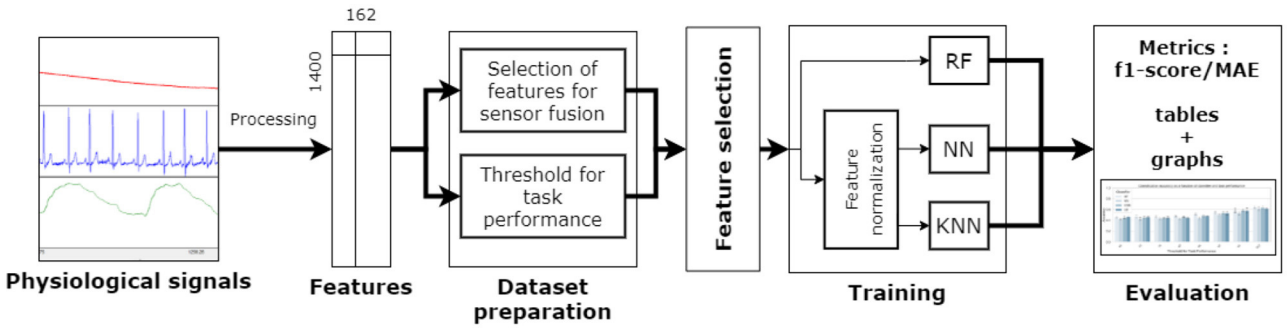


FIGURE 5 | Procedure employed for the classification and regression tasks. RF, Random Forest; NN, Neural Network; KNN, k-Nearest Neighbors.

the task difficulty as a between-subject factor (no task vs. 1-vs. 3-back). The Bonferroni method was used for adjusting the significance level ($p < 0.05$) in pairwise comparisons. The analyses were done on IBM SPSS Statistics 25.

4.2. Classification Method

This section describes the methodology used to predict the task difficulty (no task vs. low cognitive task vs. high cognitive task) and the task modality (visual cognitive task vs. auditory cognitive task), based on physiological indicators. In that regard, classification and regression tasks were both performed using machine learning techniques. As mentioned before, the effect of sensor fusion and task performance on the model's performance was also explored. The tasks performed in this study are summarized below:

- Task 1: Classification of task difficulty: effect of task performance
- Task 2: Classification of task difficulty: effect of sensor fusion
- Task 3: Regression of task difficulty: effect of task performance
- Task 4: Regression of task difficulty: effect of sensor fusion
- Task 5: Classification of task modality: effect of task performance
- Task 6: Classification of task modality: effect of sensor fusion.

For each task, the procedure employed is shown in **Figure 5**, which is similar to the one employed by Meteier et al. (2021). The following subsections explain in more detail each step of that procedure. For the classification, the model had to predict the conditions manipulated experimentally, while for the regression, the model had to predict the level of MWL on a scale between 0 and 20 (using subjective ratings as ground truth). An additional goal is to find out what are the most important features in the classification and regression processes, using an xAI technique. This might help researchers to select the most relevant physiological indicators to evaluate MWL.

4.2.1. Data Preprocessing

The process of raw physiological signals collected during the experiment was automated using the Neurokit library (Makowski et al., 2021) in a pipeline coded in Python. Raw signals from the baseline and each N-back task sequence were processed separately. Physiological data corresponding to takeover situations was used to provide the model with more training samples and potentially increase the performance. EDA, ECG, and RESP signals were all filtered with either low-pass (EDA) or band-pass (ECG and RESP) filters with adequate cut-off frequencies. The EDA signal was downsampled to 50 Hz and processed using a recent convex optimization method (Greco et al., 2016). Heartbeats were extracted from the ECG signal using a QRS-detector algorithm (Hamilton, 2002). Additional RSA features were calculated from the RESP and ECG processed signals, using the peak-to-trough (P2T) and the Porges-Bohrer methods (Lewis et al., 2012).

4.2.2. Feature Engineering and Dataset Preparation

At the end of the processing step, a large range of physiological features described in **Table 2** were computed with Neurokit (Makowski et al., 2021). For each indicator, two features were created:

- the value of the indicator while performing the N-back task (for instance, the heart rate during a task sequence)
- the difference between the value while performing the N-back task and the value during baseline (for instance, heart rate during N-back subtracted by heart rate during baseline).

The purpose of this process was to remove the physiological individual differences between drivers. Overall, 162 features from 81 indicators (10 from EDA, 48 from ECG, 16 from RESP, 7 from RSA) were calculated, for the all N-back task sequences. The size of the dataset was 162 features * 15 sequences * 80 participants = 162 x 1,400.

To test the sensor fusion, the classification with features computed from each signal alone (ECG, EDA, RESP), each possible pair of signals (EDA + ECG, EDA + RESP, ECG + RESP) and all signals combined (EDA + ECG + RESP). To investigate the effect of task performance, features from the three signals were used (EDA + ECG + RESP) and a varying threshold (from 70 to 100 by steps of 5) was applied to each task epoch. A sample (e.g., row in the dataset) was considered for training the model if the performance corresponding to that task sequence was at least higher than the chosen threshold (e.g., TaskScore in Equation 1, section 4.1.3). The number of samples considered

TABLE 2 | Indicators calculated from raw physiological signals collected from participants.

Signal	Indicator	Domain	Description
EDA	Mean raw EDA level		The mean value of filtered EDA signal
	Min raw EDA value		The minimum value of filtered EDA signal
	Max raw EDA value		The maximum value of filtered EDA signal
	Std raw EDA value		The standard deviation of filtered EDA signal
	Mean tonic EDA level		The mean value of tonic EDA signal
	Max tonic EDA value		The minimum value of tonic EDA signal
	Min tonic EDA value		The maximum value of tonic EDA signal
	Std tonic EDA value		The standard deviation of tonic EDA signal
	Mean amplitude of NS-SCRs		The mean amplitude of NS-SCRs (computed from phasic EDA signal)
	Frequency of NS-SCRs		The number of NS-SCRs per minute (computed from phasic EDA signal)
ECG/RESP	Mean Rate		The mean number of cardiac cycles per minute
	Mean	Time domain	The mean time of IBIs/BBs
	Median		The median of the absolute values of the successive differences between adjacent IBIs/BBs
	MAD		The mean absolute deviation of IBIs/BBs
	SD		The standard deviation of IBIs/BBs
	SDSD		The standard deviation of the successive differences between adjacent IBIs/BBs
	CV		The Coefficient of Variation, i.e., the ratio of SD divided by Mean
	mCV		Median-based Coefficient of Variation, i.e., the ratio of MAD divided by Median
	RMSSD		The square root of the mean of the sum of successive differences between adjacent IBIs/BBs
	CVSD		The coefficient of variation of successive differences; the RMSSD divided by Mean IBI
	HF	Frequency domain	The spectral power density pertaining to high frequency band (.15 to .4 Hz)
	SD1	Non-linear domain	Measure of the IBIs/BBs spread on the Poincar plot perpendicular to the line of identity (short-term fluctuations)
	SD2		Measure of the IBIs/BBs spread on the Poincar plot along the line of identity (long-term fluctuations)
	SD2/SD1		Ratio between long and short term fluctuations of IBIs (SD2 divided by SD1)
	ApEn		Approximate entropy
	pNN50	Time domain	The proportion of successive IBIs greater than 50 ms, out of the total number of IBIs
	pNN20		The proportion of successive IBIs greater than 20 ms, out of the total number of IBIs
	TINN		The baseline width of IBIs distribution obtained by triangular interpolation
	HTI		The HRV triangular index, measuring the total number of IBIs divided by the height of the IBIs histogram
	IQR		The interquartile range (IQR) of the RR intervals
	VHF	Frequency domain	Variability, or signal power, in very high frequency (0.4–0.5 Hz)
	HFn		The normalized high frequency, obtained by dividing the low frequency power by the total power
	LnHF		The log transformed HF
	CSI		The Cardiac Sympathetic Index
	CVI		The Cardiac Vagal Index
	CSI_modified		The modified CSI obtained by dividing the square of the longitudinal variability by its transverse variability.
ECG	S	Non-linear domain	Area of ellipse described by SD1 and SD2
	SampEn		Sample entropy
	PIP		Percentage of inflection points of the RR intervals series.
	IALS		Inverse of the average length of the acceleration/deceleration segments
	PSS		Percentage of short segments
	PAS		Percentage of IBIs in alternation segments
	GI		Guzik's Index
	SI		Slope Index
	AI		Area Index
	PI		Porta's Index

(Continued)

TABLE 2 | Continued

Signal	Indicator	Domain	Description
	C1d/C1a		Indices of respectively short-term HRV deceleration/acceleration
	SD1d/SD1a		Short-term variance of contributions of decelerations and accelerations
	C2d/C2a		Indices of respectively long-term HRV deceleration/acceleration
	SD2d/SD2a		Long-term variance of contributions of decelerations and accelerations
	Cd/Ca		Total contributions of heart rate decelerations and accelerations to HRV
	SDNNd/SDNNa		Total variance of contributions of heart rate decelerations and accelerations to HRV
RESP	Mean amplitude	Time domain	The mean respiratory amplitude.
	Mean (P2T)		Mean of RSA estimates (peak-to-trough method)
	Mean Log (P2T)		The logarithm of the mean of RSA estimates (peak-to-trough method)
	SD (P2T)		The standard deviation of all RSA estimates (peak-to-trough method)
RSA	Mean (Gates)		Mean of RSA estimates (Gates method)
	Mean Log (Gates)		The logarithm of the mean of RSA estimates (Gates method)
	SD (Gates)		The standard deviation of all RSA estimates (Gates method)
	PorgesBohrer		The Porges-Bohrer estimate of RSA, optimal when the signal to noise ratio is low, in ln(ms^2)

Those computed from both ECG and respiration (RESP) signals are grouped in the same section (ECG/RESP). IBIs, interbeat intervals; BBs, breath-to-breath intervals.

TABLE 3 | Number of samples in each class used for training the algorithms at each threshold value of task performance.

	Threshold for task performance						
	70	**75**	**80**	**85**	**90**	**95**	**100**
Task difficulty (Task 1 and 2)	453	446	442	434	393	341	254
Task modality (Task 5 and 6)	442	429	416	348	278	208	137

for training the models was hence different for each threshold value. Also, there was not an equal number of samples in each class for classifying task difficulty, because the *No Task* condition had twice fewer samples than the other classes. To address this imbalanced dataset issue, the minority classes were oversampled using the Synthetic Minority Oversampling Technique (Chawla et al., 2002). To summarize, the number of samples used for each threshold value can be found in **Table 3**.

4.2.3. Feature Normalization and Selection

A feature normalization process has been applied to feature scale sensitive models, using the RobustScaler function of the scikit learn machine-learning framework (Pedregosa et al., 2011). For each feature, the median was subtracted to all samples, which were scaled according to the interquartile range (between the first quartile and the third quartile of data distribution for each feature). For all models, a univariate feature selection process reduced the dimension of the feature space and so the computation time. The main goal of this process was also to optimize models' performance by selecting only the most relevant features. The 20 best features were selected based on univariate statistical tests, using the SelectKBest method of the scikit learn framework.

4.2.4. Selected Algorithms

The selected features are used as input of machine learning algorithms for training these models and then validating their performance. Three algorithms were selected because they can be used for both classification and regression tasks. They were implemented in Python using the scikit learn machine learning framework (Pedregosa et al., 2011). The selected algorithms were Random Forest (RF), Neural Network (NN), k-Nearest Neighbors (KNN).

4.2.5. Model Evaluation and Explanation

For each task performance threshold or combination of physiological signals, a repeated k-fold procedure was employed. The training and evaluation procedure was run 5 times, to report accurate results over several iterations. For each iteration, the dataset was randomly split into a training set (80%) and a test set (20%). To optimize the performance of models, the grid search approach was employed during the training phase. The goal was to find the set of hyperparameters that maximizes the performance of each algorithm (Claesen and De Moor, 2015). A k-fold cross-validation approach was selected to train the models. The training set was split into $k = 4$ folds, each fold acting as the validation set once. Each set of hyperparameters shown in **Table 4** was tested for each split of the dataset. The best model (e.g., the one that gave the best score over the 4 folds) was then evaluated on the test set. For the classification tasks, the weighted f1-score was used as an evaluation metric, since Task 1 and Task 2 are multi-label classification tasks (3 classes). For the regression tasks (Task 3 and 4), the mean absolute error (MAE) was computed to evaluate the performance of models. To compare the models' performance to a reference, the following baseline metrics were calculated:

- Random : a random value between 0 and 20

- MeanScale : mean value of NASA-TLX scale (10)
- MeanParticipants : the mean of mental demand score reported by participants for NASA-TLX ($M = 8.625$)
- MeanGroup : Mean of participants in each condition (no task vs. 1- vs. 3-back); the mean of mental demand score reported by participants in each condition ($M_{notask} = 3.247$, $M_{1-back} = 5.852$, $M_{3-back} = 14.099$).

Results are reported in graphs and tables, which are the best mean weighted f1-score or MAE achieved by each algorithm on the test set over the 5 iterations. The effect of sensor fusion was tested with a threshold value of 100, while the effect of task performance was tested using the three signals (EDA + ECG + RESP). To find the most relevant indicators of MWL, the most important features (e.g., physiological indicators) in the classification/regression process had to be extracted using the SHAP (SHapley Additive exPlanations) library in Python (Lundberg and Lee, 2017). By assigning an importance value to each feature for a particular prediction, it helps visualize the values of the most important features depending on the predicted class. After the training and evaluation procedure for classifying task difficulty, the best model was saved and used for generating SHAP values. The 10 most significant features were extracted, in descending order (ordered by absolute mean of SHAP value).

5. RESULTS

5.1. Statistical Validation of MWL Inducement

5.1.1. Performance on Task

The correct implication of participants in the non-driving related task was assessed using the aggregated score of task performance. Data analysis revealed only a significant effect of task difficulty on task performance [$F_{(1,76)} = 228.83$, $p < 0.001$, $\eta_p^2 = 0.75$]. Participants performed better at doing the 1-back task ($M = 97.6$, SD = 0.5%) than the 3-back task ($M = 86.2$, SD = 0.6%). Otherwise, there was no significant effect of task modality [$F_{(1,76)} = 2.90$, $p > 0.05$, $\eta_p^2 = 0.04$] and measurement time [$F_{(1,76)} = 1.14$, $p > 0.05$, $\eta_p^2 = 0.01$]. The double and triple interaction effects were not significant (Fs < 1).

5.1.2. Subjective Reports of MWL

The success of the MWL manipulation was evaluated using subjective ratings of workload from the mental demand item of thr NASA-TLX questionnaire. **Figure 6** shows the ratings of participants, depending on the modality and difficulty of the task. Data analysis revealed a significant effect of task difficulty on MWL of drivers [$F_{(2,152)} = 338.39$, $p < 0.001$, $\eta_p^2 = 0.82$]. Pairwise comparisons showed that participants found the 3-back task significantly more demanding ($M = 14.26$, SE = 0.40) than the 1-back task ($p < 0.001$; $M = 5.18$, SE = 0.38) or when performing no secondary task ($p < 0.001$; $M = 2.46$, SE = 0.39). Interestingly, the effect of measurement time (first vs. second task epoch) was significant on subjective reports of MWL from the drivers [$F_{(1,76)} = 4.57$, $p < 0.05$, $\eta_p^2 = 0.06$]. Participants reported that the first epoch of each task was significantly more demanding ($M = 7.53$, SE = 0.33) than the second one ($M = 7.07$, SE = 0.27). Otherwise, there was no significant effect of task modality [$F_{(1,76)} = 2.56$, $p > 0.05$, $\eta_p^2 = 0.03$] alone. Also, there was a significant interaction effect of task difficulty and modality [$F_{(2,152)} = 4.15$, $p < 0.05$, $\eta_p^2 = 0.05$]. Pairwise comparisons showed that participants reported that the visual 1-back task ($M = 5.52$, SE = 0.40) was significantly more demanding ($p < 0.01$) than the auditory 1-back task ($M = 4.84$, SE = 0.40), while the visual 3-back task ($M = 14.24$, SE = 0.41) was not significantly more demanding ($p < 0.05$) than the auditory 3-back task ($M = 14.28$, SE = 0.44). A significant interaction effect of task difficulty and measurement time on MWL [$F_{(2,152)} = 3.70$, $p < 0.05$, $\eta_p^2 = 0.05$] was also found. Pairwise comparisons showed that participants reported higher mental demand the first time they did not perform any secondary task (M = 3.05, SE = 0.54) than the second time ($p < 0.05$; $M = 1.86$, SE = 0.38), while it was not the case for 1-back and 3-back tasks ($p > 0.05$). Besides, the interaction effect of measurement time and modality, as well as the triple interaction effect were not significant (Fs < 1).

5.1.3. Physiological Indicators

Figure 7 shows the change in EDA tonic level, heart rate and respiratory rate of participants, depending on the task difficulty and modality. Data analysis revealed a significant effect of task modality [$F_{(1,73)} = 7.23$, $p < 0.01$, $\eta_p^2 = 0.09$] and measurement time [$F_{(1,73)} = 4.83$, $p < 0.05$, $\eta_p^2 = 0.06$] on EDA tonic level of drivers, but no significant effect of task difficulty [$F_{(2,146)} = 0.869$, $p > 0.05$, $\eta_p^2 = 0.01$]. Drivers had a higher change in EDA tonic level when performing the auditory tasks ($M = 2.78$, SE = 0.22) compared to the visual tasks ($M = 2.65$, SE = 0.20). They also showed a higher change in the second epoch of each type of task ($M = 2.82$, SE = 0.22) compared to the first one ($M = 2.61$, SE = 0.20). The double and triple interaction effects were not significant ($p < 0.05$).

Data analysis revealed a significant effect of task difficulty [$F_{(2,146)} = 8.82$, $p < 0.001$, $\eta_p^2 = 0.11$] and measurement time [$F_{(1,73)} = 37.96$, $p < 0.001$, $\eta_p^2 = 0.34$] on heart rate of drivers, but no significant effect of task modality ($F < 1$). Pairwise comparisons showed that participants that the change in drivers' heart rate was significantly higher when performing the 3-back task ($M = -0.35$, SE = 0.51) than when performing the 1-back task ($p < 0.001$; $M = -1.67$, SE = 0.50) or no task (e.g., monitoring the driving environment; $p < 0.05$; $M = -1.46$, SE = 0.51). They also had a higher heart rate in the first epoch of each type of task ($M = -0.34$, SE = 0.42) compared to the second one ($M = -1.97$, SE = 0.54). The double and triple interaction effects were not significant ($p < 0.05$).

Identically to heart rate, results show a significant effect of task difficulty [$F_{(2,146)} = 37.72$, $p < 0.001$, $\eta_p^2 = 0.34$] and measurement time [$F_{(1,73)} = 8.22$, $p < 0.001$, $\eta_p^2 = 0.10$] on respiratory rate of drivers, but no significant effect of task modality [$F_{(1,73)} = 2.30$, $p > 0.05$, $\eta_p^2 = 0.03$]. Pairwise comparisons showed that participants that the change in drivers' respiratory rate was significantly different between one condition to another ($p < 0.001$). **Figure 7** show that the change was the highest during the 3-back task, followed, respectively, by 1-back task and no task conditions. Also, participants had a higher

TABLE 4 | Hyperparameters values tested during the grid search procedure, with chosen ranges and step values for each parameter.

Classifier	Parameter name	Parameter definition	Range
RF	n_estimators	Number of trees in the forest.	[10, 257, 505, 752, 1,000]
	max_features	Number of features to consider when looking for the best split.	sqrt
	max_depth	Maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than 2 samples.	[None, 10, 40, 70, 100]
KNN	n_neighbors	Number of neighbors considered.	[5, 10, 20, 30]
	weight	weight function used in prediction.	[uniform, distance]
	algorithm	Algorithm used to compute the nearest neighbors.	[auto, ball_tree, kd_tree, brute]
NN	alpha	L2 penalty (regularization term) parameter.	[1e-4, 1] by step of 10
	hidden_layer_sizes	The number of neurons in the hidden layer.	[32, 64, 128, 256]

RBF, Radial Basis Function.

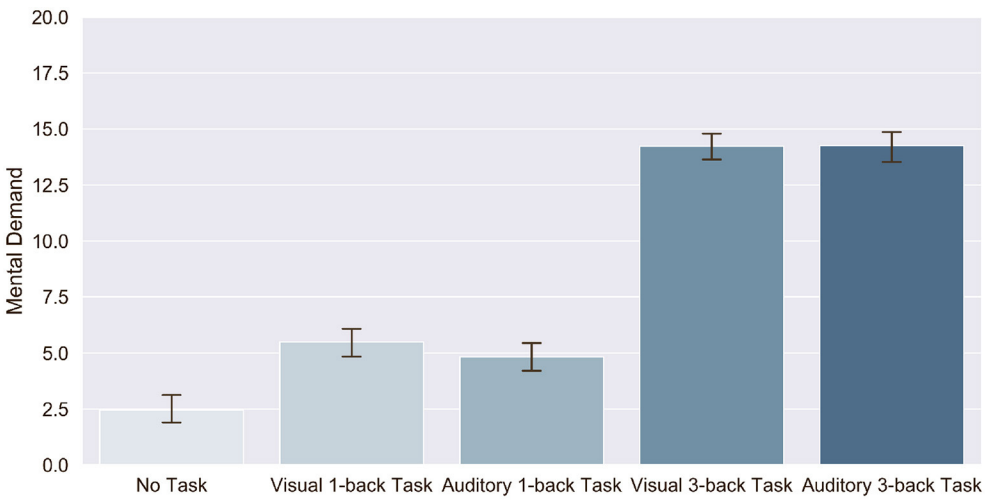


FIGURE 6 | Effect of task modality and difficulty on subjective ratings of mental demand reported after each sequence of N-back task.

respiratory rate in the first epoch of each type of task ($M = 1.23$, SE $= 0.56$) compared to the second one ($M = 0.32$, SE $= 0.48$. The double and triple interaction effects were not significant ($p < 0.05$).

5.2. Classification of Drivers' Workload Through Task Difficulty

5.2.1. Task 1 : Effect of Task Performance on Classification Accuracy

As mentioned earlier, task performance may decrease with increasing task difficulty, either because of drivers' skills or because some drivers may be tempted to abandon the task if it becomes too complicated. In this case, the physiological activation induced by the task would be reduced. For this reason, the influence of task performance on the model's accuracy for predicting task difficulty was investigated. **Table 3** (Task difficulty row) summarizes the number of samples contained in all classes for training the model at each threshold value. **Figure 8** shows the average f1-score (with standard deviation) on the test set over the 5 iterations, as a function of classifier and threshold value used for the task performance. Features were considered if the participant performed at least above the performance threshold during the task. **Table 5** summarizes the best score achieved by each classifier for each threshold value. To better understand the predictions of the best model (a Random Forest classifier with the three signals and a task performance threshold of 100), a confusion matrix is proposed in **Figure 9**. **Figures 10**, **11** show the features that had the most impact on the model predictions for predicting the MWL of drivers between the three levels. They show the SHAP values calculated with the best model for all samples of the test set.

5.2.2. Task 2 : Effect of Sensor Fusion on Accuracy

As shown in **Figure 8**, the task performance affects the physiological activation of the drivers and thus the accuracy of the models. Therefore, the effect of sensor fusion was analyzed. The performance of the models in classifying drivers' MWL as a function of task difficulty (no task, 1-back task, 3-back task) is presented in **Figure 12**. It shows the weighted average f1-score (with standard deviation) of each classifier and each

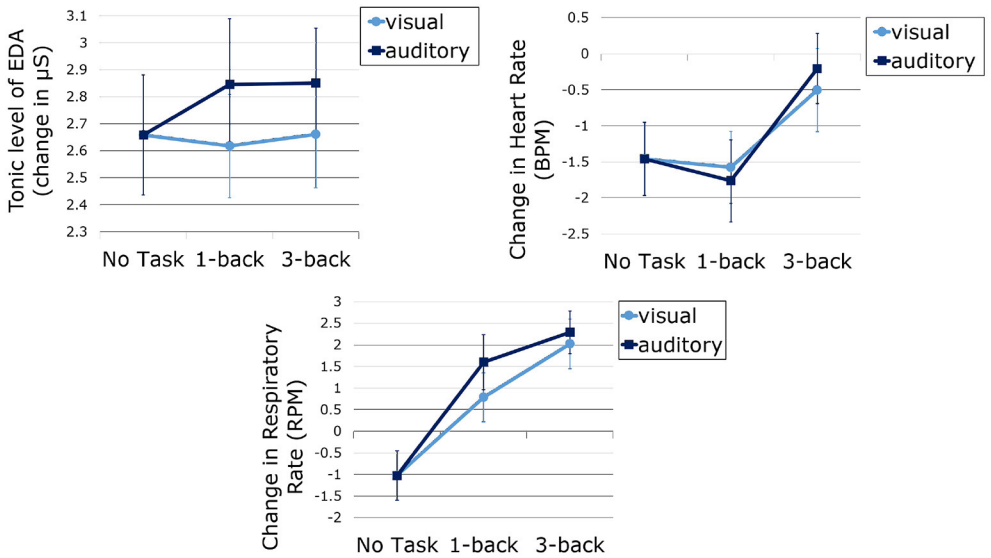


FIGURE 7 | EDA tonic level (top left), heart rate (top right) and respiratory rate (bottom) measured during the tasks and corrected with baseline, as a function of task difficulty and modality. Error bars represent standard error.

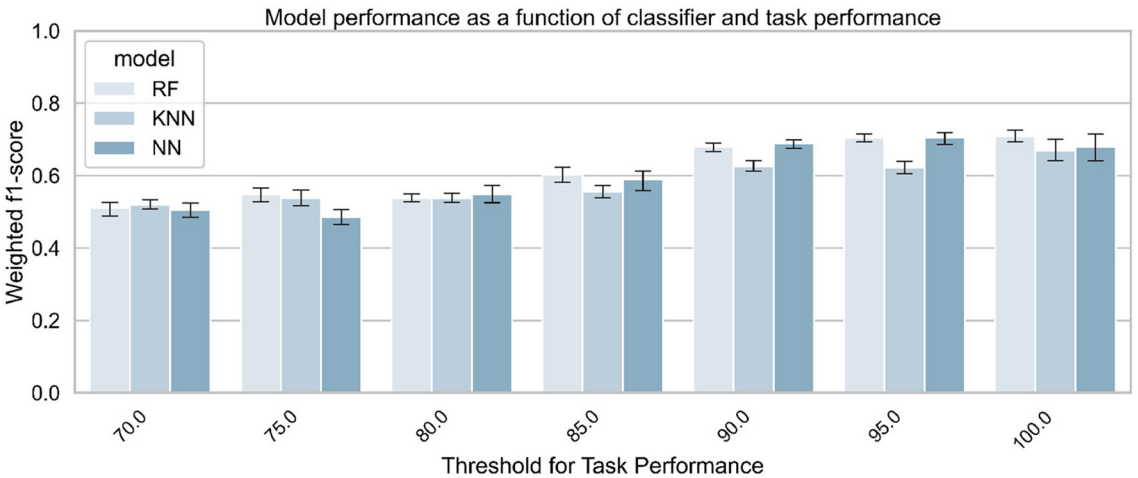


FIGURE 8 | Classifiers' performance for predicting task difficulty (no task vs. 1- vs. 3-back), as a function of classifier and task performance. The three signals (EDA + ECG + RESP) were used to train the classifiers.

signal combination on the test set over the 5 iterations. **Table 6** summarizes the best score obtained for each combination of input signals.

5.3. Regression of Drivers' Workload Using Subjective Reports

5.3.1. Task 3 : Effect of Task Performance on Regression Error

Regression tasks were performed to obtain a finer assessment of MWL. The goal was to study whether a machine learning model can assess the self-reported MWL with low error (on a scale of 0–20). First, the effect of task performance on the regression error was tested. **Figure 13** shows the model error for the MWL regression, depending on the algorithm and the threshold value used for the task performance. It shows the average MAE on the test set over the 5 iterations. As the MAE is used as a metric, this means that the lower the score, the better the model (closer to the ground truth). **Table 7** summarizes the best scores obtained by the algorithm for each threshold value, compared to various baseline metrics (defined in section 4.2.5).

5.3.2. Task 4 : Effect of Sensor Fusion on Regression Error

As with the classification tasks, the effect of sensor fusion was also investigated to see if the model performs better with a certain combination of signals. **Figure 14** shows the model

TABLE 5 | Best score achieved by the model to predict task difficulty at each threshold of task performance.

Threshold for task performance	Best classifier	f1-score [Mean (SD)]
70	KNN	0.519 (0.018)
75	RF	0.548 (0.026)
80	NN	0.549 (0.033)
85	RF	0.602 (0.026)
90	NN	0.688 (0.015)
95	NN	0.705 (0.021)
100	**RF**	**0.710 (0.022)**

The value in bold is the best score achieved by the model among all possible combinations.

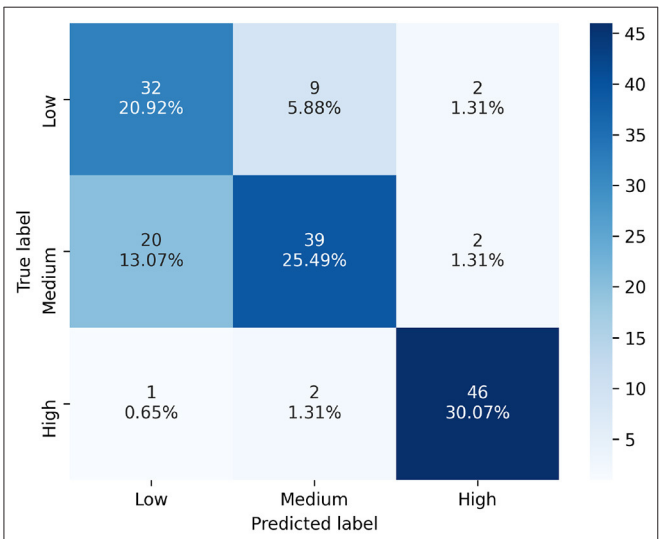


FIGURE 9 | Confusion matrix of the best model's predictions for classifying task difficulty, using the three signals (EDA + ECG + RESP) and a task performance threshold of 100. Labels : Low = No task; Medium = 1-back task; High = 3-back task.

error for MWL regression, as a function of the algorithm and the combination of signals used for training the algorithm. It shows the average error on the test set over the 5 iterations after the quadruple cross-validation training procedure. **Table 8** summarizes the best score obtained by the corresponding algorithm for each combination of signals, compared to various baseline metrics (defined in section 4.2.5).

5.4. Classification of Task Modality: Visual vs. Auditory

5.4.1. Task 5 : Effect of Task Performance on Classification Accuracy

Table 3 (Task Modality rows) summarizes the number of samples from each class that was considered for training the model at each threshold value. **Figure 15** shows the average performance of the model over 5 iterations, as a function of the classifier and the threshold value used for the task performance. **Table 9** summarizes the best score obtained by the corresponding classifier for each threshold value.

5.4.2. Task 6 : Effect of Sensor Fusion on Classification Accuracy

The accuracy of the model for the classification of the task modality (visual vs. auditory task) is presented in **Figure 16**. It shows the averages (and standard deviations) of the weighted f1 score obtained by the model for each classifier and each signal combination on the test set over the 5 iterations. **Table 10** summarizes the best result obtained for each signal combination.

6. DISCUSSION

6.1. Manipulation of MWL : Task Performance and Subjective Reports

Data analysis revealed only a significant effect of task difficulty on task performance, which is consistent with previous studies (Mehler et al., 2009, 2012). Participants were correctly implicated in the 1-back task (task score of 97.6/100), and performed worse at the 3-back task (task score of 86.2/100), which is coherent with the increase in task difficulty. Results obtained on task performance are in line with subjective reports of mental demand after the tasks. because the task difficulty had a significant effect on MWL. **Figure 6** shows that the subjective mental demand increases with task difficulty. This result also means that according to participants, performing a 1-back task is more demanding than only monitoring the environment of the car.

Besides, there was a significant effect of measurement time (first vs. second epochs) on subjective reports of MWL. The significant interaction effect of measurement time and task difficulty suggests that it was only the case while monitoring the driving environment (no task condition). Participants reported that the first sequence of *No Task* was more demanding than the second one. They might have been used to monitor the environment of the car and hence it required less mental resources throughout the experiment. Also, they might have compared with sequences of 1-back and 3-back tasks, so they have probably lowered the score associated with mental demand after the second sequence of *No Task*. Nevertheless, this may only be a subjective feeling.

Task modality did not show any significant effect on task performance, meaning that participants performed equally in auditory and visual tasks. It also did not show an effect on subjective reports of MWL. However, an interaction effect of task modality and difficulty was found. Participants felt that at the 1-back level, the visual task was significantly more demanding than the auditory task. However, this result was not consistent at the 3-back level, so it is hard to conclude this significant effect.

Since the effect of task difficulty on measures of task performance and workload was significant, we can say that the manipulation of workload at three levels was successful. Based on that, the no task, 1-back, and 3-back conditions can be considered, respectively to states of a low, medium, and high MWL in the remaining part of the manuscript.

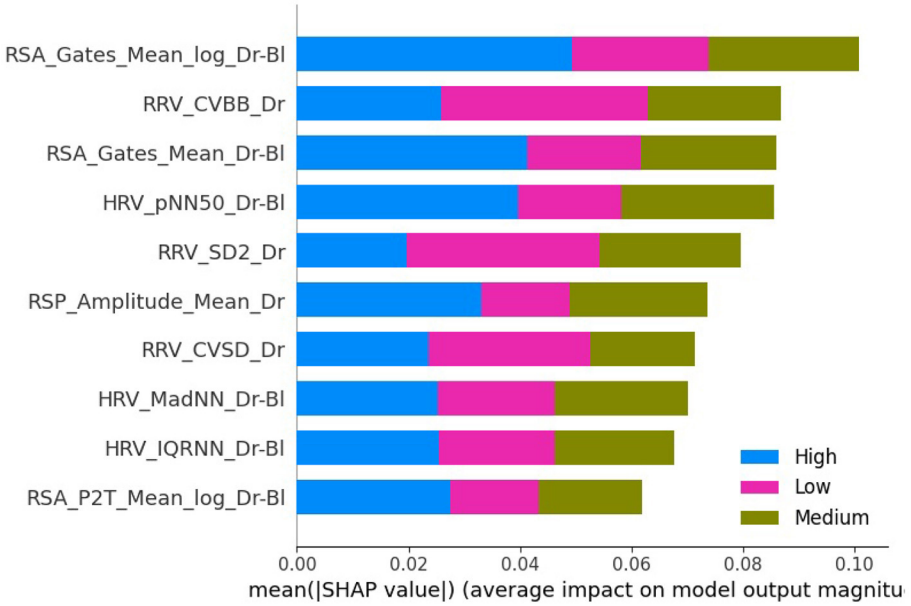


FIGURE 10 | Plot bar graph of the 10 most impacting features to predict drivers' condition based on mean absolute SHAP values, arranged in descending order. The three signals (EDA + ECG + RESP) and a threshold for task performance of 100 were used. The meaning/description of each feature can be found in **Table 2**. RRV, Respiratory rate variability; HRV, Heart Rate Variability; RSA, Respiratory Sinus Arrhythmia.

6.2. Influence of MWL on the Physiological State of Drivers

Data analysis revealed a significant effect of task difficulty on the mean heart rate and respiration rate but not on EDA. Heart rate was higher in periods of high MWL (3-back) compared to medium and low MWL, while respiration rate was different between each level of MWL. These results are in line with previous findings (Collet et al., 2009; Mehler et al., 2009, 2012), since heart and respiration rates increase with task demand (e.g., increasing workload). However, there was no difference in drivers' heart rate while monitoring the environment and performing the 1-back task. However, it is unexpected to find no significant effect of task difficulty of EDA tonic level like in previous findings (Engstrm et al., 2005; Mehler et al., 2009, 2012). This was most probably due to the low engagement of some drivers in the NDRTs, as suggested by Mehler et al. (2012) after the non-significant effect found for task difficulty on EDA in the work of Engstrm et al. (2005). This unexpected result is consistent with the claim made in the related work section that it is important to control task performance when manipulating the MWL. The non-significant difference of physiological values between *No Task* and *1-back task* is further discussed below. In addition, the tonic level of EDA was also higher on the second occurrence of each type of task, probably due to the repetition of the cognitive tasks to be performed and the demands for car pickup throughout the experiment. However, the opposite effect was found for heart and respiratory rates, which were higher in the first measurement. This could suggest a habituation effect to the task, or that heart and respiratory rates do not increase significantly with a long period of conditionally automated driving (1 h) and repeated takeover requests (5) to manage. EDA is also likely to be more sensitive to takeover requests (an audio sound was played for each request) and the tonic level of EDA may take longer to return to a "normal" state of physiological activation (Boucsein, 2012).

6.3. Classification and Regression of Drivers' Workload

To further investigate the effect of sensor fusion and task performance on the physiological state of automated vehicle drivers, classification and regression tasks were performed using machine learning techniques. For the 3-level classification task, the results show that MWL can be predicted with 71% accuracy (with f1-score as the measure) using the EDA and RESP signals as input of a random forest classifier and a task performance threshold of 100. The results are close to those obtained in some previous studies that classified MWL at only two levels (Hogervorst et al., 2014), which is encouraging for the future. The results for the regression task are consistent with those obtained for the classification. The regression showed that the level of subjective mental load reported by the participants can be predicted to plus or minus 3.195 error (on a scale of 0–20), using the 3 input signals and a task performance threshold of 100. All models tested outperformed the baseline measures, which means that the implemented model can be considered intelligent and more effective than a random prediction of mental load.

Results for both types of tasks are consistent since they show an effect of task performance on model performance. Indeed, model performance increased with better performance

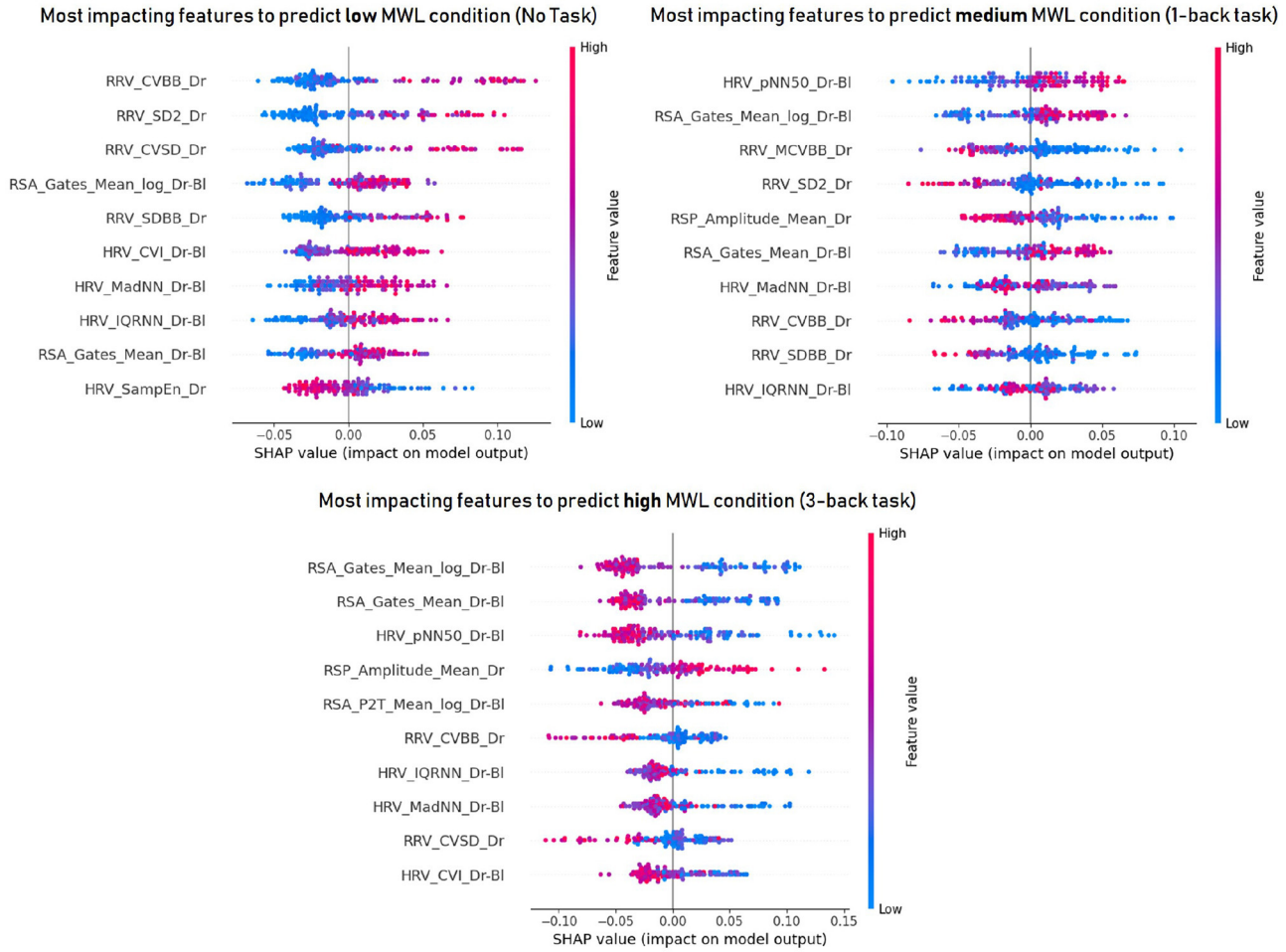


FIGURE 11 | The 10 most important features to predict drivers' mental workload at each level : low (top left), medium (top right) and high (bottom) mental workload. The three signals (EDA + ECG + RESP) and a threshold for task performance of 100 were used. A high SHAP value (to the right on the x-axis) indicates that this feature influenced positively the model to predict that class. The meaning/description of each feature can be found in **Table 2**. RRV, Respiratory rate variability; HRV, Heart Rate Variability; RSA, Respiratory Sinus Arrhythmia.

on the cognitive tasks. This result suggests that participants' physiological activation is higher when they are properly involved in a cognitive task Mehler et al. (2012). This also suggests that task performance must be controlled during experimental manipulation of the workload in order to obtain consistent results. The effect of sensor fusion was also similar for classification and regression. Model performance increases slightly with signal fusion, although the difference is small between the models using 2 or 3 signals. From the results, it is difficult to conclude that one signal is more effective in predicting mental load than another. Still, the effect of sensor fusion on models' performance are in line with a previous recent study also conducted in conditionally automated driving Meteier et al. (2021). In both studies, EDA is the input signal that performed the worst, which is also in line with the results obtained in the statistical analysis. This unexpected result can be explained by the fact that the participants were holding a tablet to perform the task, which may have induced some noise in the signal. In addition, the repetition of the takeover requests may have attenuated the increase in skin conductance due to the increase in cognitive load during the tasks. The fusion of the three signals (EDA + ECG + RESP) was always among the best results. This shows the importance of multi-modality, allowing to combine features from different signals and thus ensuring a robust evaluation of the mental load.

In this work, the f1-score obtained by the models remains relatively low. This can be explained by the difficulty of the model to distinguish between phases of low cognitive task (1-back) and phases of observation of the vehicle environment (no task). This is illustrated by the confusion matrix in **Figure 9**. This suggests that observing the vehicle environment or performing a mildly cognitive task on a digital device could induce the same level of cognitive load to the driver. Thus, this implies that drivers might be allowed to engage in mildly cognitive NDRTs in conditional automated driving, with respect to physiological activation.

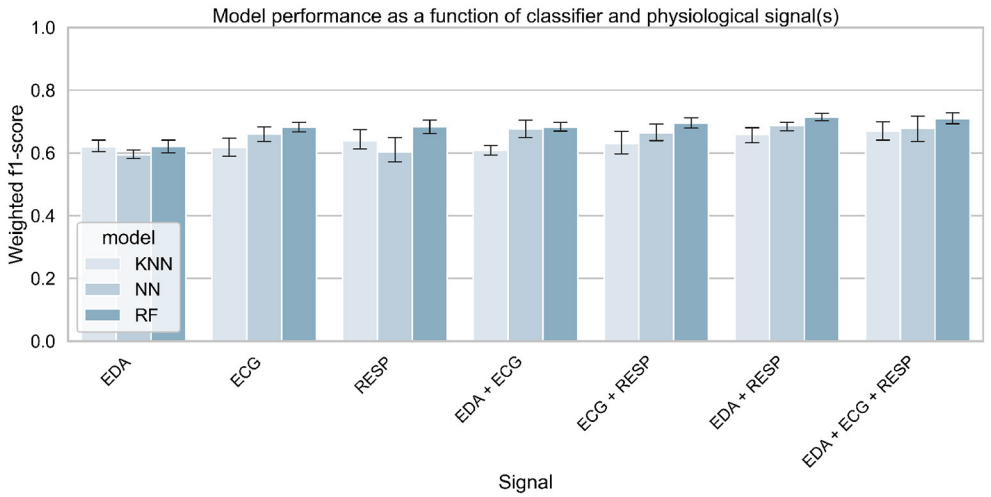


FIGURE 12 | Classifiers' performance for predicting task difficulty (no task vs. 1- vs. 3-back), as a function of classifier and selected physiological signals. A threshold for task performance of 100 was selected.

TABLE 6 | Best score achieved by the model to predict task difficulty for each combination of physiological signals.

Selected signal	Best classifier	f1-score [Mean (SD)]
EDA	RF	0.620 (0.027)
ECG	RF	0.683 (0.020)
RESP	RF	0.684 (0.028)
EDA + ECG	RF	0.681 (0.018)
ECG + RESP	RF	0.695 (0.023)
EDA + RESP	**RF**	**0.713 (0.015)**
EDA + ECG + RESP	RF	0.710 (0.022)

The value in bold is the best score achieved by the model among all possible combinations.

6.4. Relevant Indicators of Workload

In order to go even further in the explainability of the machine learning models, an explainable AI technique was applied to the best classifier to find the most relevant indicators to measure MWL. **Figure 11** shows that among the 10 indicators with the highest impact in predicting mental load, 4 are respiratory sinus arrhythmia indicators, 3 are respiratory rate variability indicators and 3 are cardiac variability indicators, which is consistent with the literature (Boyce, 1974; Muth et al., 2012; Hidalgo-Muoz et al., 2019). In particular, respiratory sinus arrhythmia (corrected to baseline) according to the Gates method (Gates et al., 2015) seems to be the most relevant indicator, especially for high mental load states. According to the results obtained in this experiment, RSA estimates decrease with increasing mental load (low values toward the right of the x-axis in **Figure 11**), which is consistent with previous studies (Boyce, 1974; Muth et al., 2012). This is associated with a decrease in cardiac variability and an increase in respiratory amplitude. Whereas, a previous study indicated that respiratory amplitude appears to remain stable with increasing MWL (Grassmann et al., 2016), the results obtained in this study suggest that participants breathed more heavily in a high mental load condition. This should be further investigated.

6.5. Classification of Task Modality

An additional goal of this work was to test whether the task modality performed by the driver could be recognized using physiological signals and machine learning. The results show that the model was only able to predict the task modality with an accuracy of 61.8% measured by the f1-score, using ECG and RESP as input signals and a threshold of 100 for the task performance. Most models tested with various combinations of thresholds for task performance and input signals have often achieved a performance of around 50%-accuracy. Hence, the effect of task performance on model performance to predict task modality is unclear. Only the threshold of 100 significantly increased model performance. These results suggest that it is difficult to predict the modality of the task performed by the driver from physiological signals alone. With the results obtained in our study, we suggest using other data sources such as cameras to predict the modality of the task performed by drivers and support them accordingly. Previous studies have shown that certain task modalities can negatively impact the driver's ability to take control of automated driving (Wandtner et al., 2018; Roche et al., 2019) and the driver's awareness of his or her environment (Meteier et al., 2020). Thus, knowing the type of task the driver is performing would optimally convey contextual information about the driving environment and thus increase situational awareness.

6.6. Limitations and Further Research

This study was conducted with young drivers (average age 24) in a simulator. This may have influenced the results obtained, as the mental workload induced in real driving conditions or with drivers of different ages is certainly not the same. Also, the scenario did not include traffic, which could have influenced the

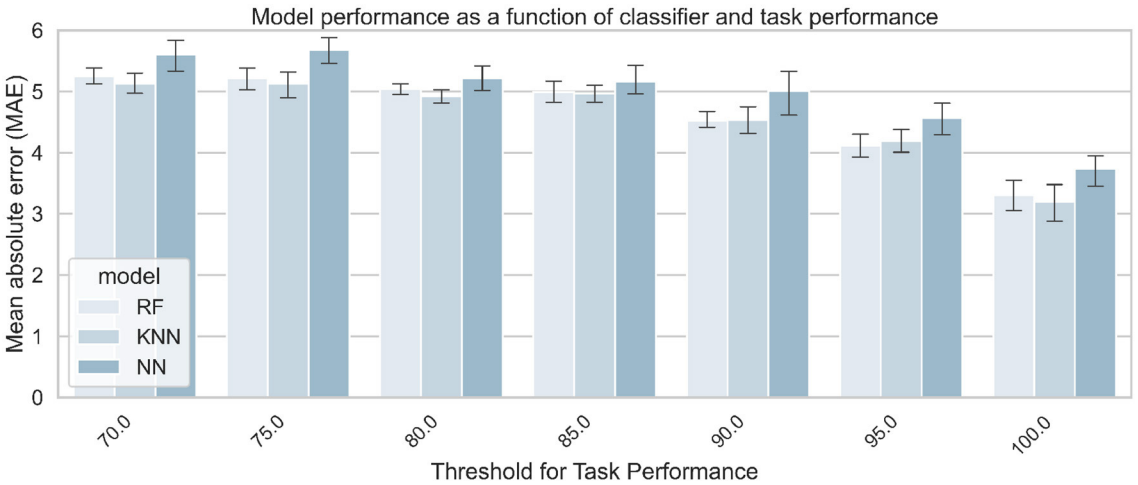


FIGURE 13 | Models' performance for predicting MWL on a 0–20 scale, as a function of algorithm and task performance. The three signals (EDA + ECG + RESP) were used to train the classifiers.

TABLE 7 | Best score achieved by the model to predict task difficulty at each threshold of task performance.

Threshold	Best model	MAE [Mean (SD)]	Random	MeanScale	MeanParticipants	MeanGroup
70	KNN	5.123 (0.208)	7.177	5.903	5.831	6.425
75	KNN	5.123 (0.277)	7.197	5.671	5.556	6.339
80	KNN	4.919 (0.146)	7.485	5.892	5.726	6.369
85	KNN	4.7968 (0.177)	7.131	5.917	5.655	6.223
90	RF	4.522 (0.166)	7.700	6.157	5.613	5.748
95	RF	4.113 (0.235)	7.748	6.592	5.854	5.328
100	**KNN**	**3.195 (0.384)**	8.085	6.912	5.934	4.438

Scores obtained for baseline metrics are also reported. The value in bold is the best score achieved by the model among all possible combinations.

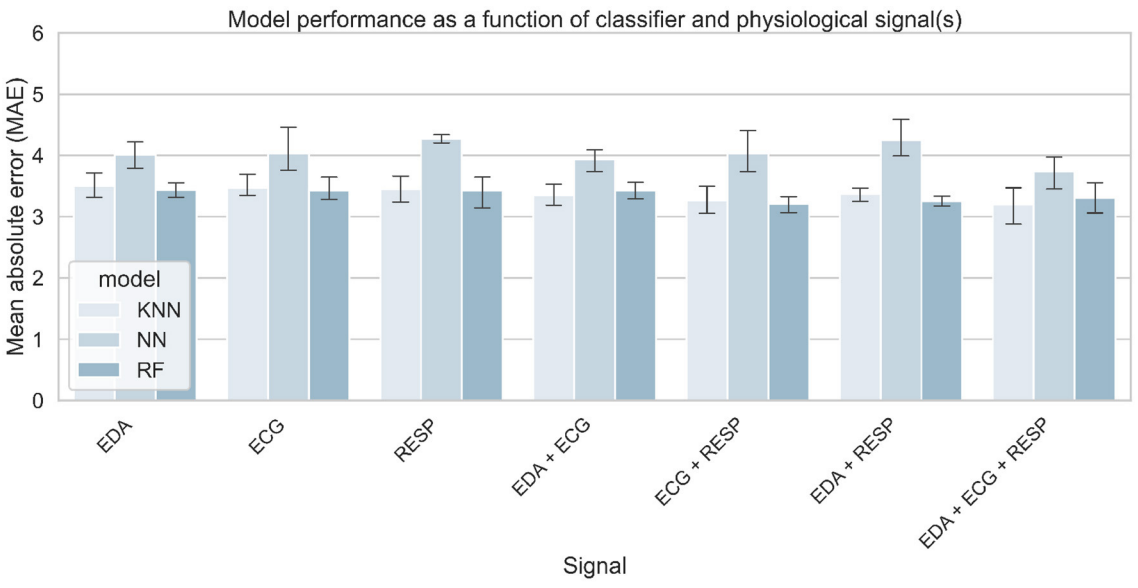


FIGURE 14 | Models' performance for predicting MWL on a 0–20 scale, as a function of selected physiological signals and algorithm. A threshold for task performance of 100 was selected.

TABLE 8 | Best score achieved by the model to predict task modality for each combination of physiological signals.

Signal(s)	Model	MAE [Mean (SD)]	Random	MeanScale	MeanParticipants	MeanGroup
EDA	RF	3.436 (0.154)	7.870	6.981	5.954	4.665
ECG	RF	3.425 (0.236)	7.905	6.527	5.562	4.180
RESP	RF	3.432 (0.329)	7.871	6.792	5.850	4.772
EDA + ECG	KNN	3.348 (0.348)	7.642	6.954	5.923	4.561
ECG + RESP	RF	3.206 (0.165)	7.634	6.696	5.691	4.267
EDA + RESP	RF	3.249 (0.105)	8.035	6.886	5.832	4.266
EDA + ECG + RESP	**KNN**	**3.195 (0.384)**	8.085	6.912	5.934	4.438

Scores obtained for baseline metrics are also reported. The value in bold is the best score achieved by the model among all possible combinations.

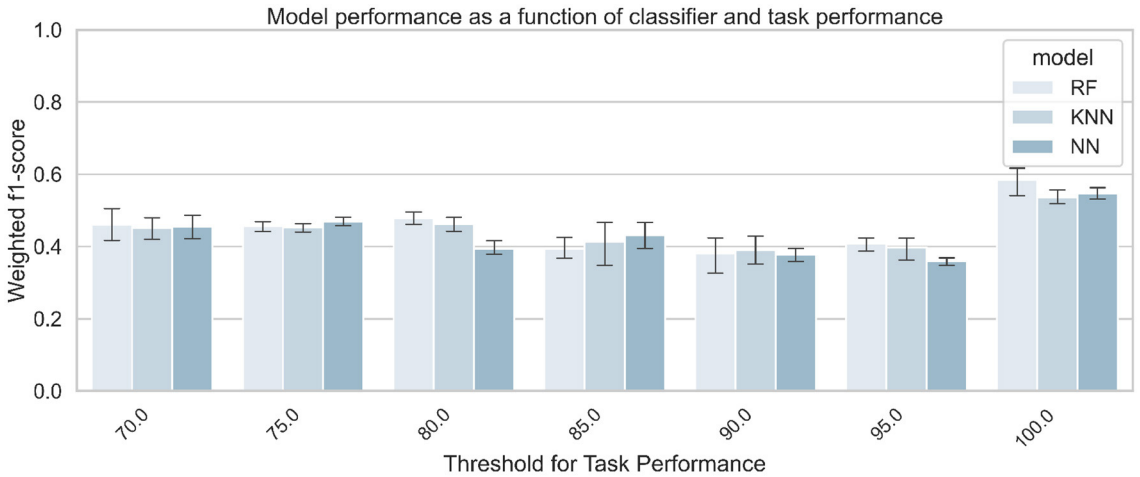


FIGURE 15 | Classification accuracy for predicting task modality (visual vs. auditory), as a function of classifier and task performance. The three signals (EDA + ECG + RESP) were used to train the classifiers.

drivers' MWL. Other factors were experimentally manipulated in this experiment but were not presented in this work. These may have influenced the participants' physiological and mental state. For example, the presence of a split-screen mobile application on the tablet for half of the participants throughout the experiment may have induced additional mental load (Meteier et al., 2020). In addition, some participants commented on the repetitive and monotonous nature of the non-driving-related task. They may have lost motivation during the experiment, which was reflected in the effect of task performance on the results. To mitigate this problem, a question could have been administered to them to subjectively measure their engagement in the NDRT.

For the non-significant effect found for task difficulty on EDA, one solution would be to take task performance into account in the statistical analysis. Another possibility would be not to take into account the periods after each takeover request, as this could have induced a large increase in EDA and thus biased the results for the non-driving-related task periods.

Regarding the classification results, we are still far from an accuracy of 100%. On the other hand, the results obtained for the regression are encouraging since the model can be considered as intelligent. However, the results obtained must be interpreted with caution. Indeed, the label used as ground truth was a subjective value. Even if this score was reported just after the task to limit recall problems, the score predicted by the model during the regression was perhaps sometimes closer to reality. A solution to this problem would be to use the performance during the task to regress the mental load instead, to assess the mental load more accurately.

To improve the results obtained for the classification and regression of mental load from physiological indicators, more

TABLE 9 | Best score achieved by the model to predict task modality at each threshold of task performance.

Threshold for task performance	Best classifier	f1-score [Mean (SD)]
70	RF	0.460 (0.050)
75	NN	0.469 (0.015)
80	RF	0.478 (0.021)
85	NN	0.431 (0.045)
90	KNN	0.391 (0.050)
95	RF	0.408 (0.023)
100	**RF**	**0.584 (0.047)**

The value in bold is the best score achieved by the model among all possible combinations.

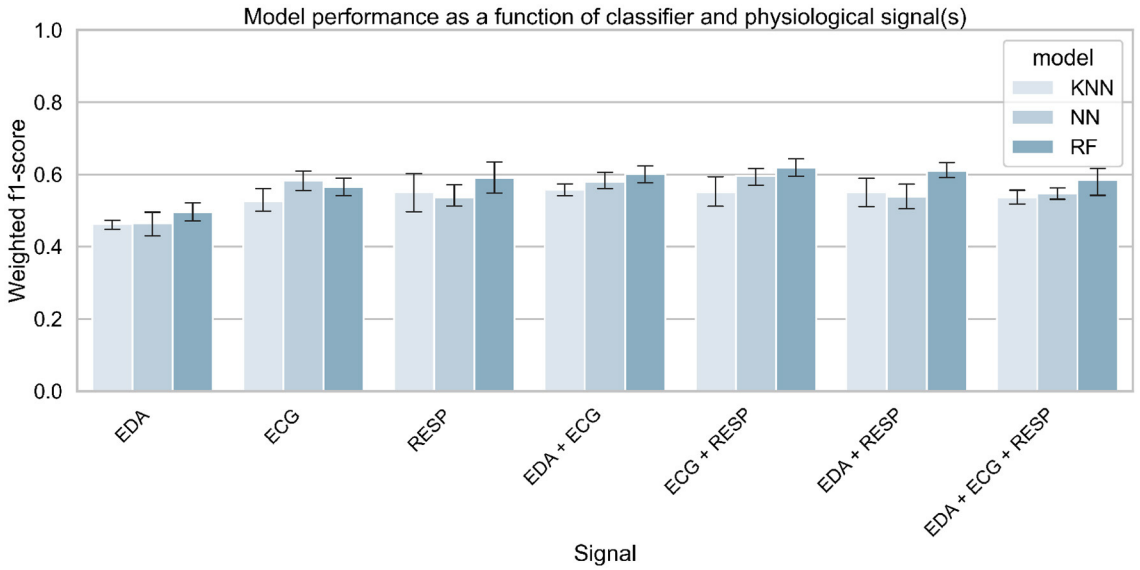


FIGURE 16 | Classification accuracy for predicting task modality (visual vs. auditory), as a function of selected physiological signals and classifier. A threshold for task performance of 100 was selected.

complex and recent models could be used, such as deep neural network architecture (Bagnall et al., 2016; Ismail Fawaz et al., 2019) or gradient boosted decision trees like XGB (Momeni et al., 2019). Data augmentation would hence be required to train models with deep architectures. This can be done using sliding windows to generate more training samples, or recent techniques of data augmentation such as Gaussian Mixture Models (GMMs) and Generative Adversarial Networks (GANs) (Hatamian et al., 2020). However, data augmentation using overlapping windows does not improve drastically models' performance to predict cognitive workload (Solovey et al., 2014; Momeni et al., 2019). This raises other research questions, such as the length of time windows used to generate the physiological indicators. Ninety second may not be the optimal time window for measuring mental load. The work of Meteier et al. (2021) shows that 4–5 min were optimal for measuring the mental load induced by a verbal task, while Solovey et al. (2014) found that 30 s gave the best results. This should be explored in future studies. The ultimate goal is to find the best trade-off between model accuracy and the time window used to predict mental load in a dynamic context such as automated driving. Another way to improve the results obtained would be to manipulate the MWL in the laboratory to limit the influence of external factors. However, the trained model would then be very efficient but less close to reality, which is less relevant for the concrete use of these intelligent models in our future cars.

TABLE 10 | Best score achieved by the model to predict task modality for each combination of physiological signals.

Selected signal	Best classifier	f1-score [Mean (SD)]
EDA	RF	0.496 (0.030)
ECG	NN	0.582 (0.035)
RESP	RF	0.591 (0.553)
EDA + ECG	RF	0.601 (0.030)
ECG + RESP	**RF**	**0.618 (0.030)**
EDA + RESP	RF	0.609 (0.027)
EDA + ECG + RESP	RF	0.584 (0.047)

The value in bold is the best score achieved by the model among all possible combinations.

7. CONCLUSION

This work studied the assessment of mental workload through physiological data in the specific context of automated driving. Three physiological signals (EDA, ECG, and respiration) from 80 subjects were collected during 1 h of conditionally automated driving in a simulator. The difficulty and modality of the task were experimentally manipulated with the N-back task. A wide range of physiological indicators was calculated from the signals collected during 15 task sequences (90 s each). Statistical analysis showed an effect of task difficulty on drivers' heart and respiratory rates, but not on the tonic level of the EDA. This could be explained by the low engagement of the drivers in the task or by the repeated requests to take over control during the experiment. A machine learning pipeline was set up, using a repeated 4-fold cross-validation approach with grid search on three algorithms. A random forest classified three different levels of mental workload with a f1-score of 0.713, using skin conductance and respiration as input signals. The drivers' subjective level of mental workload could be predicted with a mean absolute error of around 3 (on a scale of 0–20) using the three signals. In both the classification and regression tasks, the models' performance increased with task performance. This suggests the importance of controlling for task performance when using the dual-task paradigm to

experimentally manipulate workload. High engagement in the secondary task resulted in greater physiological activation and therefore helped the model to better classify or regress driver workload. In addition, the model had difficulty predicting the driver's state between monitoring the environment (no task) and performing a mild cognitive task (1-back task). The results suggest that these two tasks might induce a similar amount of physiological activation in drivers. As expected, classification of the task modality (visual or auditory) using physiological signals was not successful. Finally, the most important features in the classification process were extracted using a technique of explainable artificial intelligence. Physiological measures such as estimates of respiratory sinus arrhythmia and indicators of respiratory and heart rate variability were among the most relevant measures of mental workload, according to the results obtained in this study. This is consistent with previous literature and we suggest that these indicators should be used to assess the MWL of drivers in automated driving.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors upon request.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Internal Review Board of the Department of Psychology of the University of Fribourg. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AS, OA, EM, LA, and MW generated the idea to do this study. QM and AS created the experimental design and procedure. They also managed data collection. MC designed the driving scenario. QM and ED implemented the code to compute the indicators from the raw signals, and the classification and regression pipelines. All authors participated to the writing and revising processes.

FUNDING

This work has been supported and funded by the Hasler Foundation (Switzerland), in the framework of the AdVitam project.

ACKNOWLEDGMENTS

The authors would like to thank all the persons who contributed to this manuscript, especially Katharina Aigenbauer, Anika Dannemann, Sharon Guardini, and Aurelia Loser who helped authors for the experimental design and the data collection.

REFERENCES

Baek, H., Cho, C.-H., Cho, J., and Woo, J. (2015). Reliability of ultra-short-term analysis as a surrogate of standard 5-min analysis of heart rate variability. *Telemed. J. E-Health* 21, 404–14. doi: 10.1089/tmj.2014.0104

Baek, H., Lee, H. B., Kim, J. S., Choi, J., Kim, K. K., and Park, K. (2009). Nonintrusive biological signal monitoring in a car to evaluate a driver's stress and health state. *Telemed. J. E-Health* 15, 182–9. doi: 10.1089/tmj.2008.0090

Bagnall, A., Bostrom, A., Large, J., and Lines, J. (2016). The great time series classification bake off: an experimental evaluation of recently proposed algorithms. *Extended Version*. ArXiv, abs/1602.01711. doi: 10.1007/s10618-016-0483-9

Boucsein, W. (2012). *Electrodermal Activity*. Springer Science and Business Media, Boston, MA.

Boyce, P. R. (1974). Sinus arrhythmia as a measure of mental load. *Ergonomics* 17, 177–183. doi: 10.1080/00140137408931336

Brookhuis, K., and De Waard, D. (2001). "Assessment of drivers' workload: performance, subjective and physiological indices," in *Stress, Workload and Fatigue*, eds P. Hancock and P. Desmond (Mahwah, NJ: Lawrence Erlbaum Associates), 321–333.

Brookhuis, K., Waard, D., and Samyn, N. (2004). Effects of mdma (ecstasy), and multiple drugs use on (simulated) driving performance and traffic safety. *Psychopharmacology* 173, 440–445. doi: 10.1007/s00213-003-1714-5

Brookhuis, K. A., and de Waard, D. (2010). Monitoring drivers mental workload in driving simulators using physiological measures. *Accident Anal. Prevent*. 42, 898–903. doi: 10.1016/j.aap.2009.06.001

Bulmer, M., De Vaus, D. A., and Fielding, N. (2004). *Questionnaires*. London; Thousand Oaks, CA: Sage Publications. OCLC: 762283215.

Cacioppo, J., Tassinary, L., and Berntson, G. (2007). *Handbook of Psychophysiology, 3rd Edn*. Cambridge: Cambridge University Press.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Claesen, M., and De Moor, B. (2015). Hyperparameter search in machine learning. *arXiv:1502.02127 [cs, stat]*. arXiv: 1502.02127.

Collet, C., Clarion, A., Morel, M., Chapon, A., and Petit, C. (2009). Physiological and behavioural changes associated to the management of secondary tasks while driving. *Appl. Ergon*. 40, 1041–1046. doi: 10.1016/j.apergo.2009.01.007

Darzi, A., Gaweesh, S. M., Ahmed, M. M., and Novak, D. (2018). Identifying the Causes of drivers hazardous states using driver characteristics, vehicle kinematics, and physiological measurements. *Front. Neurosci.* 12:568. doi: 10.3389/fnins.2018.00568

De Waard, D. (1997). *The Measurement of Drivers Mental Workload* (Ph.D.) Thesis. Traffic Research Centre, University of Groningen, Haren, The Netherlands.

Dornhege, G., Millán, J. R., Hinterberger, T., McFarland D. J., and Müller, K-R. (2007). *Improving Human Performance in a Real Operating Environment through Real-Time Mental Workload Detection in Toward Brain-Computer Interfacing*. (Cambridge, MA: MIT Press), 409–422.

Engstrm, J., Johansson, E., and stlund, J. (2005). Effects of visual and cognitive load in real and simulated motorway driving. *Trans. Res. F Traffic Psychol. Behav*. 8, 97–120. doi: 10.1016/j.trf.2005.04.012

Ferreira, E., Ferreira, D., Kim, S., Siirtola, P., Roning, J., Forlizzi, J. F., and Dey, A. K. (2014). "Assessing real-time cognitive load based on psycho-physiological measures for younger and older adults," in *2014 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB)* (Orlando, FL: IEEE), 39–48.

Fisk, A. D., Derrick, W. L., and Schneider, W. (1986). A methodological assessment and evaluation of dual-task paradigms. *Curr. Psychol. Res. Rev*. 5, 315–327. doi: 10.1007/BF02686599

Gates, K. M., Gatzke-Kopp, L. M., Sandsten, M., and Blandon, A. Y. (2015). Estimating time-varying rsa to examine psychophysiological linkage of marital dyads. *Psychophysiology* 52, 1059–1065. doi: 10.1111/psyp.12428

Gawron, V. J. (2019). *Human Performance, Workload, and Situational Awareness Measures Handbook, 2-Volume Set*. Boca Raton, FL: CRC Press.

Grassmann, M., Vlemincx, E., von Leupoldt, A., Mittelstädt, J., and den Bergh, O. V. (2016). Respiratory changes in response to cognitive load: a systematic review. *Neural Plast*. 2016:8146809. doi: 10.1155/2016/8146809

Greco, A., Valenza, G., Lanata, A., Scilingo, E. P., and Citi, L. (2016). cvxEDA: a convex optimization approach to electrodermal activity processing. *IEEE Trans. Biomed. Eng*. 63, 797–804. doi: 10.1109/TBME.2015.2474131

Haapalainen, E., Kim, S., Forlizzi, J. F., and Dey, A. K. (2010). "Psycho-physiological measures for assessing cognitive load," in *Proceedings of the 12th ACM International Conference on Ubiquitous Computing-Ubicomp '10* (Copenhagen: ACM Press).

Hamilton, P. (2002). "Open source ECG analysis," in *Computers in Cardiology*, (Memphis, TN: IEEE), 101–104. doi: 10.1109/CIC.2002.1166717

Hart, S. G., and Staveland, L. E. (1988). "Development of NASA-TLX (Task Load Index): results of empirical and theoretical research," in *Advances in Psychology, volume 52 of Human Mental Workload*, eds P. A. Hancock and N. Meshkati (North-Holland), 139–183.

Hatamian, F. N., Ravikumar, N., Vesal, S., Kemeth, F. P., Struck, M., and Maier, A. K. (2020). "The effect of data augmentation on classification of atrial fibrillation in short single-lead ecg signals using deep neural networks," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Barcelona,: IEEE), 1264–1268.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (New York, NY: IEEE), 770–778. doi: 10.1109/CVPR.2016.90

Hidalgo-Muoz, A. R., Bquet, A. J., Astier-Juvenon, M., Ppin, G., Fort, A., Jallais, C., et al. (2019). Respiration and heart rate modulation due to competing cognitive tasks while driving. *Front. Hum. Neurosci*. 12:525. doi: 10.3389/fnhum.2018.00525

Hirsch, J., and Bishop, B. (1981). Respiratory sinus arrhythmia in humans: how breathing pattern modulates heart rate. *Am. J. Physiol*. 241, H620–9. doi: 10.1152/ajpheart.1981.241.4.H620

Hogervorst, M. A., Brouwer, A.-M., and van Erp, J. B. F. (2014). Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. *Front. Neurosci*. 8:322. doi: 10.3389/fnins.2014.00322

Huigen, E., Peper, A., and Grimbergen, C. A. (2002). Investigation into the origin of the noise of surface electrodes. *Med. Biol. Eng. Comput*. 40, 332–338. doi: 10.1007/BF02344216

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., and Muller, P.-A. (2019). Deep learning for time series classification: a review. *Data Min Knowl Discov*. 33, 917–963. doi: 10.1007/s10618-019-00619-1

Jaeggi, S. M., Buschkuehl, M., Etienne, A., Ozdoba, C., Perrig, W. J., and Nirkko, A. C. (2007). On how high performers keep cool brains in situations of cognitive overload. *Cogn. Affect. Behav. Neurosci*. 7, 75–89. doi: 10.3758/CABN.7.2.75

Kirchner, W. (1958). Age differences in short-term retention of rapidly changing information. *J. Exp. Psychol*. 55, 52–8. doi: 10.1037/h0043688

Kirk, R. (2013). "Latin square and related designs," in *Experimental Design: Procedures for the Behavioral Sciences, 4th Edn* (Thousand Oaks, CA: SAGE Publications, Inc.,).

Le, A. S., Aoki, H., Murase, F., and Ishida, K. (2018). A novel method for classifying driver mental workload under naturalistic conditions with information from near-infrared spectroscopy. *Front. Hum. Neurosci*. 12:431. doi: 10.3389/fnhum.2018.00431

Lewis, G. F., Furman, S. A., McCool, M. F., and Porges, S. W. (2012). Statistical strategies to quantify respiratory sinus arrhythmia: Are commonly used metrics equivalent? *Biol. Psychol*. 89, 349–364. doi: 10.1016/j.biopsycho.2011.11.009

Lundberg, S. M., and Lee, S.-I. (2017). "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems, Vol. 30*, eds I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Long Beach, CA: Curran Associates, Inc.).

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., et al. (2021). NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behav. Res. Methods*. 53, 1689–1696. doi: 10.3758/s13428-020-01516-y

Malik, M., and Terrace, C. (1996). Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. *Eur. Heart J*. 17, 354–381. doi: 10.1093/oxfordjournals.eurheartj.a014868

Mehler, B., Reimer, B., and Coughlin, J. (2012). Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task. *Hum. Factors* 54, 396–412. doi: 10.1177/0018720812442086

Mehler, B., Reimer, B., Coughlin, J., and Dusek, J. (2009). The impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *Trans. Res. Record* 2138, 6–12. doi: 10.3141/2138-02

Meteier, Q., Capallera, M., de Salis, E., Sonderegger, A., Angelini, L., Carrino, S., et al. (2020). "The effect of instructions and context-related information about limitations of conditionally automated vehicles on situation awareness," in *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI '20* (New York, NY: Association for Computing Machinery), 241–251.

Meteier, Q., Capallera, M., Ruffieux, S., Angelini, L., Abou Khaled, O., Mugellini, E., et al. (2021). Classification of drivers' workload using physiological signals in conditional automation. *Front. Psychol*. 12:268. doi: 10.3389/fpsyg.2021.596038

Momeni, N., Dell'Agnola, F., Arza, A., and Alonso, D. A. (2019). "Real-time cognitive workload monitoring based on machine learning using physiological signals in rescue missions," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Berlin: IEEE), 3779-3785.

Muth, E. R., Moss, J. D., Rosopa, P. J., Salley, J. N., and Walker, A. D. (2012). Respiratory sinus arrhythmia as a measure of cognitive workload. *Int. J. Psychophysiol*. 83, 96–101. doi: 10.1016/j.ijpsycho.2011.10.011

Pauzié, A. (2008). A method to assess the driver mental workload: The driving activity load index (dali). *IET Intell. Trans. Syst*. 2, 315–322. doi: 10.1049/iet-its:20080023

Paxion, J., Galy, E., and Berthelon, C. (2014). Mental workload and driving. *Front. Psychol*. 5:1344. doi: 10.3389/fpsyg.2014.01344

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res*. 12, 2825–2830. Available online at: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Plechawska-Wojcik, M., Tokovarov, M., Kaczorowska, M., and Zapa, D. (2019). A three-class classification of cognitive workload based on eeg spectral data. *Appl. Sci*. 9:5340. doi: 10.3390/app9245340

Reid, G., and Nygren, T. (1988). The subjective workload assessment technique: a scaling procedure for measuring mental workload. *Adv. Psychol*. 52, 185–218. doi: 10.1016/S0166-4115(08)62387-0

Roche, F., Somieski, A., and Brandenburg, S. (2019). Behavioral changes to repeated takeovers in highly automated driving: effects of the takeover-request design and the nondriving-related task modality. *Hum. Factors* 61, 839–849. doi: 10.1177/0018720818814963

Rubio, S., Diaz, E. M. C., Martin, J., and Puente, J. M. (2004). Evaluation of subjective mental workload: a comparison of swat, nasatlx, and workload profile methods. *Appl. Psychol*. 53, 61–86. doi: 10.1111/j.1464-0597.2004.00161.x

Salahuddin, L., Cho, J., Jeong, M. G., and Kim, D. (2007). "Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings," *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Lyon: IEEE), 4656–4659.

Singh, S. (2015). *Critical reasons for crashes investigated in the National Motor Vehicle Crash Causation Survey. (Traffic Safety Facts CrashoStats. Report No. DOT HS 812 115)*. Washington, DC: National Highway Traffic Safety Administration. Available Online at: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812115

Society of Automotive Engineers. (2018). *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. Version J3016_201806.* Washington, DC: SAE International. Available Online at: https://www.sae.org/standards/content/j3016_201806/

Solovey, E. T., Zec, M., Garcia Perez, E. A., Reimer, B., and Mehler, B. (2014). "Classifying driver workload using physiological and driving performance data: two field studies," in *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI '14* (Toronto, ON: ACM Press), 4057–4066.

Son, J., Oh, H., and Park, M. (2013). Identification of driver cognitive workload using support vector machines with driving performance, physiology and eye movement in a driving simulator. *Int. J. Precision Eng. Manufact.* 14, 1321–1327. doi: 10.1007/s12541-013-0179-7

Tsang, P., and Velazquez, V. L. (1996). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics* 39, 358–81. doi: 10.1080/00140139608964470

Wandtner, B., Schmig, N., and Schmidt, G. (2018). Effects of non-driving related task modalities on takeover performance in highly automated driving. *Hum. Factors* 60, 870–881. doi: 10.1177/0018720818768199

Wang, Z., Yan, W., and Oates, T. (2017). "Time series classification from scratch with deep neural networks: a strong baseline," in *2017 International Joint Conference on Neural Networks (IJCNN)*, Budapest, 1578–1585.

Wickens, C. D. (2008). Multiple resources and mental workload. *Hum. Factors* 50, 449–455. doi: 10.1518/001872008X288394

Wickens, C. D., Laux, L., Hutchins, S., and Sebok, A. (2014). Effects of sleep restriction, sleep inertia, and overload on complex cognitive performance before and after workload transition: a meta analysis and two models. *Proc. Hum. Factors Ergon.* 58, 839–843. doi: 10.1177/1541931214581177

Young, M. S., Brookhuis, K. A., Wickens, C. D., and Hancock, P. A. (2015). State of science: mental workload in ergonomics. *Ergonomics* 58, 1–17. doi: 10.1080/00140139.2014.956151

Zijlstra, F. R. H., and Van Doorn, L. (1985). *The construction of a scale to measure subjective effort*. Delft, Netherlands, 43, 124–139.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computer Science

ORIGINAL RESEARCH
published: 28 January 2022
doi: 10.3389/fcomp.2022.773256

Emotion Recognition in a Multi-Componential Framework: The Role of Physiology

Maëlan Q. Menétrey [1,2†], Gelareh Mohammadi [1,3], Joana Leitão [1] and Patrik Vuilleumier [1,4,5]*

[1] Laboratory for Behavioral Neurology and Imaging of Cognition, University of Geneva, Geneva, Switzerland, [2] Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, [3] School of Computer Science and Engineering, University of New South Wales, Sydney, NSW, Australia, [4] Department of Fundamental Neurosciences, University of Geneva, Geneva, Switzerland, [5] Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland

The Component Process Model is a well-established framework describing an emotion as a dynamic process with five highly interrelated components: cognitive appraisal, expression, motivation, physiology and feeling. Yet, few empirical studies have systematically investigated discrete emotions through this full multi-componential view. We therefore elicited various emotions during movie watching and measured their manifestations across these components. Our goal was to investigate the relationship between physiological measures and the theoretically defined components, as well as to determine whether discrete emotions could be predicted from the multicomponent response patterns. By deploying a data-driven computational approach based on multivariate pattern classification, our results suggest that physiological features are encoded within each component, supporting the hypothesis of a synchronized recruitment during an emotion episode. Overall, while emotion prediction was higher when classifiers were trained with all five components, a model without physiology features did not significantly reduce the performance. The findings therefore support a description of emotion as a multicomponent process, in which emotion recognition requires the integration of all the components. However, they also indicate that physiology *per se* is the least significant predictor for emotion classification among these five components.

Keywords: emotion, component model, autonomic nervous system, physiological responses, computational modeling

OPEN ACCESS

Edited by:
Kun Yu,
University of Technology Sydney, Australia

Reviewed by:
Jason Bernard,
McMaster University, Canada
Radoslaw Niewiadomski,
University of Trento, Italy

****Correspondence:***
Maëlan Q. Menétrey
maelan.menetrey@epfl.ch

†Present address:
Maëlan Q. Menétrey, Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *09 September 2021*
Accepted: *06 January 2022*
Published: *28 January 2022*

Citation:
Menétrey MQ, Mohammadi G, Leitão J and Vuilleumier P (2022) Emotion Recognition in a Multi-Componential Framework: The Role of Physiology. Front. Comput. Sci. 4:773256. doi: 10.3389/fcomp.2022.773256

INTRODUCTION

Emotions play a central role in human experience by changing the way we think and behave. However, our understanding of the complex mechanisms underlying their production still remains incomplete and debated. Various theoretical models have been proposed to deconstruct emotional phenomena by highlighting their constituent features, as well as the particular behaviors and particular feelings associated with them. Despite ongoing disagreements, there is a consensus at least in defining an emotion as a multicomponent response, rather than a unitary entity (Moors, 2009). This conceptualization concerning the componential nature of emotion is not only central in appraisal theories (Scherer, 2009) and constructivist theories (Barrett et al., 2007), but also found to some extent in dimensional (Russell, 2009) and basic categorical models

(Matsumoto and Ekman, 2009) that consider emotions as organized along orthogonal factors of "core affect" (valence and arousal), or as discrete and modular adaptive response patterns (fear, anger, etc.), respectively. Among these, appraisal theories, such as the Component Process Model (CPM) of emotion proposed by Scherer (1984), provide an explicit account of emotion elicitation in terms of a combination of a few distinct processes that evaluate the significance and context of the situation (e.g., relevance, novelty, controllability, etc.) and triggers a set of synchronized and interdependent responses at different functional levels in both the mind and body (Scherer, 2009). Hence, it is suggested that multiple and partly parallel appraisal processes operate to modify the motivational state (i.e., action tendencies such as approach, avoidance, or domination behaviors), the autonomic system (i.e., somatovisceral changes), as well as the somatic system (i.e., motor expression in face or voice and bodily actions). Eventually, synchronized changes in all these components—appraisal, motivation, physiology, and motor expression—may be centrally integrated in a multimodal representation (see **Figure 1**) that eventually becomes conscious and constitutes the subjective feeling component of the emotion (Grandjean et al., 2008).

Because the CPM proposes to define an emotion as a bounded episode characterized by a particular pattern of component synchronization, whereby the degree of coherence among components is a central property of emotional experience (Scherer, 2005a), it offers a valuable framework to model emotions in computationally tractable features. Yet, previous studies often relied on physiological changes combined with subjective feeling measures, either in the perspective of discrete emotion categories (e.g., fear, anger, joy, etc.) or more restricted dimensional descriptors (e.g., valence and arousal) (see Gunes and Pantic, 2010). As a consequence, such approaches have generally overlooked the full componential view of emotion. On the other hand, studies inspired by the appraisal framework have often analyzed emotional response with linear analyses and simple linear models (Smith and Ellsworth, 1985; Frijda et al., 1989; Fontaine et al., 2013). Yet, based on the interactional and multicomponent account of emotions in this framework (Sander et al., 2005), non-linear classification techniques from the field of machine learning may be more appropriate and indeed provide better performances in the discrimination of emotions (Meuleman and Scherer, 2013; Meuleman et al., 2019). However, in the few studies using such approaches, classification analyses were derived from datasets depicting the semantic representation of major emotion words, but participants were not directly experiencing genuine emotions.

In parallel, while physiology is assumed to be one of the major components of emotion, the most appropriate channels of physiological activity to assess or to differentiate a particular emotion is still debated (see Harrison et al., 2013). For example, dimensional and constructivist theories do not assume that different emotions present specific patterns of physiological outputs (Quigley and Barrett, 2014) or argue that evidence is minimal for supporting specific profiles in each emotional category, spotlighting the insufficient consistency and specificity in patterns of activation within the peripheral and central nervous systems (Wager et al., 2015; Siegel et al., 2018). It has also been advocated that an emotion emerges from an ongoing constructive process that involves a set of basic affect dimensions and psychological components that are not specific to emotions (Barrett et al., 2007; Lindquist et al., 2013). Therefore, the modulation of autonomic nervous system (ANS) activity might be tailored to the specific demand of a situation and not to a discrete emotion. Peripheral physiological state occurring during a given emotion type is therefore expected to be highly variable in its physiological nature.

In contrast, some authors argue that measures of peripheral autonomic activity may contain diagnostic information enabling the representation of discrete emotions, that is, a shared pattern of bodily changes within the same category of emotion that becomes apparent only when considering a multidimensional configuration of simultaneous measures (Kragel and LaBar, 2013). Because univariate statistical approaches, which evaluate the relationship between a dependent variable and one or more experimental independent variables, have shown inconsistent results in relating physiology measures to discrete emotions (Kreibig, 2010), the development of multivariate statistical approaches to discriminate multidimensional patterns offers new perspectives to address these issues. By assessing the correlation between both dependent and independent variables and by jointly considering a set of multiple variables, multivariate analyses can reveal a finer organization in data as compared with univariate analyses where variables are treated independently. Accordingly, several recent studies used multivariate techniques and described separate affective states based on physiological measures including cardiovascular, respiratory and electrodermal activity (Christie and Friedman, 2004; Kreibig et al., 2007; Stephens et al., 2010). Such results support theoretical accounts from both basic (Ekman, 1992) and appraisal models (Scherer, 1984) suggesting that information carried in autonomic responses is useful to distinguish between emotional states. In this view, by using the relationships between multiple physiological responses in different emotional situations, it should be possible to infer which emotion is elicited. However, empirical evidence suggests that it is still complicated to figure out from patterned physiological responses, whether ANS measures are differentiated among specific emotion categories or more basic dimensions (Mauss and Robinson, 2009; Quigley and Barrett, 2014). Moreover, it is often observed that self-reports of emotional experience discriminate between discrete emotions with a much better accuracy than autonomic patterns (Mauss and Robinson, 2009).

In sum, there is still no unanimous conclusion about distinguishable patterns of activation in ANS, due to the difficulty to identify and associate reliable response patterns to discrete emotions. As a consequence, the debate is not closed concerning the functionality of physiology during an emotional experience. Based on the CPM model, physiology is involved in shaping emotion and can contribute to differentiating emotion. However, while relevant, we hypothesize that the use of physiology alone is limited in discriminating emotion but could be better understood if integrated with the other major components of emotion. Therefore, to provide further insights about the contribution

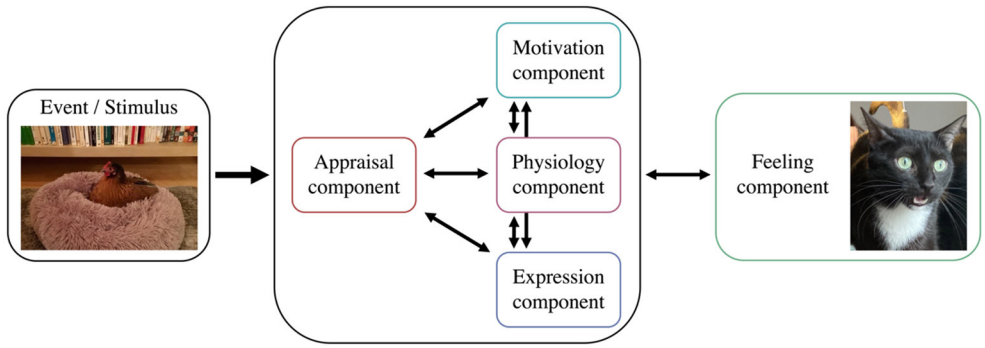


FIGURE 1 | Component model of emotion with five components. As suggested by Scherer (1984), it starts with an evaluation of an event (Appraisal component) which leads to changes in Motivation, Physiology and Expression components. Changes in all these four components modulate the Feeling component.

of physiology in emotion differentiation, we propose here to examine how a full componential model can account for the multiple and concomitant changes in physiological and behavioral measures observed during emotion elicitation. In addition, we examine the added information by each component and hypothesize that considering the synchronized changes in all components, the information in each component is already encoded in the other components. To the best of our knowledge, the present work represents one of the first attempts to investigate the componential theory by explicitly considering a combination of multiple, theoretically defined, emotional processes that occur in response to naturalistic emotional events (from cinematic film excerpts). By deploying a data-driven computational approach based on multivariate pattern classification, we aim at performing detailed analyses of physiological data in order to distinguish and predict the engagement of different emotion components across a wide range of eliciting events. On the grounds of such multicomponent response patterns, we also aim at determining to what extent discrete emotion categories can be predicted from information provided by these components, and what is the contribution of each component in such predictions. We hypothesize that a multicomponent account, as proposed by the CPM (Scherer, 1984, 2009), may allow us to capture the variability of physiological activity during emotional episodes, as well as their differentiation across major categories of emotions.

MATERIALS AND METHODS

Assuming that a wide range of emotional sequences will engage a comprehensive range of component processes, we selected a number of highly emotional film excerpts taken from different sources (see below). Physiological measures were recorded simultaneously during the initial viewing of movie clips, with no instructions other than be spontaneously absorbed by the movies. Participants were asked, during a second presentation, to fill out a detailed questionnaire with various key descriptors of emotion-eliciting episodes derived from the componential model (i.e., CoreGRID items) that assess several dimensions of appraisal, motivation, expression, physiology, and feeling experiences (Fontaine et al., 2013). We then examined whether the differential patterns of physiological measures observed across episodes could be linked to a corresponding distribution of ratings along the CoreGRID items, and whether the combined assessment of these items and physiological measures could be used together to distinguish between discrete emotions.

Population

A total of 20 French-speaking and right-handed students (9 women, 11 men) between 19 and 25 years old (mean age = 20.95, SD = 1.79) took part in the main study. All of them reported no history of neurological or psychiatric disorder, gave a written informed consent after a full explanation of the study and were remunerated. One participant completed only 2 sessions out of 4, but the data collected were nevertheless included in the study. This work was approved by Geneva Cantonal Research Committee and followed their guidelines in accordance with Helsinki declaration.

Stimuli Selection

To select a set of emotionally engaging film excerpts which could induce variations along different dimensions of the component model, a first preliminary study was conducted in separate study (for more details, see Mohammadi and Vuilleumier, 2020; Mohammadi et al., 2020). We selected a set of 139 film clips from the previous literature on emotion elicitation, matching in terms of time and visual quality (Gross and Levenson, 1995; Soleymani et al., 2009; Schaefer et al., 2010; Gabert-Quillen et al., 2015). Emotion assessment was collected in terms of discrete emotion labels and componential model descriptors. Initially, clips were evaluated over 14 discrete emotions (fear, anxiety, anger, shame, warm-hearted, joy, sadness, satisfaction, surprise, love, guilt, disgust, contempt, calm) based on a modified version of the Differential Emotion Scale (McHugo et al., 1982; Izard et al., 1993). For the component model, 39 descriptive items were selected from the CoreGRID instrument, capturing emotion features along the five components of interest: appraisal, motivation, expression, physiology, and feeling (Fontaine et al., 2013). This selection was performed based on the applicability to emotion elicitation scenarios while watching an event in a clip. The study was performed on Crowdflower, a crowdsourcing

platform, and a total number of 638 workers participated. Based on average ratings and discreteness, 40 film clips were selected for this study (for more details, see Mohammadi and Vuilleumier, 2020). Shame, warm-hearted, guilt and contempt were excluded from the list of elicited emotions because no clips received high ratings for these four emotions.

Finally, another preliminary study was conducted to isolate the highest emotional moments in each clip. To this aim, five different participants watched the full clips and rated the emotional intensity of the scene using CARMA, a software for continuous affect rating and media annotation (Girard, 2014). The five annotations were integrated to find the most intense emotional events in each time series.

The final list of film excerpts was thus represented by 4 clips for each of the 10 selected discrete emotions, with a total duration of 74 min (average length of 111 seconds per clip). Moreover, between 1 and 4 highly emotional segments of 12 seconds were selected in each film excerpt, for a total of 119 emotional segments. The list of the 40 selected films in our final dataset is presented in **Supplementary Table S1**. The duration, the initially assigned emotion label, and the number of highly emotional segments are indicated for each film excerpt.

Experimental Paradigm

The whole experiment consisted of four sessions scheduled on different days. Each session was divided into two parts, fMRI experiment and behavioral experiment, lasting for about 1 and 2 h, respectively. In the current study we focus only on the behavioral analysis and will not use the fMRI data. Stimuli presentation and assessment were controlled using Psychtoolbox-3, an interface between MATLAB and computer hardware.

During the fMRI experiment, participants were engaged in an emotion elicitation procedure using our 40 emotional film excerpts. No explicit task was required during this phase. They were simply instructed to let themselves feel and express emotions freely rather than controlling feelings and thoughts because of the experiment environment. Movies were presented inside the MRI scanner on an LCD screen through a mirror mounted on the head coil. The audio stream was transmitted through MRI-compatible earphones. Each session was composed of 10 separate runs, each presenting a film clip preceded by a 5-seconds instruction screen warning about the imminent next display and followed by a 30-seconds washout periods introduced as a low-level perceptual control baseline for the fMRI analysis (not analyzed here). Moreover, a session consisted of a pseudo-random choice of 10 unique film clips with high ratings on at least one of the 10 different pre-labeled discrete emotion categories (fear, anxiety, anger, joy, sadness, satisfaction, surprise, love, disgust, calm). This permitted to engage potentially different component processes in every session. To avoid any order effect, the presentation of all stimuli was counterbalanced.

The behavioral experiment was performed at the end of each fMRI session, in a separate room. Participants were let alone with no imposed time constraints to complete the assessment. They were asked to rate their feelings, thoughts, or emotions evoked during the first viewing of the film clips and advised not to report what might be expected to feel in general when watching such kinds of events. To achieve the emotion evaluation, the 10 film excerpts seen in the preceding session were presented on a laptop computer with LCD screen and headphones. However, the previously selected highly emotional segments (see "stimuli selection" above) were now explicitly highlighted in each film excerpts by a red frame surrounding the visual display. In order to ensure that emotion assessment corresponded to a single event and not the entire clip, the ratings were required right after each segment by pausing the clips. The assessment involved a subset of CoreGRID instrument (Fontaine et al., 2013), which is to date the most comprehensive attempt for multi-componential measurement in emotion. The set of 32 items (see **Table 1**) had been pre-selected based on their applicability to the emotion elicitation scenario with movies, rather than according to an active first-person involvement in an event. Among our set of CoreGRID items, 9 were related to the appraisal component, 6 to the expression component, 7 to the motivation component, 6 to the feeling component, and 4 to the bodily component. Participants had to indicate how much they considered that the description of the CoreGRID items correctly represented what they felt in response to the highlighted segment, using a 7-level Likert scale with 1 for "not at all" and 7 for "strongly."

Thus, each participant had to complete 119 assessments corresponding to 119 emotional segments. All responses were collected through the keyboard, for a total of 3,808 observations per participant (32 items $\times$ 119 emotional segments). Finally, they were also asked to label the segments by selecting one discrete emotion term from the list of 10 emotion categories. Therefore, the same segment may have been classified by participants into different emotion categories, and differently from the pre-labeled category defined during the pilot phase (where ratings were made for the entire film clip). In this study, we always used the subjectively experienced emotions reported by the participants as ground-truth labels for subsequent classification analyses. The frequency histogram showing the categorical emotions selected by the participants is presented in **Supplementary Figure S1**.

Physiological Data Acquisition

A number of physiological measures were collected during the first part of each session in the MRI scanner, including heart rate, respiration rate, and electrodermal activity. All the measures were acquired continuously throughout the whole scanning time. The data were first recorded with a 5,000 Hz sampling rate using the MP150 Biopac Systems software (Santa Barbara, CA), before being pre-processed with AcqKnowledge 4.2 and MATLAB 2012b.

Heart rate (HR) was recorded with a photoplethysmogram amplifier module (PPG100C). This single channel amplifier designed for indirect measurement of blood pressure was coupled to a TSD200-MRI photoplethysmogram transducer fixed on the index finger of the left hand. Recording artifacts and signal losses were corrected using endpoint function from AcqKnowledge, which interpolates the values of a selected impaired measure portion. Secondly, the pulse signal was

TABLE 1 | List of the 32 CoreGRID items.

Major components	CoreGrid items
Appraisal	
To what extent did you…	*1) think it was incongruent with your standards and ideas?*
	2) feel it was unpleasant for you?
	3) think it violated laws or socially accepted norms?
	4) think it was unpleasant for somebody (in the clip)?
	5) think it was important and relevant for the goals or needs of somebody?
	6) feel the event was unpredictable?
	7) feel the event occured suddenly?
	8) think the event was caused by chance?
	9) think the consequences were predictable?
Expression	
To what extent did you…	*10) press lips together?*
	11) close your eyes?
	12) show tears?
	13) have the jaw drop?
	14) have eyebrows go up?
	15) produce abrupt body movements?
Motivation	
To what extent did you…	*16) want to destroy something?*
	17) want to do damage, hit or say something that hurts?
	18) feel the urge to stop what was happening?
	19) want to undo what was happening?
	20) want the ongoing situation to last or be repeated?
	21) feel motivated to pay attention to what was going on?
	22) want to tackle the situation and do something?
Feeling	
To what extent did you….	*23) feel bad?*
	24) feel calm?
	25) feel good?
	26) feel strong?
	27) feel an intense emotional state?
	28) experience an emotional state for a long time?
Body	
To what extent did you…	*29) experience muscles tensing (whole body)?*
	30) have a feeling of a lump in the throat?
	31) have stomach troubles?
	32) feel warm?

Participants were asked to indicate on a 7-point Likert scale how much the descriptions represented what they felt.

exported to MATLAB and downsampled to 120 Hz. To remove scanner artifacts, a comb-pass filter was applied at 17.5 Hz. The pulse signal was then filtered with a band-pass filter between 1 and 40 Hz. Subsequently, the instantaneous heart rate was computed by identifying the peaks in the pulse signal, calculating the time intervals between them and converting this distance into beats per minute (BPM). The standard heart rate in humans goes from 60 to 100 bpm at rest. Hence, it was considered that a rate above 100 bpm was unlikely and the minimum distance between peaks will not exceed this limit. This automatic identification was manually verified by adding, changing or removing the detected peaks and possible outliers.

Respiration rate (RR) was measured using a RSP100C respiration pneumogram amplifier module, designed specifically for recording respiration effort. This differential amplifier worked with a TSD201 respiration transducer, which was attached with a belt around the upper chest near the level of maximum amplitude in order to measure thoracic expansion and contraction. Using a similar procedure as for HR preprocessing, the connect endpoint function of AcqKnowledge was first employed to correct manually the artifacts and losses of signal. After exporting the raw signal to MATLAB, it was downsampled to 120 Hz and then filtered with a band pass filter fixed between 0.05 and 1 Hz. Lastly, the signal was converted to breaths per minute using the same procedure as above. The standard respiration rate in human goes from 12 to 20 breaths per minute at rest. Since participants were performing a task inside a scanner which could be an unusual environment, the higher maximum rate was increased at 35 cycles per minute. Therefore, it was estimated that a rate above 35 was unlikely and the minimum distance between peaks will not exceed this limit. Again, this information was used in the automatic detection of the signal peaks. The respiration rate was then manually verified by looking at the detected signal peaks and corrected, with outliers being removed when it was necessary.

Electrodermal activity (EDA) was registered using an EDA100C electrodermal activity amplifier module, a single-channel, high-gain, differential amplifier designed to measure skin conductance via the constant voltage technique. The EDA100C was connected to Adult ECG Cleartrace 2 LT electrodes. Electrodes were placed on the index and the median fingers of the participants left hand. Following the manual correction of artifacts and losses of signal with the connect endpoint function on AcqKnowledge, the raw signal was exported to MATLAB. Similar to the two other physiological signals, the EDA signal was downsampled to 120 Hz. This signal, recorded by BIOPAC in microSiemes (μS), was then filtered with a 1 Hz low pass filter. An IIR (infinite impulse response) high-pass filter fixed at 0.05 Hz was applied to derive the Skin Conductance Responses (phasic component of EDA) representing the rapidly changing peaks, while a FIR (finite impulse response) low-pass filter fixed at 0.05 Hz was applied to derive the Skin Conductance Levels (tonic component of EDA) corresponding to the smooth underlying slowly-changing levels (AcqKnowledge 4 Software Guide, 2011).

Features and Normalization

MATLAB was used to select physiological values during the 12-s duration of high emotional segments. From these values, the means, variances, and ranges of each physiological

signals (HR, RR, phasic and tonic EDA) were calculated. We chose to focus specifically on the mean and variance of these physiological signals, as these are the most reliable and frequently reported features in studies associating discrete emotions and physiological responses (Kreibig, 2010). For HR and RR measures, respectively 4 and 17 responses during highly emotional segments had to be removed in one participant due to a corrupted signal, potentially induced by movements. For EDA, 267 values had to be removed due to temporary losses of signal, resulting in flat and useless measures. In particular, EDA responses of two subjects were completely removed as the EDA sensor could not capture their response. In order to handle the missing values in the physiological data, dropouts were replaced by mean value of the whole session during which the signal loss has happened (i.e., missing value imputation). Furthermore, the variance in physiological responses could be very large and different across participants. Because it was particularly important to reduce such variability in order to avoid inter-individual biases, all physiological measures were normalized within-subject using RStudio (1.1.383). To achieve this, standardized z-scores were calculated from the physiological data during the 4 sessions of each participant.

Regarding responses collected for the 32 CoreGRID items for each high emotional segment, a within-subject normalization into z-score was also performed. These normalized behavioral data and the discrete emotion labels selected by the participants for each emotional segment were incorporated to the related physiological measures. In the end, for each of the 119 emotional segments, we obtained a set of observations including 32 standardized CoreGRID items and 1 discrete emotion label, as well as 8 standardized physiological values calculated offline. However, the final dataset included 19 participants who attended all 4 sessions (19 x 119 = 2,261), while 1 participant completed 2 sessions out of 4 ($1 \times 55 = 55$). Also, for 11 participants, the assessment of one of the emotional segments did not get recorded due to a technical issue. Therefore, in total, there were (2,261 + 55 – 11 =) 2,305 sets of observations instead of the possible maximum of ($20 \times 119 =$) 2,380 (∼3% of points loss).

Predictive Analyses

To investigate the relationship between physiology and the component model descriptors, two analyses were performed. First, we examined whether the physiology measures allowed predicting component model descriptors and vice versa. Second, we assessed whether distinct features from the componential model allowed predicting discrete emotion categories and compared the value of different components for this prediction. For both analyses, multivariate pattern classifications using machine learning algorithms were undertaken to predict the variables of interest. Linear and non-linear classifiers including Logistic Regression (LR) and Support Vector Machine (SVM) with different kernels (linear, radial basis function, polynomial and sigmoid) were applied. All analyses were carried out using the RStudio statistical software, Version 1.1.383. Logistic regressions were conducted with the "caret" package, Version 6.0 and multinomial logistic regressions with the "nnet" package, Version 7.3. The binary and multiclass classifications using Support Vector Machine were conducted with the "e1071" package, Version 1.7.

First, the CoreGRID items were used as predictor variables to predict the dependent variable, which was either the mean or the variance of each physiological measure. To enable such analyses and to simplify the computational problem, the scores of the dependent variable were converted into two classes of "High" and "Low" using the median value across all the participants as a cutoff threshold. LR and SVM with linear and non-linear kernels using 10-fold cross-validation were applied. To guarantee test and training independence, each participant's assessment was included in either a test set or a training set. Conversely, similar analyses were carried out to determine whether physiology measures could encode the component model descriptors, but now using the physiology measures as independent variables in an attempt to predict the ratings of each CoreGRID item as either above or below the median.

Secondly, to examine the relationship between the component process model and discrete emotion types, multiclass classifications using SVM were performed on different combinations of CoreGRID items and physiological measures in order to predict specific emotion categories as labeled by the participants. Given the large number of classes and limited number of samples per class with too many predictors, a leave-one-subject-out cross-validation was used to guarantee a complete independence between the training and testing datasets.

For all analyses, we used a grid search method to optimize the parameters, but no significant improvement was observed, so the default parameters were kept. Moreover, SVM with radial basis function (RBF) kernel outperformed LR and other SVM models. Therefore, we will only report the result from SVM with RBF kernel.

It should be mentioned that the classes used for binary classifications were pretty balanced since they were defined based on the median value of the dependent variable, resulting in a distribution close to 50–50 split. However, in the case of multi-class classification, although the number of movie clips for each pre-labeled emotion category was balanced, the final dataset was not since we used the subjectively experienced emotions reported by each participant as emotional labels. This imbalance may have slightly affected the classification performance for some under- or over-represented categories. To account for an effect of class distribution, we reported the chance level in all comparisons as well as the confusion matrix.

RESULTS

Multivariate Pattern Analyses: Binary Classifications

Our first predictive analyses aimed to assess classification based on multivariate patterns using either physiology measures or CoreGRID items for different emotional movie segments. The variables to be discriminated were treated as binary dependent variables (High vs. Low), the classes being defined with respect to the median value. One of the main assumptions of the CPM is

that emotions rely on interdependent and synchronized changes within the five major components, suggesting that changes in any component might partly result from or contribute to changes in other components. Therefore, our initial exploratory analyses intended to examine whether physiological data (which represent an objective proxy to some of the CoreGRID items related to the physiology component) can be predicted using the CoreGRID items (which evaluate the five components of emotions), and vice versa. In other words, the idea was to investigate whether physiological changes are encoded in other emotional components. Based on the CPM, we expected to observe that physiological responses could predict not only physiology-related CoreGRID items, but also the items assessing the other components.

In the first instance, we deployed SVM classifications with 10-fold cross-validation to predict each physiological measure (mean or variance) as high or low from responses to the 32 CoreGRID items. Cross validation was applied to evaluate the generalizability of the results to some extent. This classifier yielded accuracies significantly greater than the chance rate of 50% for all physiological measures (**Table 2**). However, while these binary classifications were statistically significant and effect sizes were large, their discriminative performance remained weak (on average, about 58% of correct responses).

Conversely, using the same classification approach, the ratings of each CoreGRID item were predicted from the combination of physiological measures and physiology items in the CoreGRID questionnaire. Results from SVM showed that a majority of the CoreGRID items could be predicted significantly better than the chance level (**Table 3**), even though the classification accuracies were still relatively low (on average, about 55% of correct responses). The most reliable discrimination levels (highest t values relative to chance rate) were observed for appraisals and feelings of unpleasantness (55% of correct responses) as well as action tendencies (want to destroy / to do damage, 58% of correct responses).

Multivariate Pattern Analyses: Multiclass Classifications

Our second and main aim was to investigate whether discrete emotion categories as indicated by the participants can be predicted from ratings of their componential profiles. Here, we wanted to test whether the discrimination of discrete emotions is supported by one particular component (e.g., the appraisal component), distributed (equally) across the components, or requires the full combination of all components. The first step was to test how the entire data (physiological measures and behavioral responses) could predict discrete emotion labels. This more global pattern analysis for a multiclass variable required to go further than simple binary classification. To achieve this, a multiclass SVM classifications with leave-one-subject-out cross-validation was performed, taking the combination of within-subject normalized mean and variance measures from the four physiological signals and all behavioral responses to the CoreGRID items for the five emotional components as predictors. Applying the SVM classifiers (generated with training datasets) on separate testing datasets, we obtained an average accuracy rate of 45.4% in comparison to a rate of 17.6% for the chance level [$t_{(19)} = 10.852$, $p < 0.001$, Cohen's $d = 3.41$, 95% CI (1.75 5.06)] (**Figure 2A**).

The confusion matrix showed that five emotion categories (anger, calm, sadness and surprise) were correctly predicted more than half of the times, with an accuracy range from 55 to 59.4% (**Figure 2B**). By contrast, predictions were extremely unsuccessful for fear (misclassified as anxiety) and satisfaction (misclassified as joy or calm). However, it is worth noticing that these categorical emotions had a smaller number of instances since they were less often selected by the participants (see **Supplementary Figure S1**). Interestingly, incorrect predictions for these two emotions were still related to some extent to the target category. Indeed, mainly anxiety but also disgust and surprise were predicted instead of fear, whereas joy and calm were predicted instead of satisfaction. Love was also frequently misclassified as joy and calm.

Concomitantly, statistical measures allowing the assessment of prediction performance indicated that specificity and negative predictive value were particularly high for all emotions (**Figure 2C**). This suggests that the classification algorithm had a notable ability to correctly reject observations that did not belong to the emotion of interest, that is, to provide a good degree of certainty and reliability for true negatives. In contrast, sensitivity and positive predictive value were not as good and fluctuated substantially across the emotions, with the best performance for calm and the worst for fear and satisfaction (**Figure 2C**).

The second step consisted in investigating the added information brought by each component in the classification performance. To better identify the relation between discrete emotions and interactions of different component processes, we began by examining the effect on overall performance when one component was excluded. Five multiclass SVM classifications with leave-one-subject-out cross-validation were performed using different combinations of these components (**Figure 3**). In comparison to the accuracy rate of the complete model using physiology and all the CoreGRID items (45.4%), the accuracy rate of the reduced model without the body physiology component (4 CoreGRID items and all physiological measures) was lower but did not significantly change [45.2%, $t_{(19)} = -0.17$, $p = 0.866$, Cohen's $d = -0.01$, 95% CI (−0.13 0.11)]. These results suggest that information coming from features of the body physiology component in our study may have already been encoded in other components. On the other hand, with respect to performance with the full model, reduced models without the appraisal component [41.4%, $t_{(19)} = -4.598$, $p < 0.001$, Cohen's $d = -0.33$, 95% CI (−0.48 −0.18)], without the expression component [41.9%, $t_{(19)} = -4.358$, $p < 0.001$, Cohen's $d = -0.29$, 95% CI (−0.43 −0.15)], without the motivation component [43%, $t_{(19)} = -2.489$, $p = 0.022$, Cohen's $d = -0.21$, 95% CI (−0.37 −0.03)] or without the feeling component [44%, $t_{(19)} = -2.349$, $p = 0.029$, Cohen's $d = -0.12$, 95% CI (−0.22 −0.02)] were statistically less predictive, even though the effects sizes remained relatively small.

In addition, we examined the specific contribution in emotion classification of all major components. These contributions

TABLE 2 | Predictions of physiological changes from the 32 CoreGRID items.

Binary classification		Accuracy	Chance level	*t*	df	*p*-value	Cohen's D	95% CI
SVM classifier								
HR	Mean	0.59	0.5	7.175	9	**<0.001**	3.497	[0.76 6.23]
	Variance	0.55	0.5	4.308	9	**0.002**	1.480	[0.43 2.52]
RR	Mean	0.56	0.5	8.490	9	**<0.001**	2.950	[1.26 4.63]
	Variance	0.54	0.5	7.133	9	**<0.001**	2.246	[1.00 3.48]
Phasic EDA	Mean	0.58	0.5	8.549	9	**<0.001**	4.494	[0.81 8.17]
	Variance	0.61	0.5	12.641	9	**<0.001**	2.616	[1.70 3.53]
Tonic EDA	Mean	0.59	0.5	10.691	9	**<0.001**	5.231	[1.29 9.17]
	Variance	0.60	0.5	17.761	9	**<0.001**	7.062	[2.08 11.31]

Cross-subject binary SVM classifications. Accuracy rate represents the percentage of correct classifications. Paired t-tests were conducted to verify significant differences between SVM classifier and chance level. Bold values indicate statistically significant differences ($p < 0.05$). As estimates of effect size, we report Cohen's d and 95% confidence interval.

were assessed by predicting discrete emotion labels from each component separately. Five multiclass SVM classifications with leave-one-subject-out cross-validation were performed from the appraisal, expression, motivation, feeling, and body components (body items and all physiological measures). Since prediction performance from each of the five emotion components yielded accuracies significantly greater than the chance rate of 17.6% (**Figure 4**), we also analyzed the average sensitivity rate across the different classifiers in order to determine more precisely the power of each component to distinguish the different discrete emotions.

While the results above suggested that the percentage of correct predictions was generally similar regardless of the particular component used to train the classifiers, these additional analyses indicate that the pattern of features from specific components may yield a more reliable detection of particular emotions relative to others (**Figure 5A**). Moreover, it appeared also that some emotion classes were consistently well-discriminated by all components (e.g., calm and sadness), while others (fear, love, and satisfaction) were poorly predicted by any component. Conversely, some components could have more importance for particular emotions (e.g., surprise is well-predicted by appraisal features but not by the combination of body and physiological features, while motivation features seem best at predicting anger and joy).

Furthermore, since we observed that the body and physiology component was the least effective in discriminating discrete emotions, we also examined the sensitivity rates for each emotion and compared the performance of models using either the body-related CoreGRID items (i.e., subjective ratings), the physiological measures (i.e., objective recordings), or both information (**Figure 5B**). Consistent with the results above, we found that the sensitivity rates obtained with these models all showed a very poor discrimination for the majority of emotions, except for calm which was more successfully discriminated in comparison to predictions based on other components. Interestingly, the subjective body-related items from CoreGRID tended to surpass the objective physiology data [$t_{(19)} = 3.448$, $p = 0.002$, Cohen's $d = 0.88$, 95% CI (0.27 1.49)].

DISCUSSION

The CPM defines emotions by assuming that they are multicomponent phenomena, comprising changes in appraisal, motivation, expression, physiology, and feeling. A considerable advantage of this theory is that it offers the possibility of computational modeling based on a specific parameter space, in order to account for behavioral (Wehrle and Scherer, 2001; Meuleman et al., 2019) and neural (Leitão et al., 2020; Mohammadi et al., 2020) aspects of emotion in terms of dynamic and interactive responses among components. The current research applied multivariate pattern classification analyses for assessing the CPM framework with a range of emotions experienced during movie watching. Through this computational approach, we first investigated the links between physiology manifestations and the five emotion components proposed by the CPM to determine predictive relationships between them. Second, we investigated whether discrete emotion types can be discriminated from the multicomponent pattern of responses and assessed the importance of each component.

Assuming that physiological responses are intertwined with all components of emotion, we expected that ratings on the 32 CoreGRID features would carry information sufficient to predict corresponding physiological changes. Effectively, SVM classifications provided prediction accuracies significantly better than the chance level. However, information from the CoreGRID items did not allow a high accuracy, even though prediction was simplified by being restricted to a binary distribution. This modest accuracy may be explained by a great variability across participants, which could reduce the generalizability of classifiers when they were applied to all individuals rather than within subject. It might also reflect heterogeneity in intra-individual physiological responses among emotions with similar componential patterns. In parallel, the opposite approach to predict ratings of CoreGRID features based on the means and variances of physiological responses also yielded a performance significantly higher than chance level but still relatively low. Because each CoreGRID item focuses on quite specific behavioral features, it is however not surprising that the sole use

TABLE 3 | Predictions of individual CoreGRID item ratings from physiological responses.

Binary classification	Accuracy	Chance level	*t*	df	*p*-value	Cohen's D	95% CI
SVM classifier							
Appraisal							
Think it was incongruent with your standards and ideas	0.55	0.50	4.587	9	**0.001**	1.924	[0.43 3.41]
Feel it was unpleasant for you	0.54	0.50	4.319	9	**0.002**	2.322	[0.14 4.49]
Think it violated laws or socially accepted norms	0.55	0.50	6.011	9	**<0.001**	2.780	[0.63 4.92]
Think it was unpleasant for somebody (in the clip)	0.55	0.50	7.981	9	**<0.001**	4.409	[0.60 8.21]
Think it was relevant for the goals or needs of somebody	0.51	0.50	0.31	9	0.763	0.083	[−0.48 0.64]
Feel the event was unpredictable	0.52	0.50	2.900	9	**0.017**	1.352	[−0.002 2.71]
Feel the event occured suddenly	0.55	0.50	5.454	9	**<0.001**	2.298	[0.61 3.98]
Think the event was caused by chance	0.54	0.50	3.194	9	**0.011**	0.925	[0.19 1.65]
Think the consequences were predictable	0.55	0.50	7.949	9	**<0.001**	3.891	[0.88 6.90]
Expression							
Press lips together	0.55	0.52	5.251	9	**<0.001**	2.790	[0.32 5.25]
Close your eyes	0.56	0.54	4.019	9	**0.003**	1.279	[0.37 2.18]
Show tears	0.57	0.52	4.272	9	**0.002**	2.001	[0.29 3.70]
Have the jaw drop	0.56	0.51	4.929	9	**<0.001**	2.748	[0.18 5.31]
Have eyebrows go up	0.53	0.50	4.316	9	**0.002**	2.059	[0.28 3.83]
Produce abrupt body movements	0.56	0.54	3.069	9	**0.013**	1.106	[0.14 2.06]
Motivation							
Want to destroy something	0.58	0.54	6.843	9	**<0.001**	3.312	[0.72 5.90]
Want to do damage, hit or say something that hurts	0.58	0.52	5.223	9	**<0.001**	2.683	[0.36 4.99]
Urge to stop what was happening	0.53	0.51	1.862	9	0.095	0.727	[−0.19 1.65]
Want to undo what was happening	0.54	0.51	3.605	9	**0.005**	1.381	[0.25 2.50]
Want the ongoing situation to last or be repeated	0.54	0.53	0.925	9	0.378	0.397	[−0.53 1.33]
Motivated to pay attention to what was going on	0.53	0.50	2.719	9	**0.023**	1.476	[−0.17 3.12]
Want to tackle the situation and do something	0.54	0.53	1.329	9	0.216	0.469	[−0.31 1.25]
Feeling							
Feel bad	0.55	0.50	9.361	9	**<0.001**	4.547	[1.11 7.98]
Feel calm	0.53	0.50	2.614	9	**0.028**	1.197	[−0.06 2.45]
Feel good	0.54	0.50	3.537	9	**0.006**	1.252	[0.25 2.24]
Feel strong	0.59	0.54	4.406	9	**0.002**	1.792	[0.41 3.17]
Feel an intense emotional state	0.54	0.51	2.689	9	**0.024**	0.943	[0.05 1.82]
Experience an emotional state for a long time	0.52	0.50	2.323	9	**0.045**	0.78	[−0.02 1.59]
Body							
Experience muscles tensing (whole body)	0.54	0.51	4.202	9	**0.002**	1.905	[0.31 3.50]
Feeling of a lump in the throat	0.53	0.51	2.958	9	**0.015**	1.095	[0.11 2.07]
Have stomach troubles	0.54	0.54	1.006	9	0.340	0.428	[−0.50 1.36]
Feel warm	0.56	0.50	6.414	9	**<0.001**	2.549	[0.82 4.27]

Cross-subject binary SVM classifications. Accuracy rate represents the percentage of correct classifications. Paired t-tests were conducted to test for significant differences between SVM classifier and chance level. Bold values indicate statistically significant differences ($p < 0.05$). As estimates of effect size, we report Cohen's d and 95% confidence interval.

of physiology would be insufficient to precisely determine the ratings.

More importantly, if experiencing an emotion affects simultaneously more than one major component of emotion, one would expect that componential responses are clustered into qualitatively differentiated patterns (Scherer, 2005a; Fontaine et al., 2013). In the CPM view, an emotion arises when components are coherently organized and transiently synchronized (Scherer, 2005b). Accordingly, subjective emotion awareness might emerge as the conscious product of the feeling component generated by such synchronization (Grandjean et al., 2008). However, verbal accounts of conscious feelings may restrict the richness of emotional experience when using only declarative reports. Therefore, we anticipated that integrating the five components together into multivariate pattern analyses would provide higher accuracy rates in emotion prediction. This hypothesis was effectively confirmed, as the best prediction performances were obtained from non-linear multiclass SVM when the 32 CoreGRID items and the physiological measures were used all together in the model. Nevertheless, it is

important to note that through our one-component-out model comparisons, we found that the body and physiology component was negligible in the overall discrimination of discrete emotion labels. Indeed, prediction performances of a model without body and physiology features were not significantly different from those of the complete model, demonstrating that information derived from these data may have already been encoded in other components. However, further analyses could help to better confirm this observation.

Critically, the CPM assumes a strong causal link between appraisal and other components of emotion, since appraisal processes are the primary trigger of emotion and should account for a major part of qualitative differences in feelings (Moors and Scherer, 2013). For example, a cross-cultural study demonstrated that an appraisal questionnaire alone (31 appraisal features) could discriminate between 24 emotion terms with an accuracy of 70% (Scherer and Fontaine, 2013). In our study, we found all components provided relevant information. Moreover, although being the best predictor, the appraisal component did not provide significantly more information compared to the other components, except in comparison to the model using only body and physiology features. Overall, prediction from components did not significantly differ across emotions. At least three components were always predicting one emotion category within the same range of accuracy. These results are consistent with the assumption of a synchronized and combined engagement of these components during emotion elicitation. It is possible, however, that some results were affected by the uneven distribution of events across classes (**Supplementary Figure S1**), such as for calm (high representation) which stood out as the most recognizable state regardless of the component used, or for fear, love, and satisfaction (low representation) where all classifiers were poorly sensitive.

In line with our data indicating that physiological measures did not reliably discriminate among emotion categories, the relationship between physiological responses and emotions has long fueled conflicting views. Some authors claimed that there is no invariant and unique autonomic signature linked to each category of emotion (Barrett, 2006), or that physiological response patterns may only distinguish dimensional states (Mauss and Robinson, 2009). In contrast, because emotions imply adaptive and goal-directed reactions, they might trigger differentiated autonomic states to modulate behavior (Stemmler, 2004; Kragel and LaBar, 2013). In this vein, Kreibig (2010) reviewed the most typical ANS responses induced across various emotions and pointed to fairly consistent and stable characteristics for particular affective experiences, but without explicitly confirming a strict emotion specificity since no unique physiological pattern could be highlighted as directly diagnostic of a single emotion. It has also been shown that a single or small number of physiological indices are not able to differentiate emotions (Harrison et al., 2013). Our findings support this view by suggesting that a broader set of measures should be recorded to increase discriminative power, including physiology as well as other components.

Our study is not without limitations. First, statistical machine learning methods may be considered as uninterpretable black boxes. Indeed, SVM analysis gives no explicit clue on functional dimensions underlying classification performances. Second, these data-driven methods often need large amounts of data. We acquired data over a large number of videos and events covering a range of different emotions, but discrimination of specific patterns among the different emotion components was relatively limited with our sample of 20 participants. Third, although participants were asked to report their initial feelings during the first viewing, changes in emotional experience due to repetition or potential recall biases may not be completely excluded since each movie segment was played again before rating CoreGRID items. Fourth, we used a restricted number of CoreGRID items due to time and experimental constraints. It would certainly be beneficial to measure each component in more detailed ways by taking more features into account. For instance, motor behaviors (e.g., facial expressions) could be evaluated with direct measures such as EMG rather than self-report items. This could help to provide more objective and perhaps more discriminant measures, particularly concerning variations of pleasantness (Larsen et al., 2003). As another example, given that the appraisal component is crucial for emotion elicitation, a wider range of appraisal dimensions might allow a more precise discrimination of discrete emotions and physiological patterns. In the same way, it is also possible that the set of items selected from the original CoreGRID instrument may account for suboptimal discrimination performances (i.e., improving or degrading the classification of certain categories of emotion). Lastly, even though using film excerpts has many advantages (e.g., naturalistic and spontaneous emotion elicitation, control over stimuli and timing, standardized validation, and concomitant measurement of physiological responses), an ideal experimental paradigm should evoke first-person emotions in the participants to fully test the assumptions of the CPM framework. In other words, the only way to faithfully elicit a genuine emotion is to get participants to experience an event as pertinent for their own concerns, in order to activate the four most important appraisal features (relevance, implication, coping, normative significance) that are thought to be crucial to trigger an emotion episode (Sander et al., 2005). Viewing film excerpts is an efficient (Philippot, 1993) but passive induction technique and, therefore, the meaning of some appraisal components might be ambiguous or difficult to rate. As a result, subjective reports of behaviors and action tendencies were most likely different compared to what they would be for the same event in real life. We also cannot rule out that the correspondence found between CoreGRID items and discrete emotion labels could partially be affected by the order of the measures collected. For example, providing component-related ratings first may have activated knowledge of the emotion construct that was then used to select a label. Future research should develop more ecological scenarios that can be experienced by participants according to their self-relevance and followed by true choices of possible actions. For example, a sophisticated and ecological method was recently developed by connecting a wearable physiological sensor to a smartphone (Larradet et al., 2019). Upon detection of relevant physiological activity, the participant received a notification on her smartphone requesting to report her current emotional state. Alternatively, a study

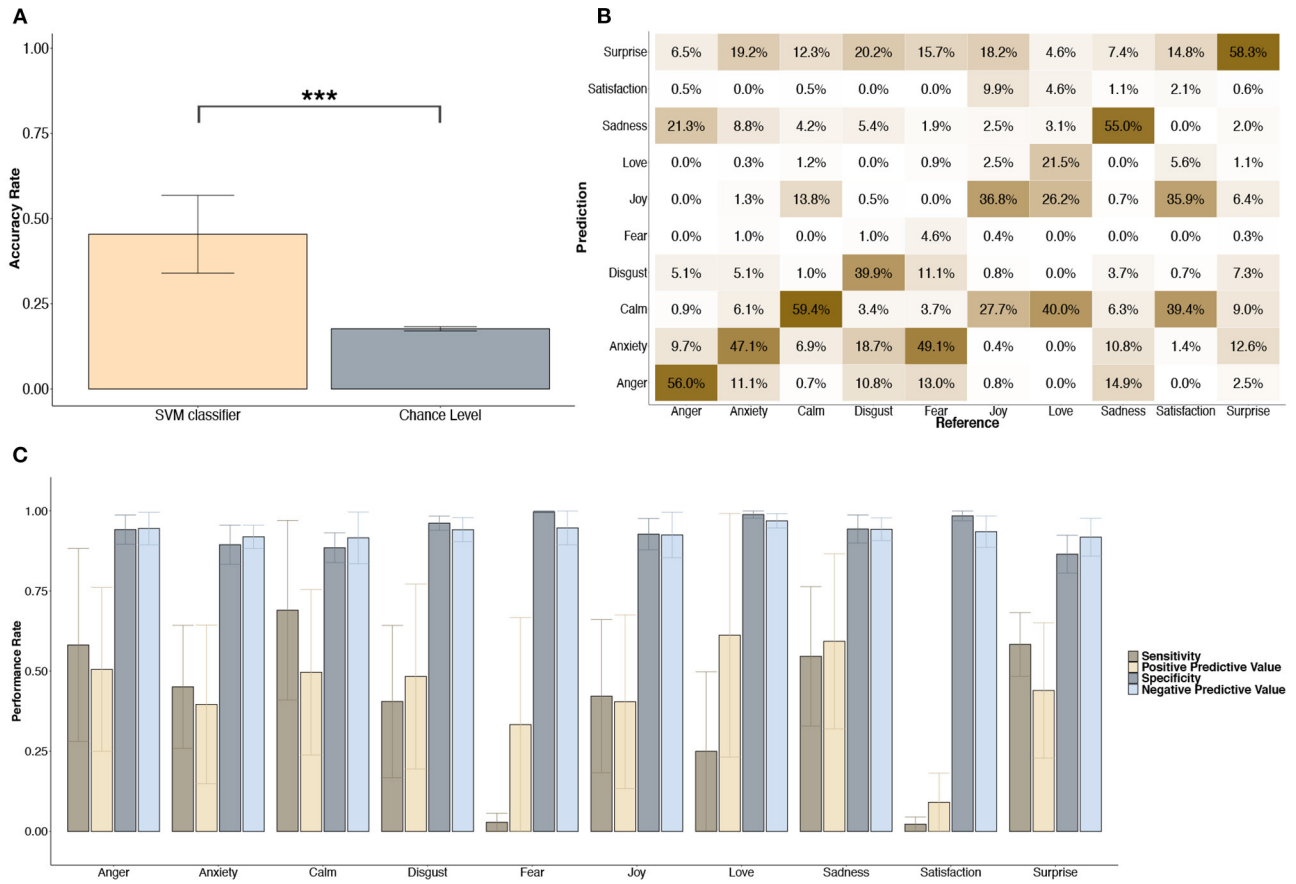


FIGURE 2 | (A) Predictions of discrete emotion labels from physiological measures and responses to the 32 CoreGRID features. Cross-subject multiclass SVM classification with leave-one-subject-out cross-validation. Accuracy rate represents the percentage of correct classifications. The error bars show the standard deviation. Paired *t*-tests were conducted to assess significant differences between SVM classifier and chance level, as highlighted by asterisks indicating the *p*-value (***$p < 0.001$). **(B)** Confusion matrix of emotion labels. The diagonal running from the lower left to the upper right represents the correct predicted emotion. **(C)** Statistical measures of classification performances across emotions. Average of statistical measures assessing the performances from the 20 classifiers. The error bars show the standard deviation.

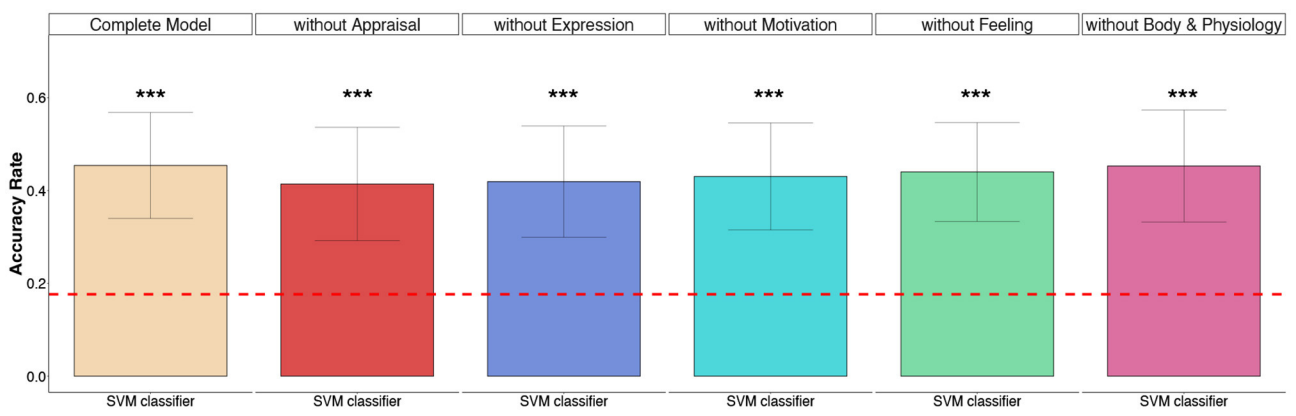


FIGURE 3 | Emotion classification using one-component-out models. Cross-subject multiclass SVM classifications with leave-one-subject-out cross-validation. Accuracy rate represents the percentage of correct classifications. The error bars show the standard deviation. Paired *t*-tests have been conducted and significant differences between SVM classifier and chance level are highlighted (asterisks indicate the *p*-value: ***$p < 0.001$).

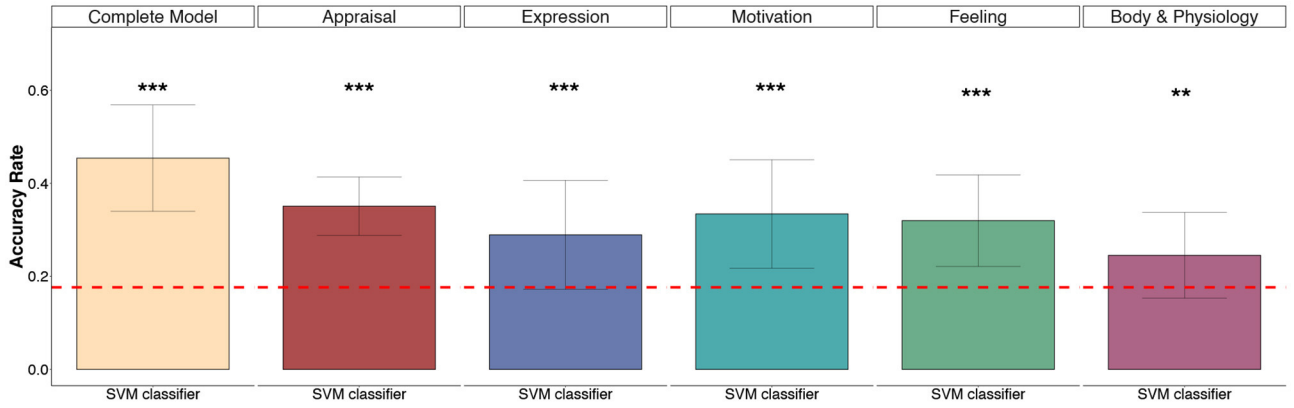


FIGURE 4 | Emotion classification using the components independently. Cross-subject multiclass SVM classifications with leave-one-subject-out cross-validation. Accuracy rate represents the percentage of correct classifications. The error bars show the standard deviation. Paired *t*-tests have been conducted and significant differences between SVM classifier and chance level are highlighted (asterisks indicate the *p*-value: $^{**}p < 0.01$, $^{***}p < 0.001$).

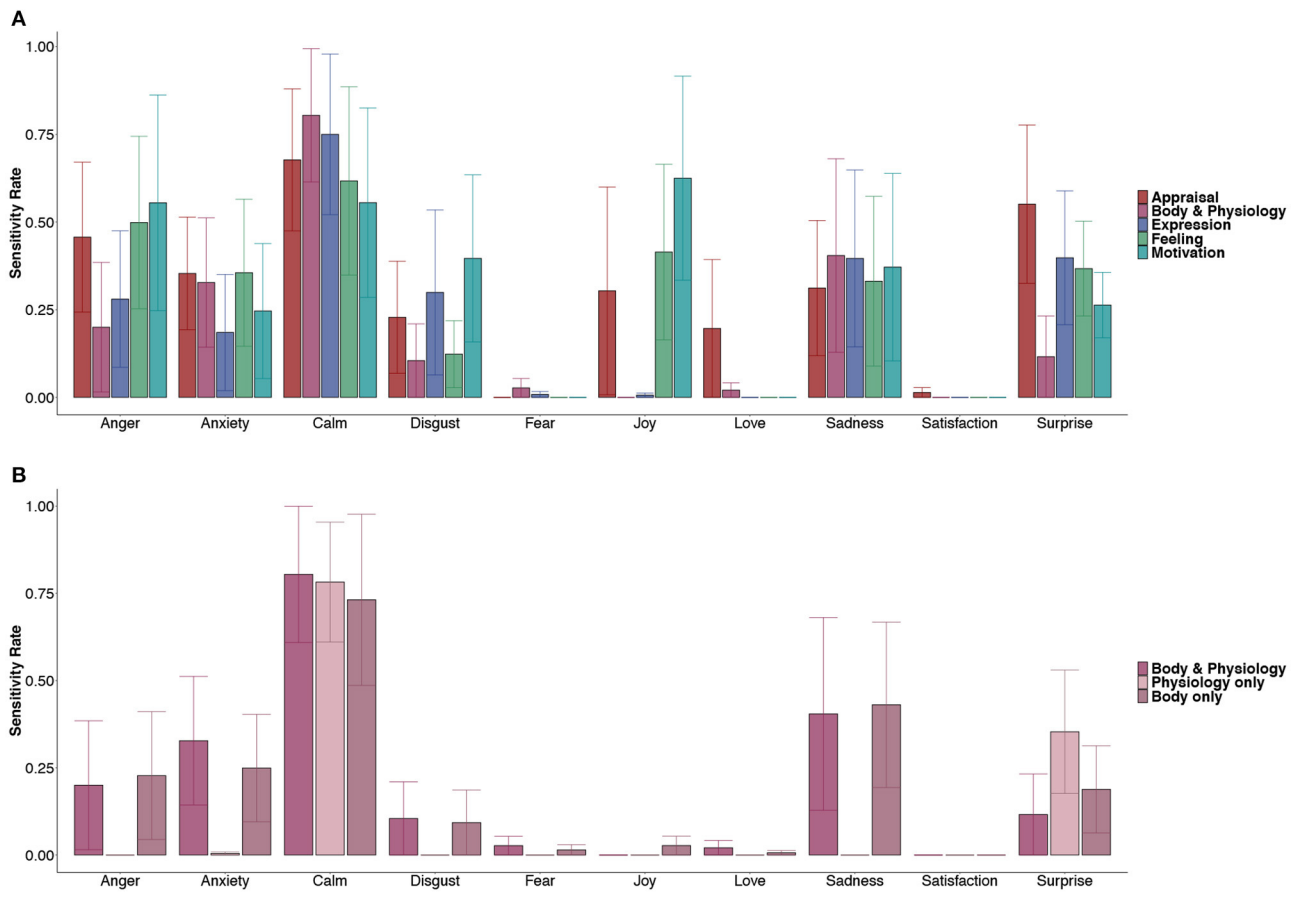


FIGURE 5 | (A) Sensitivity of emotion detection using the components independently. **(B)** Sensitivity of emotion detection using body and physiology components separately. Sensitivity rate represents the average of the sensitivity measures for each emotion label across the leave-one-subject-out cross-validation. The error bars show the standard deviation.

used virtual reality games to assess the CPM across various emotions (Meuleman and Rudrauf, 2018) and found that fear and joy were predicted by appraisal variables better than by other components, whereas these two emotions were generally poorly classified in our study. Other recent studies have also made use of (virtual reality) video games to assess appraisal and other

emotion components during brain imaging (Leitão et al., 2020) or physiological (Bassano et al., 2019) measurements.

CONCLUSION

Taken together, our results support the reliability and the interindividual consistency of CPM in the study of emotion. Multivariate pattern classification analyses generated results better than chance level (with statistical significance) to predict (1) changes in physiological measures from the 32 CoreGRID items, (2) ratings of the majority of CoreGRID items from physiological measures, and (3) discrete emotion labels that refer to conscious feelings experienced by the participants and presumably emerge from a combination of physiological and behavioral parameters. Overall, we observed, however, that physiological features were the least significant predictor for emotion classification. Yet, since our results also suggest that physiology was encoded within each of the other major components of emotion, they support the hypothesis of synchronized recruitment of all components during an emotion episode.

Further work is now required to determine why certain patterns of behavioral and physiological responses were misclassified into incorrect emotion categories and to study more deeply the links between different emotions. Similarly, it is also needed to explain the importance of various components in the recognition of different emotion categories. For instance, it would be valuable to determine whether poor discrimination stems from a too low sensitivity of the CoreGRID items and physiological measures or whether some categories of emotions simply cannot be differentiated into distinct entities with such methods, perhaps due to a high degree of overlap within the different components of emotion. Future developments allowing objective measures for each component during first-person elicitation paradigms are required to limit as much as possible the use of self-assessment questionnaires and ensure ecological validity. Overall, the current study opens a new paradigm to explore the depth of processes involved in emotion formation as well as a means of unfolding the necessary processes to be considered in developing a reliable emotion recognition system.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Geneva Cantonal Research Committee. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MQM, GM, and PV contributed to the study conception and study design. MQM and GM contributed to the data collection, data analysis and statistical analysis. JL also contributed to the data analysis. MQM, GM, JL, and PV contributed to the data interpretation, manuscript drafting and revision. All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This research was supported by a grant from the Swiss National Science Foundation (SNF Sinergia No. 180319) and the National Centre of Competence in Research (NCCR) Affective Sciences (under grant No. 51NF40- 104897). It was conducted on the imaging platform at the Brain and Behavior Lab (BBL) and benefited from the support of the BBL technical staff.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2022.773256/full#supplementary-material

REFERENCES

Barrett, L. F. (2006). Are emotions natural kinds? *Perspect. Psychol. Sci.* 1, 28–58. doi: 10.1111/j.1745-6916.2006.00003.x

Barrett, L. F., Mesquita, B., Ochsner, K. N., and Gross, J. J. (2007). The experience of emotion. *Annu. Rev. Psychol.* 58, 373–403. doi: 10.1146/annurev.psych.58.110405.085709

Bassano, C., Ballestin, G., Ceccaldi, E., Larradet, F., Mancini, M., Volta, E., et al. (2019). "A VR Game-based system for multimodal emotion data collection," in *12th Annual ACM SIGGRAPH Conference on Motion, Interaction and Games 2019* 38, 1–3. doi: 10.1145/3359566.3364695

Christie, I. C., and Friedman, B. H. (2004). Autonomic specificity of discrete emotion and dimensions of affective space: a multivariate approach. *Int. J. Psychophysiol.* 51, 143–153. doi: 10.1016/j.ijpsycho.2003.08.002

Ekman, P. (1992). An argument for basic emotions. *Cogn. Emot.* 6, 169–200. doi: 10.1080/02699939208411068

Fontaine, J. R. J., Scherer, K. R., and Soriano, C. (2013). "The why, the what, and the how of the GRID instrument," in *Components of emotional meaning: a sourcebook*, eds J. R. J. Fontaine, K. R. Scherer, and C. Soriano (Oxford: Oxford University Press), 83–97. doi: 10.1093/acprof:oso/9780199592746.003.0006

Frijda, N. H., Kuipers, P., and Ter Schure, E. (1989). Relations among emotion, appraisal, and emotional action readiness. *J. Pers. Soc. Psychol.* 57, 212. doi: 10.1037/0022-3514.57.2.212

Gabert-Quillen, C. A., Bartolini, E. E., Abravanel, B. T., and Sanislow, C. A. (2015). Ratings for emotion film clips. *Behav. Res. Methods* 47, 773–787. doi: 10.3758/s13428-014-0500-0

Girard, J. (2014). CARMA: software for continuous affect rating and media annotation. *J. Open Res. Softw.* 2, e5. doi: 10.5334/jors.ar

Grandjean, D., Sander, D., and Scherer, K. R. (2008). Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. *Consc. Cogn.* 17, 484–495. doi: 10.1016/j.concog.2008.03.019

Gross, J. J., and Levenson, R. W. (1995). Emotion elicitation using films. *Cogn. Emot.* 9, 87–108. doi: 10.1080/02699939508408966

Gunes, H., and Pantic, M. (2010). Automatic, dimensional and continuous emotion recognition. *Int. J. Synth. Emot.* 1, 68–99. doi: 10.4018/jse.2010101605

Harrison, N. A., Kreibig, S. D., and Critchley, H. D. (2013). "A two-way road," in *The Cambridge Handbook of Human Affective Neuroscience Efferent and Afferent Pathways of Autonomic Activity in Emotion*, eds J. Armony, and P. Vuilleumier (Cambridge: Cambridge University Press), 82–106. doi: 10.1017/CBO9780511843716.006

Izard, C. E., Libero, D. Z., Putnam, P., and Haynes, O. M. (1993). Stability of emotion experiences and their relations to traits of personality. *J. Pers. Soc. Psychol.* 64, 847–860. doi: 10.1037/0022-3514.64.5.847

Kragel, P. A., and LaBar, K. S. (2013). Multivariate pattern classification reveals autonomic and experiential representations of discrete emotions. *Emotion* 13, 681–690. doi: 10.1037/a0031820

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: a review. *Biol. Psychol.* 84, 394–421. doi: 10.1016/j.biopsycho.2010.03.010

Kreibig, S. D., Wilhelm, F. H., Roth, W. T., and Gross, J. J. (2007). Cardiovascular, electrodermal, and respiratory response patterns to fear- and sadness-inducing films. *Psychophysiology* 44, 787–806. doi: 10.1111/j.1469-8986.2007.00550.x

Larradet, F., Niewiadomski, R., Barresi, G., and Mattos, L. S. (2019). "Appraisal theory-based mobile app for physiological data collection and labelling in the wild," in *Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers* (London: ACM), 752–756. doi: 10.1145/3341162.3345595

Larsen, J. T., Norris, C. J., and Cacioppo, J. T. (2003). Effects of positive and negative affect on electromyographic activity over zygomaticus major and corrugator supercilii. *Psychophysiology* 40, 776–785. doi: 10.1111/1469-8986.00078

Leitão, J., Meuleman, B., Van De Ville, D., and Vuilleumier, P. (2020). Computational imaging during video game playing shows dynamic synchronization of cortical and subcortical networks of emotions. *PLoS Biol.* 18, e3000900. doi: 10.1371/journal.pbio.3000900

Lindquist, K. A., Siegel, E. H., Quigley, K. S., and Barrett, L. F. (2013). The hundred-year emotion war: are emotions natural kinds or psychological constructions? Comment on Lench, Flores, and Bench (2011). *Psychol. Bull.* 139, 255–263. doi: 10.1037/a0029038

Matsumoto, D., and Ekman, P. (2009). "Basic emotions," in *The Oxford Companion to Emotion and the Affective Sciences*, eds D. Sander and K. R. Scherer (New York, NY: Oxford University Press), 69–72.

Mauss, I. B., and Robinson, M. D. (2009). Measures of emotion: a review. *Cogn. Emot.* 23, 209–237. doi: 10.1080/02699930802204677

McHugo, G. J., Smith, C. A., and Lanzetta, J. T. (1982). The structure of self-reports of emotional responses to film segments. *Motivat. Emot.* 6, 365–385. doi: 10.1007/BF00998191

Meuleman, B., Moors, A., Fontaine, J., Renaud, O., and Scherer, K. (2019). Interaction and threshold effects of appraisal on componential patterns of emotion: a study using cross- cultural semantic data. *Emotion* 19, 425–442. doi: 10.1037/emo0000449

Meuleman, B., and Rudrauf, D. (2018). Induction and profiling of strong multi-componential emotions in virtual reality. *IEEE Trans. Affect. Comput.* 12, 189–202. doi: 10.1109/TAFFC.2018.2864730

Meuleman, B., and Scherer, K. R. (2013). Nonlinear appraisal modeling: an application of machine learning to the study of emotion production. *IEEE Trans. Affect. Comput.* 4, 398–411. doi: 10.1109/T-AFFC.2013.25

Mohammadi, G., Van De Ville, D., and Vuilleumier, P. (2020). Brain networks subserving functional core processes of emotions identified with componential modelling. bioRxiv [*Preprint*]. doi: 10.1101/2020.06.10.145201

Mohammadi, G., and Vuilleumier, P. (2020). "A multi-componential approach to emotion recognition and the effect of personality," in *IEEE Transactions on Affective Computing*. doi: 10.1109/TAFFC.2020.3028109

Moors, A. (2009). Theories of emotion causation: a review. *Cogn. Emot.* 23, 625–662. doi: 10.1080/02699930802645739

Moors, A., and Scherer, K. R. (2013). "The role of appraisal in emotion," in *Handbook of Cognition and Emotion*, eds M. Robinson, E. Watkins, and E. Harmon-Jones (New York, NY: Guilford Press), 135–155.

Philippot, P. (1993). Inducing and assessing differentiated emotion-feeling states in the laboratory. *Cogn. Emot.* 7, 171–193. doi: 10.1080/02699939308409183

Quigley, K. S., and Barrett, L. F. (2014). Is there consistency and specificity of autonomic changes during emotional episodes? Guidance from the Conceptual Act Theory and psychophysiology. *Biol. Psychol.* 98, 82–94. doi: 10.1016/j.biopsycho.2013.12.013

Russell, J. A. (2009). Emotion, core affect, and psychological construction. *Cogn. Emot.* 23, 1259–1283. doi: 10.1080/02699930902809375

Sander, D., Grandjean, D., and Scherer, K. R. (2005). A system approach to appraisal mechanisms in emotion. *Neural Netw.* 18, 317–352. doi: 10.1016/j.neunet.2005.03.001

Schaefer, A., Nils, F., Sanchez, X., and Philippot, P. (2010). Assessing the effectiveness of a large database of emotion-eliciting films: a new tool for emotion researchers. *Cogn. Emot.* 24, 1153–1172. doi: 10.1080/02699930903274322

Scherer, K. R. (1984). "On the nature and function of emotion: a component process approach," in *Approaches to Emotion*, eds K. R. Scherer and P. Ekman (Hillsdale, NJ: Erlbaum), 293–317.

Scherer, K. R. (2005a). What are emotions? And how can they be measured? *Soc. Sci. Inform.* 44, 695–729. doi: 10.1177/0539018405058216

Scherer, K. R. (2005b). "Unconscious processes in emotion: the bulk of the iceberg," in *Emotion and Consciousness*, eds L. F. Barrett, P. M. Niedenthal, and P. Winkielman (New York, NY: Guilford Press), 312–334.

Scherer, K. R. (2009). The dynamic architecture of emotion: evidence for the component process model. *Cogn. Emot.* 23, 1307–1351. doi: 10.1080/02699930902928969

Scherer, K. R., and Fontaine, J. R. J. (2013). "Driving the emotion process: the appraisal component, in Components of emotional meaning: a sourcebook," in *Components of Emotional Meaning*, eds J. R. J. Fontaine, K. R. Scherer, and C. Soriano (Oxford: Oxford University Press), 186–209. doi: 10.1093/acprof:oso/9780199592746.003.0013

Siegel, E. H., Sands, M. K., Van den Noortgate, W., Condon, P., Chang, Y., Dy, J., et al. (2018). Emotion fingerprints or emotion populations? A meta-analytic investigation of autonomic features of emotion categories. *Psychol. Bull.* 144, 343–393. doi: 10.1037/bul0000128

Smith, C. A., and Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *J. Pers. Soc. Psychol.* 48, 813. doi: 10.1037/0022-3514.48.4.813

Soleymani, M., Chanel, G., Kierkels, J. J., and Pun, T. (2009). Affective characterization of movie scenes based on content analysis and physiological changes. *Int. J. Semant. Comput.* 3, 235–254. doi: 10.1142/S1793351X09000744

Stemmler, G. (2004). "Physiological processes during emotion," in *The Regulation of Emotion*, eds P. Philippot and R. S. Feldman (Mahwah, NJ: Erlbaum).

Stephens, C. L., Christie, I. C., and Friedman, B. H. (2010). Autonomic specificity of basic emotions: evidence from pattern classification and cluster analysis. *Biol. Psychol.* 84, 463–473. doi: 10.1016/j.biopsycho.2010.03.014

Wager, T. D., Kang, J., Johnson, T. D., Nichols, T. E., Satpute, A. B., and Barrett, L. F. (2015). A bayesian model of category-specific emotional brain responses. *PLoS Comput. Biol.* 11, e1004066. doi: 10.1371/journal.pcbi.1004066

Wehrle, T., and Scherer, K. R. (2001). "Towards computational modeling of appraisal theories," in *Appraisal Processes in Emotion: Theory, Methods, Research*, eds K. Scherer, A. Schorr, and T. Johnstone (New York, NY: Oxford University Press), 350–365.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computer Science

ORIGINAL RESEARCH
published: 31 January 2022
doi: 10.3389/fcomp.2021.804917

STEP-UP: Enabling Low-Cost IMU Sensors to Predict the Type of Dementia During Everyday Stair Climbing

Catherine Holloway[1,2,3], William Bhot[2,3], Keir X. X. Yong[4], Ian McCarthy[5], Tatsuto Suzuki[5], Amelia Carton[4], Biao Yang[5], Robin Serougne[5], Derrick Boampong[5], Nick Tyler[6], Sebastian J. Crutch[4], Nadia Berthouze[1] and Youngjun Cho[1,2,3]*

[1] UCL Interaction Centre, University College London, London, United Kingdom, [2] Global Disability Innovation Hub, University College London, London, United Kingdom, [3] Department of Computer Science, University College London, London, United Kingdom, [4] Dementia Research Centre, Institute of Neurology, University College London, London, United Kingdom, [5] Pedestrian Accessibility and Movement Environment Laboratory, Department of Civil, Environmental and Geomatic Engineering, University College London, London, United Kingdom, [6] Universal Composition Laboratory, University College London, London, United Kingdom

OPEN ACCESS

Edited by:
Hans Hallez,
KU Leuven, Belgium

Reviewed by:
Peter Karsmakers,
KU Leuven Campus Geel, Belgium
Haoru Su,
Beijing University of Technology, China

***Correspondence:**
Catherine Holloway
c.holloway@ucl.ac.uk

Specialty section:
This article was submitted to Mobile and Ubiquitous Computing, a section of the journal Frontiers in Computer Science

Received: *29 October 2021*
Accepted: *28 December 2021*
Published: *31 January 2022*

Citation:
Holloway C, Bhot W, Yong KXX, McCarthy I, Suzuki T, Carton A, Yang B, Serougne R, Boampong D, Tyler N, Crutch SJ, Berthouze N and Cho Y (2022) STEP-UP: Enabling Low-Cost IMU Sensors to Predict the Type of Dementia During Everyday Stair Climbing. Front. Comput. Sci. 3:804917. doi: 10.3389/fcomp.2021.804917

Posterior Cortical Atrophy is a rare but significant form of dementia which affects people's visual ability before their memory. This is often misdiagnosed as an eyesight rather than brain sight problem. This paper aims to address the frequent, initial misdiagnosis of this disease as a vision problem through the use of an intelligent, cost-effective, wearable system, alongside diagnosis of the more typical Alzheimer's Disease. We propose low-level features constructed from the IMU data gathered from 35 participants, while they performed a stair climbing and descending task in a real-world simulated environment. We demonstrate that with these features the machine learning models predict dementia with 87.02% accuracy. Furthermore, we investigate how system parameters, such as number of sensors, affect the prediction accuracy. This lays the groundwork for a simple clinical test to enable detection of dementia which can be carried out in the wild.

Keywords: health—clinical, wearable computers, empirical study that tells us about people, lab study, dementia

INTRODUCTION

The rate of people living with dementia is increasing. Alzheimer's Disease (AD) is the most common cause of dementia and is often seen as simply part of the aging process and something which will affect most people (International Alzheimer's Disease, 2019) as the average living age increases. AD is a progressive disease which affects a person's memory and therefore their ability to conduct activities of daily living independently which decreases their quality of life (Gale et al., 2018). However, AD is not a single disease type, instead there is the typical presentation and a number of atypical presentations (Graff-Radford et al., 2021). Posterior Cortical Atrophy (PCA) is one such atypical presentation which typically results in "a progressive, often striking, and fairly selective decline in visual-processing skills and other functions that depend on the parietal, occipital, and occipitotemporal regions of the brain" (Crutch et al., 2012). Different types of AD may often be misdiagnosed until quite advanced. This is indeed the case for PCA where the atypical vision-based symptoms present themselves at an early age (typically emerging during 50–65 years old) leading

to a simple vision-problem diagnosis (Crutch et al., 2012). Therefore, it is important to develop methods that can identify AD regardless of its type so that people with rare forms can efficiently get the treatment they need. We do this by building on previous studies into everyday walking tasks detection.

People with typical Alzheimer's Disease (tAD) have characteristic issues when navigating their everyday environments (McCarthy et al., 2019) with a noticeable general decline in gait patterns (Valkanova and Ebmeier, 2017). Previous lab-based research has demonstrated differences in gait parameters such as step-time and walking speed between people with dementia and age-matched controls (Marquis et al., 2002; Waite et al., 2005; Wang et al., 2006; Verghese et al., 2007; Cedervall et al., 2014; Rosso et al., 2017). These studies indicate that the decline is linked to both phenotype and stage of the disease (Allali et al., 2016; Castrillo et al., 2016; Del Campo et al., 2016; McCarthy et al., 2019; Yong et al., 2020). Furthermore, a noticeable decline in gait is thought to predate other cognitive decline (Hall et al., 2000). Therefore, a decline in gait appears to be an appropriate biomarker for the detection of dementia (Montero-Odasso, 2016). However, it is important to move out of the laboratory setting to in-the-wild settings for clinical tools to better aid persons with disability (Holloway and Dawes, 2016). In the recent disability interactions manifesto (Holloway, 2019) the need for in-the-wild data collection was clearly stated. Such data sets were deemed essential to ensure future technologies to aid persons with disabilities such as dementia in living more independently.

This work is part of a wider investigation of gait and spatial navigation in people with dementia in a living lab environment, which specifically focuses on both people with tAD and PCA. Within the field of dementia there is a need for research in living labs, which move beyond highly controlled lab-based settings (Duff, 2020; Schneider and Goldberg, 2020). The living labs serve as a stepping-stone to full in-the-wild testing (Alavi et al., 2020). Full in-the-wild testing for dementia could reduce the stress of clinical tests for patients and allow for continuous monitoring of decline. Therefore, in this research we aim to pave the way to in-the-wild detection of dementia by discriminating people with dementia from controls in a living lab. Furthermore, we include a rare form of dementia—PCA—that is often missed by clinicians, demonstrating the benefits of this approach to dementia detection. The evidence-based discrimination of dementia, particularly its atypical presentations, not only has clinical applications, but also addresses a key desire of health and social-care professionals for better understanding of rarer presentations of dementia, for appropriate evidence-based assessment (McIntyre et al., 2019). Our apparatus uses low-cost, unobtrusive devices to discriminate dementia, which not only increases the applicability of our research, but also has not been achieved before. Furthermore, we analyze system parameters that led to accurate discrimination, which could aid future research seeking to extend this research or deploy it in the wild.

Therefore, in this paper we focus specifically on the question—can wearable, low-cost, unobtrusive devices be used to detect AD regardless of its presentation? In answering this question, we contribute the following:

- Demonstrate the feasibility of discriminating controls from people with two types of dementia [the more typical Alzheimer's disease (tAD) and a rare form of dementia—Posterior Cortical Atrophy (PCA)] in a simulated real-world environment—a staircase. To do this we analyzed data from a low-cost, IMU system using machine learning classifiers. The developed analysis software tools are available at https://github.com/williambhot/detecting_dementia_stairs.
- Examine different system parameters and the direction of traversal that promote accurate discrimination of dementia.
- Release a data set of IMU data from people with tAD, older adults and people with PCA to foster this work in the research community.
- Discuss use cases for the proposed system.

While the primary aim of this study is to discriminate both the rare PCA and more typical Alzheimer's Disease from healthy controls, we also analyze differences in the detection of these two types of the disease by analyzing the performance of a ternary model that seeks to discriminate the two types of dementia from each other as well as from controls.

We believe that this research, could provide a key stepping-stone in enabling potential applications in detecting dementia such as a screening tool for healthcare workers and practitioners, general self-screening and support tool. Nevertheless, further research would be required before this is possible to address some of the limitations of this study (such as generalization issues) and full in-the-wild testing. We discuss this further in section Discussion.

RELATED WORK

Posterior Cortical Atrophy

PCA is a rare early-onset syndrome which presents with visual complaints and is most commonly caused by Alzheimer's disease (AD) pathology. PCA has been identified as a distinct clinical syndrome as opposed to just AD with specific, noticeable visual deficits (Mendez et al., 2002). It also affects literacy, numeracy and gesture (Crutch et al., 2016). People with PCA, as opposed to typical AD (tAD) have better language and memory abilities (Crutch et al., 2016; Firth et al., 2019), but these come at the cost of a greater understanding of the disease and higher levels of depression (Mendez et al., 2002). Specific interventions need to be developed for people with PCA which help overcome the difficulties they face in visual tasks and help aid better mental health (Mendez et al., 2002). However, such interventions can only be developed once the disease has been detected and detection is often delayed due to the atypical symptoms compared to tAD and the early onset of the disease (Crutch et al., 2012; Graff-Radford et al., 2021).

Detecting rare forms of dementia like PCA with confidence is not an easy task. People often notice something going wrong with their eyes, e.g., being unable to see a shuttlecock once it has landed on the ground but being able to see it when in flight. The first stop for people following these visual oddities is to visit the optician or GP. It is rare that the symptoms as presented are immediately associated with a form of AD. More generally health and social care practitioners are often unaware of, and find it

difficult to appreciate that forms of dementia can affect people's visual abilities (McIntyre et al., 2019).

Dementia Detection

Previous work in the detection of dementia has ranged from mobile-based automatic speech recognition tools (e.g., Shibata et al., 2018; Tröger et al., 2018) to oculomotor performance during web browsing and multimodal interactions with computer avatars (Cano et al., 2017). However, to date these screening tools remain proofs of concept rather than clinical tools.

Previous research has identified that changes in gait are sensitive to dementia, even at early disease stages (Hall et al., 2000), and during the transitional stage between normal cognitive decline and dementia also known as Mild Cognitive Impairment (Gwak et al., 2018; Halloway et al., 2019; Schaat et al., 2020). It was found that a decline in gait predates observable cognitive changes associated with dementia, and gait continues to decline with the progression of dementia (Marquis et al., 2002; Waite et al., 2005; Wang et al., 2006; Verghese et al., 2007; Cedervall et al., 2014). By comparing the gait of healthy age-matched controls to that of people with dementia, clinical research has identified that changes in the pace, rhythm and variability of gait are associated with the decline into dementia (Verghese et al., 2007). Researchers have found people with dementia to have a lower natural walking speed (Marquis et al., 2002; Waite et al., 2005; Wang et al., 2006; Verghese et al., 2007), lower cadence, shorter stride length, shorter swing times and longer stance times as well as longer double support times (Verghese et al., 2007). Furthermore, studies have also shown that variability in gait is higher amongst people with dementia, who lack rhythmic and consistent gait (Verghese et al., 2007).

While previous clinical research has helped to identify the changes in gait that occur during the decline into dementia, this research has ignored two important factors that would allow such knowledge to be used for detection of the disease in the wild. Firstly, previous research relies heavily on experiments conducted in laboratory settings that do not mirror the complexities of the real-world environments through which people with dementia must navigate (McCarthy et al., 2019). These laboratory experiments usually involve monitoring the gait of participants while they walk along a straight, uninclined path for a short distance and use full biomechanics models to determine changes in gait (Marquis et al., 2002; Waite et al., 2005; Wang et al., 2006; Verghese et al., 2007). For example, many use electronic walkways with inbuilt pressure sensors (Verghese et al., 2007; Wittwer et al., 2013; Callisaya et al., 2017) or motion capture systems (Cedervall et al., 2014). The form factor, complicated setup procedures and price of these measurement systems limit their use in real world environments. Secondly, while some previous studies have analyzed different types of dementia (Mc Ardle et al., 2020), previous studies ignore the differences between types of dementia and either focus on one type of dementia (Wittwer et al., 2013; Cedervall et al., 2014; Callisaya et al., 2017) or consider dementia without looking at its type (Marquis et al., 2002; Wang et al., 2006). Furthermore, to our knowledge, gait of people with PCA has only been analyzed by previous research in this line of investigation (Carton et al., 2016; Ocal et al., 2017; Yong et al., 2018, 2020; McCarthy et al., 2019; McCarthy et al., Unpublished[1]). This research has found that some patients with dementia show a consistent pattern of hesitation (which can be identified from step times) when navigating complex routes (McCarthy et al., 2019; Yong et al., 2020). However, it was not possible within that task to identify patterns which could be used for predictive purposes. We believe that the regular pattern offered by stairs will help to regularize these irregularities within the gait pattern which would then allow for successful detection of tAD and PCA. Once the feasibility of this approach is established, it will enable a low-cost detection device to be added to footwear. This could enable the detection of dementia in the wild, minimizing stressful laboratory tests, and promoting data-driven methods for appropriate detection of dementia for both typical AD and the rarer PCA. Furthermore, the ability of the device to detect the typical Alzheimer's disease (tAD) provides the final product with a much wider number of use cases. The unobtrusive, low-cost nature of such a device enables its deployment in high-risk populations to continuously monitor changes in risk of developing dementia.

MATERIALS AND METHODS

In this section, we present the proposed STEP-UP framework and technical details.

Data Collection Protocol

Participants' gait was monitored using Inertial Measurement Units (IMUs) while they climbed a staircase in the living lab environment. This living lab was co-designed by clinical, engineering and computer science researchers, with inputs from patients. The IMUs used were MTw (Xsens Technologies B.V., The Netherlands). They are comprised of an accelerometer, a gyroscope, and a magnetometer (however, the magnetometer was not used for this study). Each participant had a sensor attached to the outside of each heel with the long axis being horizontal, as well as a sensor on the back of the pelvis attached orthogonally to the sensors on the heels (**Figure 1**). Participants were asked to walk up or down a short flight of stairs consisting of four steps (the dimensions of each step were 23 × 112 × 25 cm, H × W × D) (**Figure 1**) in a variety of environmental conditions. These environmental conditions included different lighting levels (low: 20 lux; high: 190 lux) and either the presence or absence of visual cues (i.e., hazard tape over the edge of steps). Each participant was asked to attempt 16 versions of the trial (twice for each combination of conditions—dim light/bright light, visual cues/no visual cues—in the upwards and downwards direction). No constraints were imposed on the way of descending or ascending the stairs. The ordering of trials was randomized for each participant (see **Figure 2A**).

Participants

Participants were from one of three groups—the group with PCA [containing 11 participants—6 female and 5 male—of age 64.6

[1]McCarthy, I. D., Suzuki, T., Holloway, C., Poole, T., Frost, C., Carton, A., et al. (Unpublished). *Gait Assessment of People with Alzheimer's Disease Traversing Routes of Varying Complexity*.

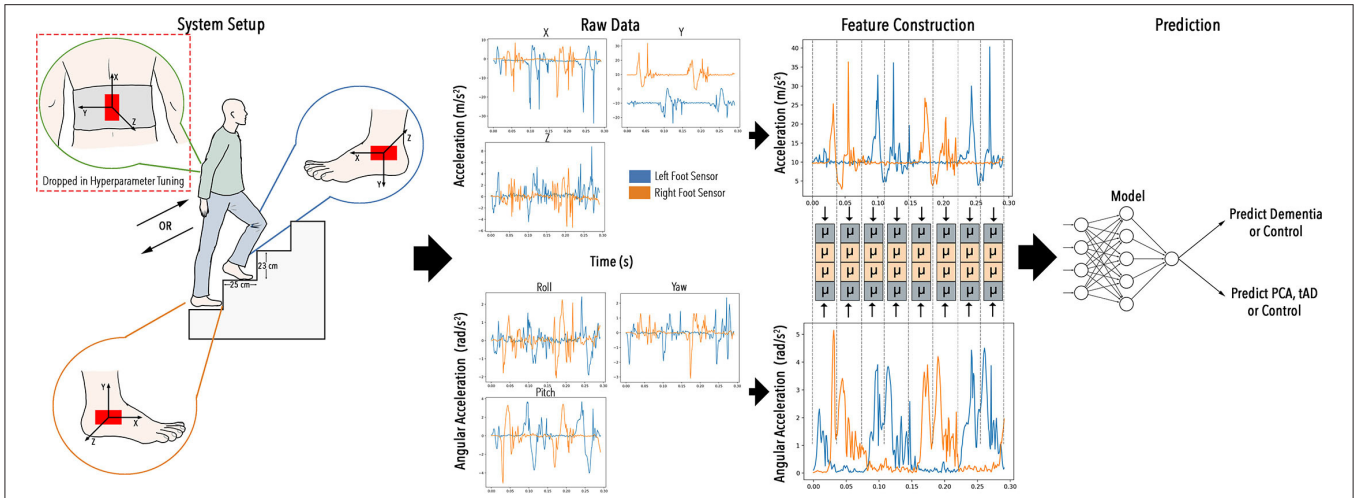


FIGURE 1 | Project STEP-UP: to enable low-cost and wearable IMU sensors to infer dementia types in the wild whilst climbing stairs.

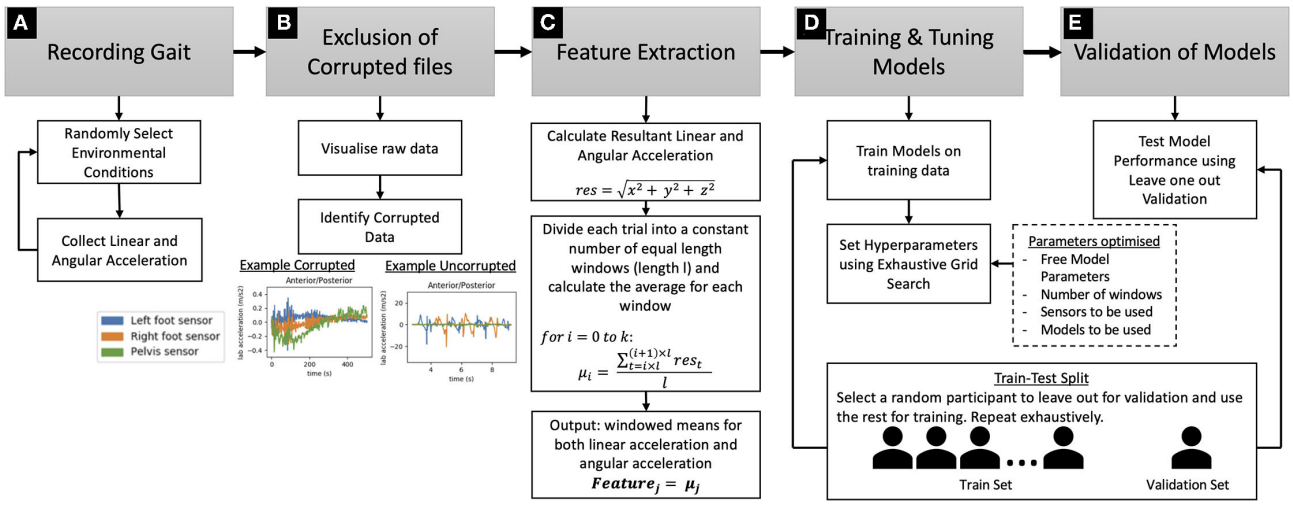


FIGURE 2 | Technical details of Step-up framework: **(A)** Gait Recording procedure using wearable IMUs, **(B)** Procedure for the exclusion of corrupted files, **(C)** Feature extraction procedure using windowed averaging, **(D)** Model training and tuning procedure, **(E)** Validation procedure using Leave One Out Validation. The procedure for splitting the dataset into training and testing sets is shown under **(D,E)**.

± 5.9 years, height 168.92 ± 6.49 cm, weight 68.22 ± 13.31 kg, with Mini Mental State Examination (MMSE) score 18.6 ± 6.1], the group with tAD (containing 10 participants—6 female and 4 male—of age 66.2 ± 5.0 years, height 167.91 ± 11.82 cm, weight 66.21 ± 5.03 kg, with MMSE score 18.6 ± 5.0) and the control group consisting of age matched participants with no diagnosed form of dementia (containing 14 participants—6 female and 8 male—of age 64.2 ± 4.1 years, height 172.36 ± 13.21 cm, weight 73.23 ± 15.23 kg). The experimental design of having a control group of healthy age-matched participants is the standard experimental protocol used in this field (Callisaya et al., 2017; McCarthy et al., 2019). MMSE tests were only conducted on people with dementia, and not on control participants. One-way ANOVAs demonstrated that there were no statistically significant differences between the groups in age [$F_{(2,32)} = 0.506$; $p = 0.61$], weight [$F_{(2,30)} = 0.404$; $p = 0.67$] or height [$F_{(2,31)} = 0.580$; $p = 0.57$]. Furthermore, a student's t-test showed that there was no difference between MMSE scores for participants in the PCA and tAD conditions [$t_{(18)} = 0$; $p = 1$]. Ethical approval for the study was provided by the National Research Ethics Service Committee London Queen Square, and written informed consent was obtained from all 35 participants.

Pre-processing and Classification Strategy

The data was processed in Python 3.7 (Python Programming Language, RRID:SCR_008394) using standard data processing libraries including NumPy (NumPy, RRID:SCR_008633), SciPy (SciPy, RRID:SCR_008058),

TABLE 1 | The dataset before removing the corrupted files compared to the dataset after this removal.

Group	Number of trials (before removal)	Number of trials (after removal)
Control	208	207
PCA	159	150
tAD	160	159
Total	527	516

Pandas (Pandas, RRID:SCR_018214), Matplotlib (MatPlotLib, RRID:SCR_008624) and Scikit Learn (scikit-learn, RRID:SCR_002577). The data pre-processing and classification strategy is shown in **Figure 2**. This process included hyperparameter optimization on the models to select the best parameters and analysis of how direction of traversal and different system setups affected the performance of this model. This section summarizes the methods we used to achieve this. The software tools we developed are released to foster this work in the research community (https://github.com/williambhot/detecting_dementia_stairs).

Exclusion of Participants

On visualizing the IMU data—acceleration and gyroscope data—data for some trials was found to be corrupted. Visualizing the raw data from these trials showed only noise and no evidence of cyclic, step-like motion (**Figure 2B**). Therefore, these trials were removed from further analysis.

This resulted in the removal of 11 trials from a total of 527 trials (**Table 1**). After removing excluded trials, 40.12% of trials were controls, 29.07% were in the PCA condition and 30.81% were in the tAD condition. Up-sampling was conducted on the trials from the different conditions before training any models, so that the models did not overfit to these differences in the frequencies in the groups.

Dead Reckoning and Gait Parameters

Initially we tried to calculate velocity and displacement from the IMU data using a dead-reckoning technique with a zero-offset to account for sensor drift (Ojeda and Borenstein, 2007; Park and Suh, 2010). Using this we calculated gait parameters that have been previously associated with dementia such as lower walking speed (Marquis et al., 2002; Waite et al., 2005; Wang et al., 2006; Verghese et al., 2007) and shorter stride length (Verghese et al., 2007). However, we found that in our current set up it was not possible to conduct dead reckoning with a high enough degree of accuracy for calculating the gait parameters required. We attribute this to the experimental setup as well as issues with controlling the task across participants, especially those with more advanced dementia. See the discussion for more details on this.

Lower-Level Features

Considering the difficulty of conducting dead-reckoning and calculating gait parameters in a system designed to be useable in the real world, we propose more low-level features that, from a low-cost IMU system, can be more easily designed for real-world use. This involved calculating the vector length of the 3d linear and angular acceleration to obtain the resultant linear and angular acceleration (see **Figure 2C**):

$$R = \sqrt{x^2 + y^2 + z^2}$$

These two signals—resultant linear acceleration and resultant angular acceleration—were then split into a constant number of windows (k) and the averages of each window (μ_i where i is the number of the window) were used as the features. The windows were calculated in the following way—across the entire dataset, the same number of windows (k) were used and in a single trial these windows were of the same length (l), however, across multiple trials window length was different (see **Figure 2C**):

$$\mu_i = \frac{\sum_{t=i\times l}^{(i+1)\times l} R_t}{l}$$

Where $i \in [0, k)$ is the number of the current window varying between 0 and $k-1$, k is the total number of windows and t is the current sample for the linear or angular acceleration.

These windowed averages were used as the feature values, allowing a constant number of features for each trial, while providing the model with information from different sections of the trial. The primary reason for using this approach was to have a constant number of features for all trials, which is required by many Machine Learning models. The number of windows was set using hyperparameter optimization. Specifically, different numbers of windows were experimented with, but it was found that models using a multiple of four windows achieved a higher performance than others and specifically eight windows yielded the best performance (**Figure 3**). One reason for this could be that there were four steps in the staircase and, therefore, setting the number of windows to a multiple of four provides an approximate way to separate the data based on steps, assuming each step is traversed in approximately the same amount of time in a single trial. However, every participant did not take the same amount of time on each step, and several participants waited for a while on some steps. Therefore, for these participants segmenting the data in this way would not segment the trial by steps. Nevertheless, this was not our motivation for doing this, but rather it was to segment the trial into an equal number of windows so that models that required a fixed number of features could be employed.

Machine Learning Models

We assessed the ability of different machine learning models to classify the data, including decision trees (Random Forest and Gradient Boosting Models) and Multi-Layer Perceptron (MLP) models. To this end, we fit the models to the data and evaluated the models' ability to generalize by testing it on unseen data (see the following section). Furthermore, we chose the parameters of this model through hyper-parameter optimization discussed later (see **Figure 2D**).

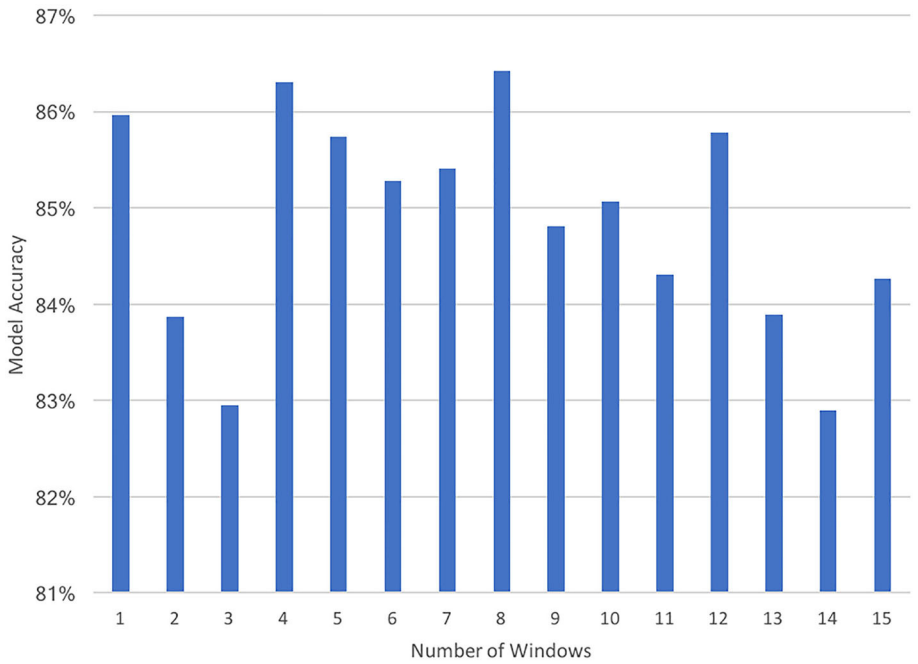


FIGURE 3 | A plot of the prediction accuracies of the Random Forest Classifier when using different numbers of windows (1–15) for constructing the features.

Two variants of all the models were fit to the data—a binary model to discriminate dementia from control participants and a ternary model to discriminate between controls, tAD and PCA participants. While we were able to discriminate people with dementia from control participants, we were unable to discriminate PCA from tAD with high accuracy (see section Results for more details). We suggest that this is because the gait of the two types of dementia was similar to each other and therefore could not be discriminated using these low-level features (see discussion for more details).

Nevertheless, given features (μ_i; *where* $i \in [0, k)$) the models learnt a mapping (Γ) from features to the probability (p) of this data belonging to the different classes (c; *where* $c = \{control, dementia\}$ *or* $c = \{control, PCA, tAD\}$). This is as follows:

$$p\,(c|\mu_0, \ldots, \mu_k) = \Gamma(\mu_0, \ldots, \mu_k)$$

Based on the value of this probability for each class, the most likely class for that data can then be ascertained as the class with the maximum probability.

Evaluation of Models

A Leave-One-Person-Out (also called leave-one-subject-out, LOSO) cross validation was used to evaluate the generalization capabilities of our predictions (see **Figure 2E**). In this method, the model is trained on the data from all but one participant (Cho et al., 2019). Predictions are then made on the data from the remaining participant to gauge how well the model performs on unseen data from a participant on which it has not been trained. As data from each model are not independent from one another, the Cochran's Q test was used to determine the significance of the overall accuracy of each model. This was done using the dichotomous "true" or "false" prediction for each fold. A pairwise *post-hoc* Dunn test with Bonferroni adjustments was used to test for differences between models. All statistical tests were run with a significance level of $\alpha = 0.05$ and were conducted using IBM SPSS V25 (IBM SPSS Statistics, RRID:SCR_019096).

Furthermore, we report accuracy and F1 scores for all models. These are calculated by exhaustively leaving each participant out (as explained above), training the model on the remaining participants and evaluating the model on the participant left out. The accuracy and F1 score were then calculated across all these folds of the data. The accuracy was calculated as the number of correctly classified trials over the total number of trials. F1 scores with respect to each class were calculated as:

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

Hyper-Parameter Optimization

The hyper-parameters for all models were chosen using hyperparameter optimization—a standard method in Machine Learning for systematically choosing the parameters of the model that are not directly learnt. All the models were tuned for this study using a type of hyper-parameter tuning—exhaustive grid search (Buitinck et al., 2013) in which variations of the model are run repeatedly using different values of the hyper-parameters, that have been identified manually. The hyper-parameters chosen for the model for the final analyses were the parameters that

TABLE 2 | Values of the hyper-parameters (for each model) that yielded the highest performance and were used in all analyses.

Model	Parameter name	Binary parameters	Multiclass parameters
Gradient boosting	Number of trees	80	70
	Maximum depth of trees	1	3
	Minimum samples in leaf nodes	2	2
	Learning rate	0.15	0.05
Random forest	Number of trees	120	120
	Maximum depth of trees	None	3
	Minimum samples in leaf nodes	5	2
MLP	Number of units in hidden layer	8	8
		Non-linearity	Logistic/ sigmoid function
	Maximum number of iterations	750	750
	Learning rate	0.0002	0.0002

produced the best performance while conducting the grid search (**Table 2**). This approach was also used for selecting the number of windows to use in constructing the features (see **Figure 3**).

Direction of Traversal and System Analysis

A secondary aim of the study was to identify the components of the system that promote a high classification accuracy. This involved analyzing: the importance of the three sensors, the importance of the different features and the importance of the direction of traversal of the stairs.

For the analysis of the importance of the sensors, the performance of different variants of the models was analyzed. These variants of the models used features from different combinations of the sensors. The importance of the different features was analyzed using the tree-based models (i.e., the Random Forest and Gradient Boosting models), firstly, because they provide methods for determining the importance of features in making a prediction and secondly, due to their high performance. This analysis was done, by calculating the reduction in impurity (or error) that each node (or partition) provides weighted by the probability of reaching that node in the tree and then averaged over all trees to give the final metric of importance. Therefore, importance represents how well the feature portioned the data into the relevant classes weighted by the likelihood of this feature being used in classifying a datapoint. The analysis of traversal direction was done by training the model on all the data, then separating predictions into those made on trials in the upward direction and those made in the downward direction and calculating the accuracy on these subsets separately.

To understand which sensors were most effective a Kruskal-Wallis H-test was conducted and pairwise *post-hoc* Dunn tests with Bonferroni adjustments were used to determine which sensors to use in further analyses. Finally, a Friedman's Two-Way Analysis of Variance was conducted to understand the importance of features and the influence of upwards and downwards traversal.

TABLE 3 | Results from a representative run of the models for detecting the dementia (PCA/tAD).

Model	Accuracy (%)	F1 score (wrt the control class) (%)	F1 score (wrt the dementia PCA/tAD class) (%)
Gradient boosting	86.05	82.78	88.27
Random forest	87.02	83.14	88.38
MLP	86.63	82.71	87.75

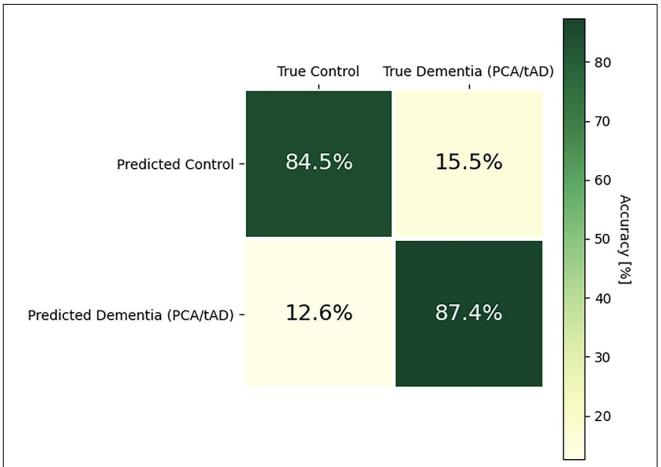


FIGURE 4 | A confusion matrix for the binary Random Forest model.

RESULTS

Prediction Results

This section presents the results achieved in detecting whether participants had dementia as well as the type of dementia.

In the binary models, trained to discriminate people with dementia from controls, the Random Forest Classifier was the most successful at predicting the presence of dementia, which it accurately did in 87.02% of cases (see **Table 3**; **Figure 4** for more details). Furthermore, the F1 score with respect to control class was 83.14 and 88.38% with respect to the dementia class, both of which were higher than the same for any other model. The Cochran's Q test confirmed the differences between the performance of the models, $\chi^2\ (4, N = 516) = 47.56, p < 0.001$.

In the case of the ternary type-based classification (Control vs. tAD vs. PCA), the MLP classifier outperforms all other classifiers and accurately predicts the type of dementia in 68.22% of cases. Furthermore, the F1 score with respect to the control class was 83.72%, 64.8% with respect to the PCA class, and 47.69% with respect to the tAD class. The Cochran's Q test confirmed that there were differences between the performance of the models, $\chi^2\ (4, N = 516) = 47.56, p < 0.001$.

Furthermore, analyzing the confusion matrix of the winning model (the MLP classifier) in the ternary case suggests that the model misclassifies more often between the two types of dementia than with controls (see **Table 4**; **Figure 5**). This could be because people with dementia share some similar

TABLE 4 | Results from a representative run of the models for detecting the type of dementia.

Model	Accuracy	F1 score (wrt the control class)	F1 score (wrt the PCA class)	F1 score (wrt the tAD class)
MLP	68.22%	83.72%	64.8%	47.69%

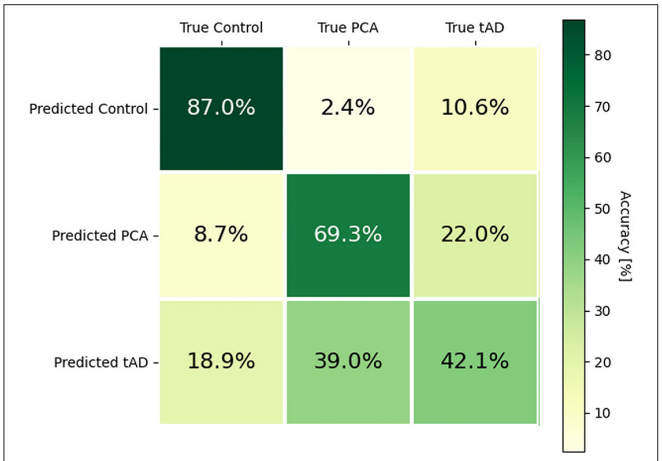


FIGURE 5 | A confusion matrix for the ternary MLP model.

symptoms no matter the type and therefore their gait is much more similar to each other than to that of controls. Moreover, it is more common for the model to confuse participants with tAD with the control group than it is for the model to confuse participants with PCA with the control group. This could be because PCA affects visual processing more than tAD, and therefore the effects of this disease are more prominent in a trial such as this. This trend has also been identified by previous research done in the same program of work at Pedestrian Accessibility Movement Environment Laboratory (PAMELA), which found that participants with early stage PCA performed worse than people with tAD (Yong et al., 2020). Therefore, because the gait of participants with PCA is more easily distinguishable from "normal" gait than the gait of participants with tAD, the model does not confuse PCA with controls as often as it confuses tAD with controls.

In summary, these models could enable an in-the-wild screening tool for dementia, allowing people to conduct an initial screening, with reasonably high accuracy, before potentially receiving a clinical test to verify this. However, further research is required before this is possible, particularly in the case of the type-based classification where accuracy for the two types of dementia is lower than that for controls, suggesting that the current system may be sensitive to dementia, but not its type. See the discussion for more details.

Direction of Traversal and System Analysis

Analysis of Number of Sensors

A Kruskal-Wallis H test showed that there was a statistically significant difference in the importance of the sensors, $\chi^2(6) = 157.13$, $p < 0.001$. Specifically, we tested the performance across model variants that used all different combinations of sensors (left foot; right foot; pelvis; left foot and right foot; left foot and pelvis; right foot and pelvis; left foot, right foot and pelvis). *Post-hoc* analysis showed the best performing combination was found to be the left and right foot sensor features together. These together gave a mean rank of 163.22 and an average accuracy of 85.94%. In contrast the worst performance was given by the pelvis features alone which had a mean rank of 13.00 and an accuracy of 74.45%. The importance of the placement and number of sensors, as given by the resulting accuracy, are given in **Table 5**.

TABLE 5 | Average accuracies of Binary Gradient Boosting Classifiers using different sensors.

Position of sensors used	Accuracy (%)
Left foot	81.99
Right foot	84.78
Pelvis	74.45
Left, right foot	85.94
Left foot, pelvis	81.90
Right foot, pelvis	83.25
Left foot, right foot, pelvis	83.41

The table shows the average accuracies (across 25 samples) of the Binary Gradient Boosting classifier when using the data from different combinations of the sensors to construct the features.

The importance of the feet sensors in predictions could be explained simply because gait, which is heavily based on steps, can be more easily deduced from the movement of the feet, than the pelvis. Therefore, the accuracy of the model that uses a sensor on each foot is significantly higher than the others. Furthermore, it is interesting to note that the model that uses all three sensors yields a significantly lower accuracy than the model that uses only just two sensors—one on each foot. A potential reason for this is that given the data from each foot sensor, the pelvis sensor provides little additional useful information. Therefore, this information does not enhance the performance of the model, but could allow the model to identify trends that exist in the training set (or a subset of it) but do not generalize to other cases, causing the model to overfit to the training data.

The rest of the analyses (presented in this paper) used only the sensors attached to the feet as these produced the best performance. This analysis shows that when the data from sensors is processed independently of each other, sensors attached to participants' feet are more informative for making predictions.

These results of this analysis could not only be interesting to clinicians, and other researchers aiming to build similar systems, but also means that the sensor system can be truly unobtrusive as it does not require a pelvis sensor that can cause discomfort, thereby allowing its use in the wild. See the Discussion for more information about this.

The Importance of Features

Further analysis of the models was conducted to better understand how features from the gyroscope and the

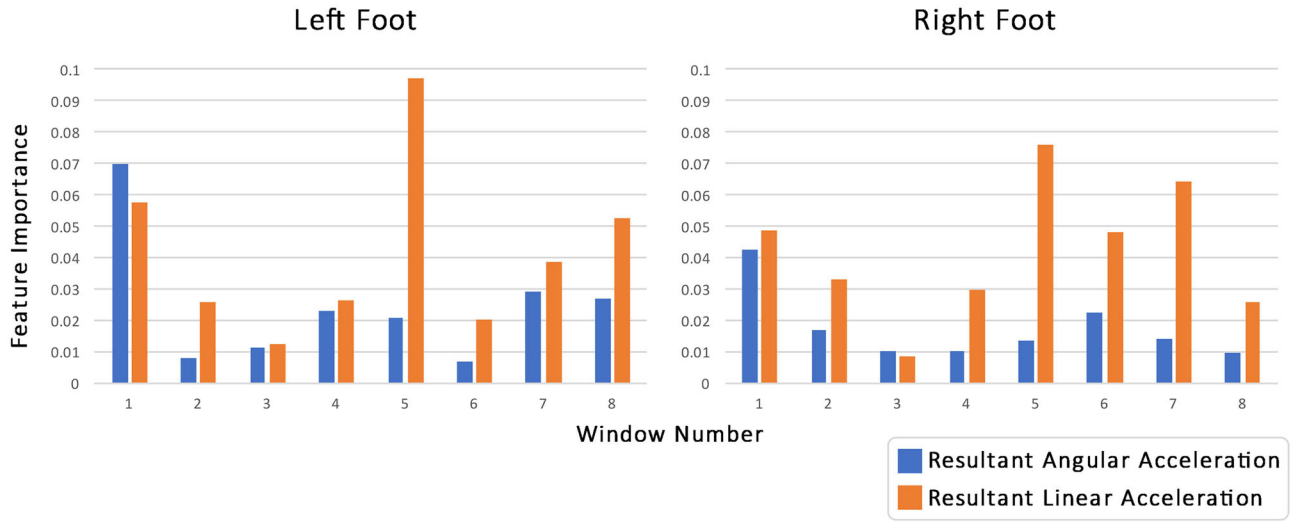


FIGURE 6 | The importance of the features for the Random Forest Classifier when predicting dementia. The features used were the windowed averages (number of windows 8) of linear acceleration (blue bars) and angular acceleration data (orange bars) for both the left (left hand side) and right sensor (right hand side). Feature importance was calculated as the reduction in impurity (or error) that each node (or partition) provides weighted by the probability of reaching that node in the tree and then averaged over all trees.

accelerometer contributed to the overall prediction (**Figure 6**). This was analyzed by looking at the feature importance, using the tree-based models. Feature importance was calculated as the reduction in impurity (or error) that each node (or partition) provides weighted by the probability of reaching that node in the tree and then averaged over all trees. A Kruskal-Wallis H test showed that linear acceleration was statistically more important than angular acceleration $\chi^2(31) = 795.47$, $p < 0.001$. While there is no conclusive explanation for this it is possible that this occurs because acceleration and velocity are directly related. Therefore, acceleration provides the model with useful information about the speed of a participant, the points when the foot is at rest, and how quickly the participant progresses through the trial. These have been identified by previous research (Verghese et al., 2007; Cedervall et al., 2014; Carton et al., 2016; Castrillo et al., 2016; Del Campo et al., 2016; Montero-Odasso, 2016) as factors that help distinguish participants with dementia from those without.

Furthermore, it appears (**Figure 6**; **Table 6**) that if we divide the trial into two halves (windows 1–4 and 4–8, respectively), then the second half appears more important generally for the model. To analyze this further the importance of the linear accelerations and the angular accelerations for the 4 windows in the two halves were summed together for each sensor and each type of acceleration. A second Kruskal-Wallis *H*-test was applied followed by pairwise *post-hoc* Dunn tests with Bonferroni adjustments. Each of the pairwise comparisons was significant. The importance of the linear acceleration in the second half of the trial was found to be significantly greater than that of the first ($p = 0.014$), which in turn was found to be significantly greater than the angular acceleration in the last half ($p < 0.001$). The angular acceleration in the first half was the least important and significantly less than the angular acceleration in the second half ($p = 0.014$).

TABLE 6 | Results of hypothesis testing comparing the linear and angular acceleration in the first (windows 1–4) and second (windows 5–8) halves of the trial.

	First half (%)	Second half (%)	*p*-value
Linear acceleration	24.17	42.25	<0.001
Angular acceleration	19.20	14.37	<0.001

This analysis was conducted on all tree-based models (in both the binary and multi-class settings) which provide easy ways to calculate and analyze the importance of features, as well as being among the best performing models, and the trends identified across all these tree-based models were similar. Therefore, this analysis identified the most informative components of the trial for distinguishing participants with dementia from controls, however, further research is required to provide an explanation for why these trends occur.

The Effect of Traversal Direction

The analysis of the direction of traversal of the stairs that helps distinguish people with dementia from controls is presented in this section. The mean accuracy of the upward or downward directions are given in **Table 7**. This suggested that for people with dementia the binary models were more accurate in the upwards direction as compared to the downwards direction.

To analyze this further, the same analysis was conducted in the multiclass setting with accuracies split according to the class. The results of this analysis are summarized in **Table 8**.

TABLE 7 | Results of hypothesis testing comparing the prediction accuracies attained in the upward and downward directions.

Model	Upward accuracy (%)	Downward accuracy (%)
Random forest	86.77	85.31
Gradient boosting	86.97	86.08
MLP	89.29	82.79

TABLE 8 | The average accuracies (across 25 samples) of the better performing models for predicting dementia phenotype.

Model	Upwards accuracy			Downwards accuracy		
	Control (%)	PCA (%)	tAD (%)	Control (%)	PCA (%)	tAD (%)
Random forest	80.12	56.43	51.1	79.46	63.11	41.16
Gradient boosting	79.03	61.95	45.90	89.69	71.37	30.03
MLP	79.42	74.32	49.5	92.19	71.9	34.89

A Friedman's Two-Way Analysis of Variance was conducted which proved there was a significant difference between the models and between up and down conditions $\chi^2(17) = 415.41$, $p < 0.001$. Pairwise analysis across two independent variables (models and up/down) was not conducted as it was thought to be over analysis of the data. However, from **Table 8** it can be seen that in the multiclass tree-based models the percentage of the trials that were correctly classified as PCA is generally higher in the downward direction, which is in contrast to the results found for classifying dementia with binary models. This could be attributed to the fact that on the way down, the stairs are not directly in participants' line of sight when looking forward and, therefore, it is harder for them to process this information. Alternatively, it could be that descending stairs is less physically demanding, but the consequence of falling is greater when descending, causing anxiety in the participants.

While this analysis provides interesting insights into which direction of traversal is more informative for predicting dementia, the varied results across different models led to this analysis being inconclusive. Moreover, further research is required to provide an explanation for these differences.

The analysis of the importance of features and the direction of traversal provides some initial insights into how the gait of people with dementia (both PCA and tAD) could differ from that of controls, which may be informative to healthcare workers and patients. However, further analysis is required into the varied results and generalizability of these findings to other environments. See the Discussion for more details.

DISCUSSION

This section discusses the contributions made, current limitations and future possible use cases of the STEP-UP system.

Detection and Discrimination of Dementia

While previous research has helped to identify the changes in gait that occur during the decline into dementia, the research has ignored two important factors that would allow such analyses to be used in the real world. Firstly, previous research relies heavily on experiments conducted in laboratory settings, using technologies such as optical systems that cannot be used in the real-world (Verghese et al., 2007; Wittwer et al., 2013; Callisaya et al., 2017) and treadmills which constrain the way of walking to a straight line. This limits the applications of this research as people hoping to use this method to screen for early cues of dementia would need to be subjected to these laboratory tests. Secondly, previous research often ignores different types of Alzheimer's focusing instead on tAD. The use of low-cost wearable technology offers the opportunity to gather data about people's ability to conduct everyday tasks, including climbing or descending stairs as they go about their life. Previous research (Plant and Barton, 2020) suggests that data from everyday life are more informative about a person's disease than data in clinical assessment laboratory where people may attempt to over control their behavior. In addition, as such sensors get integrated into people's clothes and accessories, early detection of possible problems (especially rarer types of dementia like PCA) could be detected before people purposely look for a dementia assessment.

Our study has demonstrated the feasibility of deploying low-cost sensors to measure gait patterns for predicting dementia (both tAD and a rarer type of dementia: PCA) in everyday tasks of climbing and descending stairs. We have achieved this by focusing on low-level input features and investigating their non-linear mapping onto types of dementia and controlled groups with supervised classifiers. This is of critical importance when it comes to low-cost systems being used in the real world as calculating hand-engineered high-level gait features (e.g., Verghese et al., 2007) is often infeasible and requires high level controls. Also, low-level features used with artificial neural networks have been shown repeatedly to have higher robustness for other sensing modalities (Kostek et al., 2004; Cho et al., 2019).

In this research we analyzed the detection of dementia as compared to healthy participants, however, real-world deployment could enable larger datasets. This could further lead to an improvement in the performance not only on the detection of dementia cues but also on discriminating between different types of dementia. Moreover, the inclusion of more varied data such as that of participants with Mild Cognitive Impairement or early stages of dementia could enable this system to be used by these populations, allowing for early-stage detection. While we did not look at these populations, previous research analyzing gait using similar methods and measures has found that gait is sensitive to early signs of dementia and can predict cognitive decline (Marquis et al., 2002; Waite et al., 2005; Wang et al., 2006; Verghese et al., 2007; Cedervall et al., 2014; Gwak et al., 2018; Halloway et al., 2019; Schaat et al., 2020). Therefore, deployment of this system in real-world settings could enable dementia detection in everyday settings which could bring several use cases and potential benefits. While in-depth analysis of this is

left to future research, some of the potential future examples are discussed below:

Screening Tool for Healthcare Workers and Practitioners

A screening tool which could be deployed in clinical settings or as an at-home test can be developed. The clinical tool could be used by community healthcare workers as well as general practitioners to enable easy detection of typical and atypical presentations of Alzheimer's disease. Carers' wellbeing can often be neglected, however they are often under considerable stress (Gilhooly et al., 2016). The amount of stress carers experience decreases with acceptance of the diagnosis and social support networks, and is increased with wishful thinking, denial and avoidance strategies (Gilhooly et al., 2016). An early diagnosis gives more time for acceptance and support networks to be established. These benefit the person diagnosed, their families and carers. It could be that beyond the benefits of simple screening we could also investigate ways of developing support tools for the carers, which could be linked to the stage of dementia of the person for whom they are caring.

General Self-Screening

As sensors are increasingly integrated into our daily activities (e.g., sensor in shoes for running, imaging for fitness tracking) and used to quantify our wellbeing (Cho et al., 2017; Cho, 2021), such sensors could be used together to detect and identify cues of decline and dementia. Our results provide some insights on how the sensors could be used in the wild. Firstly, our research found that the presence of dementia is more easily detected during upwards stair climbing, suggesting that the gait of people with dementia is more abnormal during upwards stair climbing. The same sensors placed on the shoes could first detect upward stair climbing (Formento et al., 2014) and data from this activity can be prioritized for more accurate predictions. Similarly, the sensors could also detect long periods of activity and even fatigue or pain (Wang et al., 2019) and consider such variables when evaluating the assessment tool outcome. Finally, as any motor activity modeling suffers from people's idiosyncrasy, such models could take advantage of the long history of sensor data gathered from the person to build personal models of what is a normal pattern (given the physical ability including vision of the person) and hence detect possible sudden declines that may indicate such underlying causes of dementia and even atypical causes.

Support Tool for Patients

It would seem feasible to also develop the ability to classify deteriorations in a person's condition following diagnosis. This would need a larger data set collected in the wild. Once developed decline in gait such as those detected by lab-based studies (e.g., Verghese et al., 2007; Callisaya et al., 2017) could be detected as people conduct their daily activities and be directly linked to clinical care pathways. This would enable person-centered care to be established, rather than simply asking people to return for appointments based on standard time predictions of decline.

An important perspective is on the effect of different combinations of sensors on the detection performance. Our research found that of all combinations of the sensors, models using only the sensors attached to the feet performed best. This led to us dropping the pelvis sensor from further analyses. Additionally, a sensor constantly attached to a person's pelvis may cause discomfort. Therefore, our research suggests that a truly unobtrusive system could be built simply with sensors attached to people's shoes. Furthermore, the support tool could be further developed to be predictive of decline, providing further support to people with dementia and their care givers.

Limitations

Despite promising results, there is room for improvement. We discuss points to help the deployment of such a system.

Discriminating Type

While the model has shown a good performance (from LOSO cross-validation) in the multi-class classification (Control vs. tAD vs. PCA), we have found lower performance in discriminating the two types of dementia when samples from the controlled group are not considered in the classification task. This can provide insights. First, this could be related to the fact that the gait of the two subtypes of dementia was very similar to each other, suggesting that gait is sensitive to dementia as a whole, but less sensitive to the type of dementia. This could suggest that different measures may be required to provide a more comprehensive diagnosis. For example, in PCA vision is predominantly affected with memory often being (initially) unaffected. Second, the data from healthy participants could play an essential role in discriminating patterns associated with each dementia type. Third, when it comes to the dementia detection task (dementia vs. control), the proposed system results in a very high accuracy of 87.02%.

Generalization Issues and Dataset

Another potential limitation in this study is that models might be overfit to the data, reducing its ability to generalize to unseen data. While we prevented this as much as possible by using LOSO validation, ensuring the model was not only tested on unseen data but on data from an unseen participant. However, all the data from all participants was collected on the same staircase using the same system setup to collect the data. Therefore, these models may not generalize to other environments, other staircases or other IMU systems. This may limit the direct application of this system to the real-world diagnosis of dementia. Therefore, further research is required to prove the generalizability of this research to other environments and system implementations.

Another related issue was that it was more difficult to achieve a high degree of control in the task especially in people with dementia. This may have resulted in patients taking breaks in the middle of the task, not initially standing in the correct start position, etc. Therefore, the model might use these artifacts to discriminate patients from controls rather than their gait. Nevertheless, these behaviors are symptoms of dementia that should generalize across patients.

Furthermore, in this study we only compared the gait of participants with dementia to healthy age-matched controls. Therefore, this model may be overfit to distinguishing healthy and unhealthy participants and may not be able to distinguish dementia from other diseases with similar presentations or

people with a bad physical condition. Therefore, this requires further research and fine-tuning of this issue. We believe that the deployment of this system in the real-world would enable overcoming these overfitting issues by allowing more varied data to be tested.

CONCLUSION

This research demonstrates the feasibility of automatically detecting both the more typical Alzheimer's Disease (tAD) as well as a rarer and distinct form of dementia—Posterior Cortical Atrophy (PCA)—based on gait in a real world-environment. To this end, we propose the use of low-level features based on windowed averaging of data from a low-cost, unobtrusive IMU system. These features are easy to calculate from a small number of IMU sensors, enabling their use in a real-world system. We also demonstrate that these features can be used with Machine Learning models to predict dementia with 87.02% accuracy. Furthermore, we demonstrate that a sensor placed on each foot is sufficient for this analysis. Lastly, we demonstrate the models are better able to discriminate people with dementia from healthy controls when they are climbing up stairs, suggesting that people with dementia find it harder to climb up stairs.

Therefore, this research concludes that machine learning analysis of IMU data, gathered from a person's gait in a real-world environment, could unobtrusively be used to assess the risk of having dementia. Once further researched, a system such as this could provide an initial assessment of the risk of having a certain type of dementia before conducting any clinical tests, thereby streamlining and enhancing the diagnostic process. Therefore, not only are these results interesting from a research perspective, but also have potential real-world applications.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by National Research Ethics Service Committee London Queen Square. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CH: conception of analysis, acquisition of data, drafting manuscript, analysis of data, and built use cases. WB: conception of analysis, analysis of data, drafting manuscript, and built use cases. KY: conception and design of experimental protocol, acquisition of data, and assisted drafting manuscript. IM and TS: acquisition of data and advised on data analysis. AC and BY: acquisition of data. RS, DB, and NT: conception and design of experimental protocol. SC: conception and design of experimental protocol and assisted drafting manuscript. NB: drafting manuscript, advised on analysis of data, and built use cases. YC: overall technical supervision, drafting manuscript, advised on analysis of data, and built use cases. All authors contributed to the article and approved the submitted version.

FUNDING

The Dementia Research Center is an Alzheimer's Research UK Co-ordinating Center and was supported by Alzheimer's Research UK, Brain Research Trust, and the Wolfson Foundation. This work was also supported by the NIHR Queen Square Dementia Biomedical Research Unit and by an Alzheimer's Research UK Senior Research Fellowship (ART-SRF2010-3) and ESRC/NIHR (ES/L001810/1) and EPSRC (EP/M006093/1) grants to SC. KY is funded by the Alzheimer's Society, grant number 453 (AS-JF-18-003).

ACKNOWLEDGMENTS

We would like to thank participants for their patience and goodwill in taking part.

REFERENCES

Alavi, H. S., Denis, L., and Yvonne, R. (2020). The five strands of living lab: a literature study of the evolution of living lab concepts in HCI. *ACM Trans. Comput. Hum. Interac.* 27, 26. doi: 10.1145/3380958

Allali, G., Annweiler, C., Blumen, H. M., Callisaya, M. L., De Cock, A.-M., et al. (2016). Gait phenotype from mild cognitive impairment to moderate dementia: results from the GOOD initiative. *Eur. J. Neurol.* 23, 527–541. doi: 10.1111/ene.12882

Buitinck, L., Gilles, L., Mathieu, B., Fabian, P., Andreas, M., Olivier, G., et al. (2013). API design for machine learning software: experiences from the scikit-learn project. *arXiv[Preprint].arXiv:1309.0238.*

Callisaya, M. L., Launay, C. P., Srikanth, V. K., Verghese, J., Allali, G., and Beauchet, O. (2017). Cognitive status, fast walking speed and walking speed reserve—the gait and Alzheimer interactions tracking (GAIT) study. *GeroScience* 39, 231–239. doi: 10.1007/s11357-017-9973-y

Cano, L. A. M., Beltrán, J., Navarro, R., García-Vázquez, M. S., and Castro, L. A. (2017). Towards early dementia detection by oculomotor performance analysis on leisure web content," in *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. UbiComp 17. Maui, Hawaii: Association for Computing Machinery, 800–804.

Carton, A. M., Yong, K. X., Peters, A., Ocal, D., Kaski, D., Gonzalez, A. S., et al. (2016). Effects of dementia-related visual impairement on route following in posterior cortical atrophy and typical Alzheimer's disease. *Alzheimer Dement.* 12(Suppl. 7), P257–58. doi: 10.1016/j.jalz.2016.06.461

Castrillo, A., Olmos, L. G., Rodríguez, F., and Duarte, J. (2016). Gait disorder in a cohort of patients with mild and moderate Alzheimer's disease. *Am. J. Alzheimer's Dis. Other Dement.* 31, 257–262. doi: 10.1177/1533317515603113

Cedervall, Y., Halvorsen, K., and Åberg, A. C. (2014). A longitudinal study of gait function in people with Alzheimer disease. *Gait Post.* 39, S138. doi: 10.1016/j.gaitpost.2014.04.198

Cho, Y. (2021). "Rethinking eye-blink: assessing task difficulty through physiological representation of spontaneous blinking," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery), 1–12.

Cho, Y., Julier, S. J., and Bianchi-Berthouze, N. (2019). Instant stress: detection of perceived mental stress through smartphone photoplethysmography and thermal imaging. *JMIR Mental Health.* 6, 140. doi: 10.2196/10140

Cho, Y., Julier, S. J., Marquardt, N., and Bianchi-Berthouze, N. (2017). Robust tracking of respiratory rate in high-dynamic range scenes using mobile thermal imaging. *Biomed. Optics Express.* 8, 4480–4503. doi: 10.1364/BOE.8.004480

Crutch, S. J., Lehmann, M., Schott, J. M., Rabinovici, G. D., Rossor, M. N., and Fox, N. C. (2012). Posterior cortical atrophy. *Lancet Neurol.* 11, 170–178. doi: 10.1016/S1474-4422(11)70289-7

Crutch, S. J., Yong, K. X., and Shakespeare, T. J. (2016). Looking but not seeing: recent perspectives on posterior cortical atrophy. *Curr. Direct. Psychol. Sci.* 25, 251–260. doi: 10.1177/0963721416655999

Del Campo, N., Payoux, P., Djilali, A., Delrieu, J., Hoogendijk, E. O., Rolland, Y., et al. (2016). Relationship of regional brain β-amyloid to gait speed. *Neurology* 86, 36–43. doi: 10.1212/WNL.0000000000002235

Duff, K. (2020). Cognitive composites in AD trials? Drinking the kool-aid and paying the price? *Alzheimer Dement. Diagn. Assess. Dis. Monit.* 12, e12011. doi: 10.1002/dad2.12011

Firth, N. C., Primativo, S., Marinescu, R. V., Shakespeare, T. J., Suarez-Gonzalez, A., Lehmann, M., et al. (2019). Longitudinal neuroanatomical and cognitive progression of posterior cortical atrophy. *Brain* 142, 2082–2095. doi: 10.1093/brain/awz136

Formento, P. C., Acevedo, R., Ghoussayni, S., and Ewins, D. (2014). Gait event detection during stair walking using a rate gyroscope. *Sensors (Basel, Switzerland)* 14, 5470–5485. doi: 10.3390/s140305470

Gale, S. A., Acar, D., and Daffner, K. R. (2018). Dementia. *Am. J. Med.* 131, 1161–1169. doi: 10.1016/j.amjmed.2018.01.022

Gilhooly, K. J., Gilhooly, M. L., Sullivan, M. P., McIntyre, A., Wilson, L., Harding, E., et al. (2016). A meta-review of stress, coping and interventions in dementia and dementia caregiving. *BMC Geriat.* 16, 106. doi: 10.1186/s12877-016-0280-8

Graff-Radford, J., Yong, K. X., Apostolova, L. G., Bouwman, F. H., Carrillo, M., Dickerson, B. C., et al. (2021). New insights into atypical Alzheimer's disease in the era of biomarkers. *Lancet Neurol.* 20, 222–234. doi: 10.1016/S1474-4422(20)30440-3

Gwak, M., Woo, E., and Sarrafzadeh, M. (2018). "The role of accelerometer and gyroscope sensors in identification of mild cognitive impairment," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)* (Anaheim, CA), 434–38.

Hall, C. B., Lipton, R. B., Sliwinski, M., and Stewart, W. F. (2000). A change point model for estimating the onset of cognitive decline in preclinical Alzheimer's disease. *Stat. Med.* 19, 1555–66. doi: 10.1002/(SICI)1097-0258(20000615/30)19:11/12<1555::AIDSIM445>3.0.CO;2-3

Halloway, S., Arfanakis, K., Wilbur, J., Schoeny, M. E., and Pressler, S. J. (2019). Accelerometer physical activity is associated with greater gray matter volumes in older adults without dementia or mild cognitive impairment. *J. Gerontol. B.* 74, 1142–1151. doi: 10.1093/geronb/gby010

Holloway, C. (2019). Disability interaction (DIX): a manifesto. *Interactions* 26, 44–49. doi: 10.1145/3310322

Holloway, C., and Dawes, H. (2016). Disrupting the world of disability: the next generation of assistive technologies and rehabilitation practices. *Healthcare Technol. Lett.* 3, 254–256. doi: 10.1049/htl.2016.0087

International Alzheimer's Disease (2019). *World Alzheimer Report 2019: Attitudes to Dementia | Alzheimer's Disease International.* Available online at: https://www.alz.co.uk/research/world-report-2019 (accessed September 20, 2019).

Kostek, B., Szczuko, P., and Zwan, P. (2004). "Processing of musical data employing rough sets and artificial neural networks," in *Rough Sets and Current Trends in Computing*, eds S. Tsumoto, R. Słowiński, J. Komorowski, and Jerzy W. Grzymała-Busse. Lecture Notes in Computer Science. (Berlin, Heidelberg: Springer), 539–48.

Marquis, S., Moore, M. M., Howieson, D. B., Sexton, G., Payami, H., Kaye, J. A., et al. (2002). Independent predictors of cognitive decline in healthy elderly persons. *Arch. Neurol.* 59, 601–606. doi: 10.1001/archneur.59.4.601

Mc Ardle, R., Del Din, S., Galna, B., Thomas, A., and Rochester, L. (2020). Differentiating dementia disease subtypes with gait analysis: feasibility of wearable sensors? *Gait Posture* 76, 372–76. doi: 10.1016/j.gaitpost.2019.12.028

McCarthy, I., Suzuki, T., Holloway, C., Poole, T., Frost, C., Carton, A., et al. (2019). Detection and localisation of hesitant steps in people with Alzheimer's disease navigating routes of varying complexity. *Healthcare Technol. Lett.* 6, 42–47. doi: 10.1049/htl.2018.5034

McIntyre, A., Harding, E., Yong, K. X., Sullivan, M. P., Gilhooly, M., Gilhooly, K., et al. (2019). Health and social care practitioners' understanding of the problems of people with dementia-related visual processing impairment. *Health Soc. Care Commun.* 27, 982–990. doi: 10.1111/hsc.12715

Mendez, M. F., Ghajarania, M., and Perryman, K. M. (2002). Posterior cortical atrophy: clinical characteristics and differences compared to Alzheimer's disease. *Dement. Geriatr. Cogn. Disord.* 14, 33–40. doi: 10.1159/000058331

Montero-Odasso, M. (2016). Gait as a biomarker of cognitive impairment and dementia syndromes. Quo Vadis? *Eur. J. Neurol.* 23, 437–438. doi: 10.1111/ene.12908

Ocal, D., Yong, K., McCarthy, I., Suzuki, T., Suzuki, A., Boampong, D., et al. (2017). Effects of ground lighting uniformity and clutter on navigational ability in posterior cortical atrophy and typical Alzheimer's disease. *Alzheimers Dement.* 13(Suppl. 7): P534–35. doi: 10.1016/j.jalz.2017.06.635

Ojeda, L., and Borenstein, J. (2007). "Non-GPS navigation with the personal dead-reckoning system - art. No. 65610C," in *Proceedings of SPIE - The International Society for Optical Engineering* (Orlanda, FL).

Park, S. K., and Suh, Y. S. (2010). A zero velocity detection algorithm using inertial sensors for pedestrian navigation systems. *Sensors (Basel, Switzerland)* 10, 9163–9178. doi: 10.3390/s101009163

Plant, D., and Barton, A. (2020). Adding value to real-world data: the role of biomarkers. *Rheumatology* 59, 31–38. doi: 10.1093/rheumatology/kez113

Rosso, A. L., Verghese, J., Metti, A. L., Boudreau, R. M., Aizenstein, H. J., Kritchevsky, S., et al. (2017). Slowing gait and risk for cognitive impairment: the hippocampus as a shared neural substrate. *Neurology* 89, 336–342. doi: 10.1212/WNL.0000000000004153

Schaat, S., Koldrack, P., Yordanova, K., Kirste, T., and Teipel, S. (2020). Real-time detection of spatial disorientation in persons with mild cognitive impairment and dementia. *Gerontology* 66, 85–94. doi: 10.1159/000500971

Schneider, L. S., and Goldberg, T. E. (2020). Response to peer commentaries: composite cognitive and functional measures for early stage Alzheimer's disease trials. *Alzheimers Dement. Diagn. Assess. Dis. Monit.* 12, e12024. doi: 10.1002/dad2.12024

Shibata, D., Wakamiya, S., Ito, K., Miyabe, M., Kinoshita, A., and Aramaki, E. (2018). "VocabChecker: measuring language abilities for detecting early stage dementia," in *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, Companion. Tokyo, Japan: Association for Computing Machinery.

Tröger, J., Linz, N., König, A., Robert, P., and Alexandersson, J. (2018). "Telephone-based dementia screening I: automated semantic verbal fluency assessment," in *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*. PervasiveHealth18, New York, NY, USA: Association for Computing Machinery.

Valkanova, V., and Ebmeier, K. P. (2017). What can gait tell us about dementia? Review of epidemiological and neuropsychological evidence. *Gait Posture* 53, 215–223. doi: 10.1016/j.gaitpost.2017.01.024

Verghese, J., Wang, C., Lipton, R. B., Holtzer, R., and Xue, X. (2007). Quantitative gait dysfunction and risk of cognitive decline and dementia. *J. Neurol. Neurosurg. Psychiatry* 78, 929–935. doi: 10.1136/jnnp.2006.106914

Waite, L. M., Grayson, D. A., Piguet, O., Creasey, H., Bennett, H. P., and Broe, G. A. (2005). Gait slowing as a predictor of incident dementia: 6-year longitudinal data from the Sydney older persons study. *J. Neurol. Sci. Vasc. Dement.* 229–230, 89–93. doi: 10.1016/j.jns.2004.11.009

Wang, C., Olugbade, T. A., Mathur, A., De, C., Williams, A. C., Lane, N. D., et al. (2019). "Recurrent network based automatic detection of chronic pain protective behavior using MoCap and SEMG Data," in *Proceedings of the 23rd International Symposium on Wearable Computers*, 225–30. ISWC 19. London, United Kingdom: Association for Computing Machinery.

Wang, L., Larson, E. B., Bowen, J. D., and van Belle, G. (2006). Performance-based physical function and future dementia in older people. *Arch. Intern. Med.* 166, 1115–1120. doi: 10.1001/archinte.166.10.1115

Wittwer, J. E., Webster, K. E., and Hill, K. (2013). Effect of rhythmic auditory cueing on gait in people with Alzheimer disease. *Arch. Phys. Med. Rehabil.* 94, 718–724. doi: 10.1016/j.apmr.2012.11.009

Yong, K. X., McCarthy, I. D., Poole, T., Ocal, D., Suzuki, A., Suzuki, T., et al. (2020). Effects of lighting variability on locomotion in posterior cortical atrophy. *Alzheimers Dement. Transl. Res. Clin. Interven.* 6, e12077. doi: 10.1002/trc2.12077

Yong, K. X., McCarthy, I. D., Poole, T., Suzuki, T., Yang, B., Carton, A. M., et al. (2018). Navigational cue effects in Alzheimer's disease and posterior cortical atrophy. *Ann. Clin. Transl. Neurol.* 5, 697–709. doi: 10.1002/acn3.566

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

BRIEF RESEARCH REPORT
published: 02 February 2022
doi: 10.3389/fpsyg.2021.781487

How the Brunswikian Lens Model Illustrates the Relationship Between Physiological and Behavioral Signals and Psychological Emotional and Cognitive States

Judee K. Burgoon[1], Rebecca Xinran Wang[2], Xunyu Chen[2], Tina Saiying Ge[2] and Bradley Dorn[2]*

[1] Center for the Management of Information, University of Arizona, Tucson, AZ, United States, [2] Management Information Systems, University of Arizona, Tucson, AZ, United States

OPEN ACCESS

Edited by:
Francois Bremond, Institut National de Recherche en Informatique et en Automatique (INRIA), France

Reviewed by:
Colin Hesse, Oregon State University, United States
Michal Balazia, Research Centre Inria Sophia Antipolis Méditerranée, France

***Correspondence:**
Judee K. Burgoon
judee@email.arizona.edu

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Psychology

Received: *22 September 2021*
Accepted: *10 December 2021*
Published: *02 February 2022*

Citation:
Burgoon JK, Wang RX, Chen X, Ge TS and Dorn B (2022) How the Brunswikian Lens Model Illustrates the Relationship Between Physiological and Behavioral Signals and Psychological Emotional and Cognitive States. Front. Psychol. 12:781487. doi: 10.3389/fpsyg.2021.781487

Social relationships are constructed by and through the relational communication that people exchange. Relational messages are implicit nonverbal and verbal messages that signal how people regard one another and define their interpersonal relationships—equal or unequal, affectionate or hostile, inclusive or exclusive, similar or dissimilar, and so forth. Such signals can be measured automatically by the latest machine learning software tools and combined into meaningful factors that represent the socioemotional expressions that constitute relational messages between people. Relational messages operate continuously on a parallel track with verbal communication, implicitly telling interactants the current state of their relationship and how to interpret the verbal messages being exchanged. We report an investigation that explored how group members signal these implicit messages through multimodal behaviors measured by sensor data and linked to the socioemotional cognitions interpreted as relational messages. By use of a modified Brunswikian lens model, we predicted perceived relational messages of dominance, affection, involvement, composure, similarity and trust from automatically measured kinesic, vocalic and linguistic indicators. The relational messages in turn predicted the veracity of group members. The Brunswikian Lens Model offers a way to connect objective behaviors exhibited by social actors to the emotions and cognitions being perceived by other interactants and linking those perceptions to social outcomes. This method can be used to ascertain what behaviors and/or perceptions are associated with judgments of an actor's veracity. Computerized measurements of behaviors and perceptions can replace manual measurements, significantly expediting analysis and drilling down to micro-level measurement in a previously unavailable manner.

Keywords: nonverbal communication, relational communication, dominance, affection, involvement, trust, similarity, nervousness

INTRODUCTION: RELATIONAL COMMUNICATION AND THE BRUNSWIKIAN LENS MODEL

Relational communication forms the architecture through which social relationships are constructed. As expressed by Hawes (1973), "communication functions not only to transmit information but to bind symbol users (p. 15)." Through ubiquitous verbal and nonverbal relational messages, people reciprocally signal the nature of their interpersonal relationships. Implicit signals express how people regard one another and how they gauge the ongoing status of their interpersonal relationships (Guerrero et al., 2017). The signals form non-orthogonal, generic message themes known as *topoi* (Burgoon and Hale, 1984). Drawn from a synthesis of literature and theorizing from multiple social science disciplines, these *topoi* are universal forms of expressions between humans. They represent the fundamental meanings that define how people relate to one another along such dimensions as dominance, affection, involvement, composure, similarity, and trust.

One way to understand the cognitive and emotional components of relational communication is through the application of a Brunswikian lens model (e.g., Bernieri et al., 1996; Scherer, 2003; Hartwig and Bond, 2011) in which objective *distal indicators* contribute to psychological judgments, also called *proximal percepts,* which are imbued with cognitive or emotional overtones that hold a predictive relationship with outcomes such as deception or credibility. The Brunswikian lens model (**Figure 1**) brings insight into how relational communication can be expressed either through psychological perceptions or through the kinesic, vocalic and linguistic signals that create those meanings. Some people relate to one another according to the concrete, objective signals, such as "my partner stood seven feet away from me and did not touch me." Others relate to one another according to the meanings such signals express, such as, "my partner was detached and cold." These alternative layers of expression can be combined to convey the cognitive and emotional meanings being encoded (expressed) and decoded (deciphered and interpreted). The Brunswikian lens model shows how the different aspects of the signaling process can be combined. The distal, objective signals that can be measured and factored with automated computer tools can be linked to the psychological perceptual judgments that represent relational message themes. These subjective percepts in turn predict communicative outcomes such as successful identification of another's deception or credibility.

Our demonstration of the lens model comes from a deception project conducted in eight different locations (three in the United States and five in diverse international locations). Groups of 5–8 participants played a game called Resistance, during which they carried out a series of decisions to win (or lose) missions and thus to win (or lose) the game. Those who intended to sabotage the missions employed deception and misdirection, which enabled them to win the game. The interest here is in the automatically measured, objective signals emitted by participants. These formed meaningful clusters that were "read" and responded to as relational messages. We illustrate how a modified Brunswikian lens model combines collections of concrete, objective behaviors to form subjective cognitive and emotional states that represent relational communication. Various relational communication themes in turn predict various social outcomes. Put differently, multimodal distal signals link to proximal percepts of relational messages that, in turn, predict outcomes such as the accurate identification of veracity.

METHODS

Sample

College-age participants (N = 695; mean age = 22 years) from universities in 3 United States states (Arizona, California, and Maryland), and 5 international ones (Israel, Zambia, Fiji, Singapore and Hong Kong) were recruited to participate in an interactive social game called Resistance in exchange for payment for their time and possible bonuses. Universities were ones where local and national IRBs approved participation. The Human Research Protection Office of the United States Army Research Laboratory served as the IRB for the United States institutions and approved the project. The diverse international sample was intended to test the generalizability and universality of findings (see Ting-Toomey et al., 2000, regarding various cultural styles). However, comparisons among the eight locations failed to show significant differences, apart from Fijians expressing more dominance, and sample sizes within United States locations were too small to compare cultural differences, so we have omitted cultural comparisons (see Dunbar et al., 2021; Giles et al., 2021 for the cultural comparisons).

Procedures

A detailed description of the game is found in Dorn et al. (2021). An ice-breaker activity introducing one another established a baseline for players' behaviors and perceptions of one another. The games consisted of participants conducting a series of make-believe missions. Teams of up to eight players selected a leader, approved the composition of the teams, then voted for the missions to succeed or fail. Players had been randomly assigned the role of Villager or Spy. Villagers were expected to vote for missions to succeed. Those designated as Spies were expected to engage in occasional deception to cause missions to fail. Spies knew one another's identity; Villagers did not.

After every other round, players rated other team members on 7-point Likert (1932) format scales measuring each other's relational communication (see below). The ultimate winners of the game (Spies or Villagers) were determined by which team won the most rounds (see Dorn et al., 2021, for more details). Players also received bonuses if chosen as the leader or a team member.

Nonverbal audiovisual signals (described below) were captured by tablet computers in front of each player, a 360-degree overhead camera and a webcam on the side that recorded the group as a whole. The audiovisual recordings became the basis for kinesic (body language) and vocalic analysis. The audio signals were translated into text for linguistic analysis.

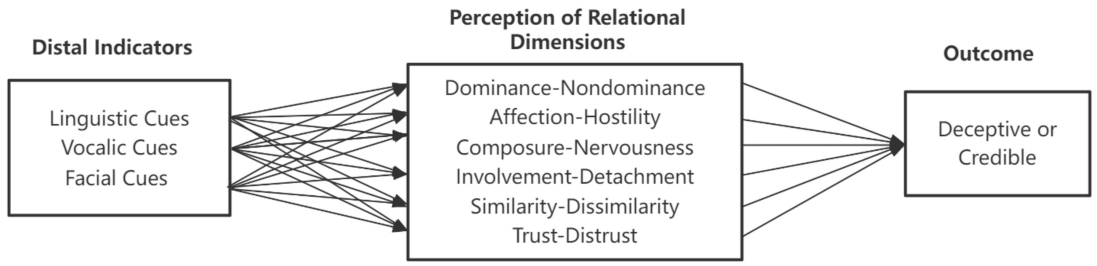


FIGURE 1 | Brunswikian lens model of relational communication.

Affective and Cognitive Measures

The measures that gauged players' emotional and cognitive states were self-report items from the Relational Communication Scale (RCS; Burgoon and Hale, 1987). These generic themes are context-independent. They represent fundamental dimensions along which people identify how they relate to one another and regard themselves in the context of their interpersonal relationships, without regard to the actual verbal content being expressed. The RCS includes 12 non-orthogonal dimensions, 6 of which were measured here: dominance-nondominance, liking-dislike, involvement-detachment, similarity-dissimilarity, composure-nervousness, and trust-distrust. Coefficient alpha reliabilities were 0.91, 0.89, 0.84, 0.78, 0.84, and 0.91, respectively. Some dimensions that were expected to vary across the time course of the game were measured periodically; others that were expected to be more stable were measured at its conclusion.

Outcomes/Attributions

Attributions were based on theories of how people relate to one another and use linguistic, kinesic, and vocalic features to express those relationships. Some features appear in multiple relational messages because relational messages are comprised of constellations of nonverbal and verbal signals. For example, lip corner puller that forms smiles appear in liking, composure, involvement, and trust. The typical compositions of these relational message *topoi* can be found in Burgoon et al. (2022).

Table 1 lists the message themes investigated here and the significant linguistic, vocalic and facial features that emerged for each relational dimension. The linguistic features are a small subset of lexical and syntactic features chosen to illustrate their role in conveying relational message themes measured by SPLICE software (Moffitt et al., 2012). The acoustic features are ones that are measured by OpenSmile (Eyben et al., 2010), an open-source software. The facial features are Action Units and combinations measured by the OpenFace software (Baltrušaitis et al., 2015, 2018), also an open-source software program.

RESULTS

Significant indicators are listed in **Table 1**. Complete statistical results are reported in the **Supplementary Material**. Here we summarize main findings.

Dominance-Nondominance

A central theme defining interpersonal relationships is dominance: who is more powerful, who is more subservient, and whether relationships are more egalitarian. In Burgoon and Dunbar (2006), a number of macro-level strategies are outlined for exhibiting power, dominance, and status or their bipolar opposites. In the current analysis we are more concerned with micro-level nonverbal and verbal behaviors through which those strategies are enacted.

As with previous studies (Zhou et al., 2004; Pentland et al., 2021), dominant players talked more often, for a longer duration, and were more likely to contribute to the conversation. Unexpectedly, mean pitch did not correlate with perceptions of dominance. Rather, the standard deviation of pitch had a significant effect on the player's perceived dominance, indicating dominant individuals talk with more variability in pitch. Further, HNR, which is the proportion of harmonic sound to noise in the voice in decibels (Pentland et al., 2021), was also significant. Higher mean level and lower variability of HNR correlated with a higher perceived dominance. The face was a very active site for signaling dominance or non-dominance. The eye and mouth region were the most involved as dominance signals; language choice played a lesser role.

Affection-Hostility

Whereas dominance represents the vertical aspect of human relations, affection represents the horizontal dimension. Whether called affiliation, liking, positivity, or valence, this dimension is meant to capture the positive to negative sentiment individuals express toward one another. Many of the behaviors associated with expressions of liking are part of other expressions as well, including expressions of immediacy. Immediacy is an amalgam of proxemic, kinesic, vocalic and linguistic features that signal psychological closeness or distance (Burgoon et al., 1985, 2022). In the case of this game, in which seating location, facing and body orientation, and proxemic behaviors were fixed and therefore excluded from consideration, we looked instead for facial pleasantness, smiling, expressivity and other facial signals of positive affect. Predicted vocalic indicators of liking were pitch variety, relaxed laughter, and rapid turn-switches, while linguistic indicators were predicted to include inclusive language like first person plurals and positive affect language.

Results showed numerous facial expression features correlating with liking and dislike, especially in the mouth, cheek, nose and brow regions. Vocally, only duration of turns-at-talk was positively associated with liking, and mean shimmer (a measure of vocal hoarseness) was negatively associated with liking. Pitch, loudness and other aspects of voice quality did not matter. Longer sentences, less hedging, and (unexpectedly), more dysfluencies were associated with perceived liking.

TABLE 1 | Significant linguistic, vocalic, and facial cues of dominance, affection, composure, involvement, similarity, and trust ($p < 0.1$).

Constructs	Linguistic Cues	Vocalic Cues	Facial Cues
Dominance-Non-dominance	Number of Words (+)	Turn-at-talk duration (+) Standard deviation of pitch (+) Average harmonic-to-noise ratio (+) Standard deviation of harmonic-to-noise ratio (–)	Mean cheek raiser (–) Mean lid tightener (+) Mean lip corner puller (+) Variance of brow lowerer (+) Variance of upper lip raiser (+) Variance of dimpler (–) Max inner brow raiser (+) Max outer brow raiser (–) Max brow lowerer (–) Max cheek raiser (+) Max lip corner puller (–) Max dimpler (+)
Affection-Hostility	Number of sentences (+) Hedge ratio (–)	Turn-at-talk duration (+) Average shimmer (–)	Mean cheek raiser (–) Mean dimpler (+) Mean lip tightener (+) Variance of brow lowerer (+) Variance of nose wrinkler (–) Variance of lip tightener (–) Max inner brow raiser (+) Max brow lowerer (–) Max cheek raiser (+) Max lid tightener (–) Max nose wrinkler (+) Max lip corner puller (–)
Composure-Nervousness	Disfluency ratio (–)	Average loudness (+) Average shimmer (–)	Mean upper lip raiser (–) Mean lip stretcher (+) Mean blink (+) Variance of brow lowerer (+) Variance of lip stretcher (–) Max brow lowerer (–) Max nose wrinkler (+) Max chin raiser (–)
Involvement-Detachment	Number of words (+) Number of sentences (+)	Turn-at-talk duration (+) Average shimmer (–)	Mean cheek raiser (–) Mean lid tightener (+) Mean nose wrinkler (+) Mean lip corner puller (+) Variance of brow lowerer (+) Variance of dimpler (–) Max brow lowerer (–) Max cheek raiser (+) Max lid tightener (–) Max dimpler (+)
Similarity-Dissimilarity	Number of sentences (+) Number of words (–)	Standard deviation of harmonic-to-noise ratio (+) Average shimmer (–) Standard deviation of shimmer (+)	Mean inner brow raiser (–) Mean outer brow raiser (+) Mean cheek raiser (–) Mean lip corner puller (+) Mean lip tightener (+) Variance of inner brow raiser (+) Variance of outer brow raiser (–) Variance of brow lowerer (+) Variance of cheek raiser (+) Variance of lip tightener (–) Variance of jaw drop (+) Max lid tightener (–) Max chin raiser (–)

(Continued)

TABLE 1 | (Continued)

Constructs	Linguistic Cues	Vocalic Cues	Facial Cues
Trust-Distrust	Number of sentences (+)	Turn-at-talk duration (+) Average shimmer (–)	Mean cheek raiser (–) Mean jaw drop (–) Variance of nose wrinkler (–) Variance of jaw drop (+) Max brow lowerer (–) Max lip corner puller (–) Max dimpler (+) Max lip suck (–)

Positive (and negative) signs in the parentheses indicate significant positive (or negative) unstandardized beta weights in regression analyses between the behavioral cue and the focal relational message construct.

Composure-Nervousness

Composure in the case of relational messages means signaling that one is comfortable, at ease and relaxed in the other's presence. Composure is manifested as facial and postural relaxation. Acoustically, composure presents as a more expressive and pleasant voice. The bipolar opposites of composure are signals of nervousness. In addition to higher anxiety being associated with speech dysfluencies like stuttering (Ezrati-Vinacour and Levin, 2004), nervousness may present in the form of rigid faces, voices, posture and heads; gaze avoidance; fidgeting or other adaptor (self-touching) gestures; softer vocal amplitude; higher pitch; more dysfluencies; and shorter and fewer turns-at-talk. Additionally, nervousness often conveys detachment or unpleasantness (Burgoon et al., 2021).

Results in this experiment showed that more fluent speakers were perceived as more composed, with higher average loudness and lower average shimmer, indicating that those who speak more loudly and less hoarsely are perceived as more composed; conversely, dysfluent, quieter and hoarser voices conveyed discomfort. In terms of facial behaviors, perceived composure (or nervousness) was positively (or negatively) associated with several features in the brow, eye, lip and chin regions, confirming the expectation that nervousness is shown particularly in the upper and lower action units of the face.

Involvement-Detachment

Involvement is a relational message that can have positive or negative connotations. Dillard et al. (1999) proposed that involvement is an intensifier dimension between competing meanings of dominance or affiliation, which could alter which set of features is associated with involvement. Coker and Burgoon (1987) analyzed over 50 features that could be associated with involvement, most either value-neutral or more tilted in favor of a positive sentiment.

Here, results showed that higher perceived involvement was associated with more words, sentences and longer turns-at-talk duration, indicating that perceived involvement increased with participation in the group conversation. Findings from the audio channel are consistent with Coker and Burgoon (1987), which showed greater involvement corresponded to fewer silences in speech, more vocal warmth and relaxation, but no effect of disfluency. Average magnitude and variability of pitch and loudness were non-significant, contrary to a previous finding that higher pitch, pitch range, and voice intensity are indicative of conversational involvement (Oertel et al., 2011). Meanwhile, perceived involvement was negatively associated with average shimmer. Additionally, significant facial cues included many in the eye, brow and cheek regions. Thus, facial activation played a significant role in expressing involvement.

Similarity-Dissimilarity

Interpersonal similarity measures the degree to which people share like attitudes, beliefs, personal characteristics, experiences, and so forth (Burgoon and Hale, 1984). Similarity promotes communication and bolsters influence (Krishnan and Hunt, 2021).

The results here showed that, linguistically, the number of sentences was a significant contributor to perceived similarity, while the number of words curiously detracted. Vocally, variability in shimmer had a positive effect on the similarity ratings, while mean shimmer was negatively related. Thus, less overall shimmer but more variability in shimmer expressed similarity. Additionally, perceived similarity was positively associated with the standard deviation of HNR (Harmonic to Noise Ratio), again a signal of variability. It is worth noting that two behavioral indicators, number of sentences and average shimmer, affected the similarity ratings and the trust ratings in the same direction, implying the close relationship between these two relational dimensions. The face model revealed a rich set of significant correlates with similarity, many involving variability or maximums and signifying that more active faces were read as greater similarity.

Trust-Distrust

As the glue that holds society together, trust plays an essential role in interpersonal (Golembiewski and McConkie, 1975) and commercial (Morgan and Hunt, 1994) relationships and consequently has attracted abundant scholarly attention. Trust fosters cooperation (Balliet and Van Lange, 2013) and reduces costs of social transactions (Dyer and Chu, 2003). Though the concept of trust has been investigated extensively, defining the construct remains a challenging task due to its multi-contextual nature. A typology derived from various definitions (McKnight and Chervany, 2000) suggests that benevolence, integrity, competence, and predictability are the defining characteristics of trust. A rich set of verbal and nonverbal cues, such as smile

(Centorrino et al., 2015), eye contact or gaze aversion (Bayliss and Tipper, 2006), facial expressivity (Krumhuber et al., 2007), voice pitch (McAleer et al., 2014), prosody dynamics (Chen et al., 2020), verbal politeness (Lam, 2011) and use of technical terms (Joiner et al., 2002) have been reported to convey interpersonal trust and promote cooperative behavior.

In the current study, we found that the greater number of sentences enhanced a participant's perceived trustworthiness, though the total amount of speech (i.e., words) had no such effect. The vocalic model showed that turn-at-talk duration, which contributes to the total amount of speech, also boosted perceived trustworthiness, corroborating the positive effect of sentence quantity. Meanwhile, average shimmer had a negative effect on perceived trustworthiness, indicating a less hoarse voice with less breathiness can stimulate trust. The face model produced mixed results. While speaking activity (reflected by the variance of jaw drop) and maximum magnitude of dimpler (a lower face muscle movement driven by smiling) increased perceived trustworthiness, the average level of cheek-raising, jaw-dropping, variance of nose-wrinkling, and maximum level of brow-lowering, lip corner-pulling and lip-sucking all negatively affected trust. Apparently, too much activity and adaptor behavior in the lip and cheek region diminished trust, contrary to the benefit of such vocal and facial activity in expressing involvement and similarity.

Perceived Veracity

One way to analyze the effect of the six relational dimensions on the outcome of perceived veracity is to use two-stage least squares regression with deception manipulation (i.e., players' role) as an instrumental variable. We operationalized perceived deceptiveness as the percentage of Villagers who regarded a player as a Spy. Results in the **Supplementary Material** show that the regression coefficients for all the relational dimensions are significantly negative, suggesting that players with higher perceived dominance, affection, composure, involvement, similarity (with Villager raters), and trustworthiness are less often judged as deceivers. Composure and affection have the largest effect sizes. Thus, players whose relational communication includes nonverbal and verbal signals that convey the least nervousness and engender the most liking are least likely to be suspected as Spies. This analysis demonstrates how the Brunswikian lens model links distal communication signals to meaningful psychological and emotional percepts of interaction to social outcomes of that interaction (e.g., perceived veracity).

DISCUSSION

Interactants in social contexts send and interpret relational messages using a broad array of verbal and nonverbal behaviors. Applying a modified Brunswikian lens model, we investigated how individuals form proximal percepts based on multimodal behavioral indicators.

We undertook the current approach to illustrate how multimodal signals can be combined to predict some focal variable of interest. Our indicators were not intended to be exhaustive but rather a sampling that could be incorporated into a Brunswikian lens model and thus demonstrate how perceptual and objective variables can be combined to predict whatever outcome is of interest, in this case, deception. Objective distal indicators combine to form proximal percepts; subjective percepts predict outcomes. Modeling social behavior in this manner makes clear the importance of distinguishing objective indicators from subjective perceptions. Distal indicators usually represent more objective, discrete, and microscopic variables that are often regarded as ground truth, whereas percepts are the subjective, macroscopic, interpretive layer of judgments that are formed from the distal cues. Percepts are the intermediate judgment that predicts outcomes of interest. In the case of deception, distal clues might include objective behaviors such as eye blinks and immobile facial muscles that lead to the percept nervousness and thus to the conclusion that the speaker's frozen, impassive face conveys deceptiveness.

The Brunswikian lens model is a very flexible model that permits choosing few or many indicators of a given type (e.g., facial expressiveness signals), depending on the research question of interest. It also permits beginning with the most distal physical and physiological indicators, then working to the more proximal interior psychological and emotional states to arrive at a predicted behavioral outcome, or instead beginning with the psychological emotional and cognitive states, such as emotional stress and cognitive overload, then working backward to the objective behaviors that account for those cognitive-emotional states. Either the distal indicators or proximal percepts can be used to predict ultimate attributions. Here, where our interest was in deception, the analysis showed that relational messages are one way to conceptualize the implicit social meanings that are the percepts predicting deceptiveness.

Important from a communication (Subrahmanian et al., 2021) standpoint is that all three modalities—linguistic, vocalic and kinesic–contribute variance to the final prediction. The model encourages deeper investigation into what objective indicators contribute to the relational *topoi* that are so deeply embedded in the process of interpersonal communication. An example: A member of a decision-making group may characterize another member's communication as involved, expressing commonality and similarity, and engendering trust. But these interpretive characterizations leave unanswered what behaviors contribute to those perceptions. AI models can probe what distal signals combine to form these relational messages and lead to perceptions that another is credible or deceptive.

Our findings open up many avenues for future CS research into relational communication. First, the CS community could apply state-of-the-art machine learning methods to predict relational messages. These predictions would facilitate a better understanding of dynamic human interactions. They might show, for instance, how certain actions lead to distrust among group members and account for deterioration of a sense of homophily and liking as the group's interaction unfolds. Or they might identify what group members' behaviors promote trust and ultimately, to favorable decisions. Such analysis could assist with decision making scenarios such as business negotiations or discussions of pandemic relief programs. One possible

direction is to make inferences on multiple non-orthogonal relational messages through transfer learning (Zhuang et al., 2020). Another direction would be to apply time series analysis to model long interactions, which would allow predictions of dynamic changes in these relational messages over time. Besides making predictions, recent developments in explainable artificial intelligence (Adadi and Berrada, 2018) would help interpret the models and benefit the social science community in identifying more nuanced behavioral indicators of relational messages and in developing relevant theories. Presenting intelligible explanations also increases users' trust (Gunning et al., 2019). These advancements in CS research present exciting opportunities to further investigate relational messages during human interactions and create synergy between the CS and social science communities.

Second, it would be of great value for the CS community to develop more powerful tools for analyzing behaviors of multiple modalities. Besides the linguistic, vocalic, and facial features, other physiological and behavioral signals, such as gestures and posture, would also be valuable to investigate. In addition, an integrated tool for processing speech, voice, and video in real-time would be beneficial. Although real-time speech (Gao et al., 2019), voice (Acharya et al., 2018), and video processing (Ananthanarayanan et al., 2017) and their integration (Kose and Saraclar, 2021) have been widely studied in computer science, the analysis of physiological and behavioral signals in psychological, emotional, and cognitive states and relational messages presents a new and interesting path, especially for real-time applications (e.g., decision support in business negotiations). Another useful future direction is to harness the power of computer-based techniques to perform real-time audio and video quality checks for better data inputs in a non-laboratory setting. Although we have taken extensive actions to ensure the quality of data collected in labs, unexpected factors, such as uneven lights and background noise, may distort the data collected in the field or in online experiments. A real-time data input quality checker would provide guidance on high-quality data collection and reduce the influence from unforeseen human and environmental matters. We urge further developments in these automated tools for better data collection and analysis.

Although computer scientists and social scientists routinely call for more cross-disciplinary collaboration, such lip service is rarely accompanied by true integration of the work. The Brunswikian lens model offers a productive vehicle for creating that collaboration and integration.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because the project is in progress. Upon completion of the grant, the United States Army Research Office (project sponsor) is committed to making the multimodal, de-identified data available. Contact should be made to the respective investigators for the data of interest. Requests to access the datasets should be directed to VS Subrahmanian, Computer Science, Northwestern University, Evanston, IL, United States.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The Army Research Laboratory Human Research Protection Office. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JB along with other investigators designed the experiments, planned the data analysis, and wrote a portion of the current manuscript. RW, XC, and TG worked collaboratively to conduct data analysis and wrote one relational theme section. BD conducted much of the data set preparation and initial analysis, wrote software to conduct the game, and traveled internationally to conduct the game. All authors contributed to the article and approved the submitted version.

FUNDING

This project is part of a Multi-University Research Initiative. The five-year project is a collaboration among the University of Arizona, Dartmouth University, University of California, Santa Barbara, Rutgers University, Stanford University, and University of Maryland to investigate Socio-Cultural Attitudinal Networks. The research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-16-1-0342 (PIs: VS Subrahmanian and Judee Burgoon). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.781487/full#supplementary-material

REFERENCES

Acharya, J., Patil, A., Li, X., Chen, Y., Liu, S. C., and Basu, A. (2018). A comparison of low-complexity real-time feature extraction for neuromorphic speech recognition. *Front. Neurosci.* 12:160. doi: 10.3389/fnins.2018.00160

Adadi, A., and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 6, 52138–52160. doi: 10.1109/access.2018.2870052

Ananthanarayanan, G., Bahl, P., Bodík, P., Chintalapudi, K., Philipose, M., Ravindranath, L., et al. (2017). Real-time video analytics: the killer app for edge computing. *Computer* 50, 58–67. doi: 10.1109/mc.2017.3641638

Balliet, D., and Van Lange, P. A. (2013). Trust, conflict, and cooperation: a meta-analysis. *Psychol. Bull.* 139, 1090–1112. doi: 10.1037/a0030939

Baltrušaitis, T., Mahmoud, M., and Robinson, P. (2015). "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *Proceedings of the 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 6, (Ljubljana: IEEE), 1–6.

Baltrušaitis, T., Zadeh, A., Lim, Y. C., and Morency, L.-P. (2018). "Openface 2.0: facial behavior analysis toolkit," in *Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, (Xi'an : IEEE), 59–66.

Bayliss, A. P., and Tipper, S. P. (2006). Predictive gaze cues and personality judgments: should eye trust you? *Psychol. Sci.* 17, 514–520. doi: 10.1111/j.1467-9280.2006.01737.x

Bernieri, F. J., Gillis, J. S., Davis, J. M., and Grahe, J. E. (1996). Dyad rapport and the accuracy of its judgment across situations: a lens model analysis. *J. Pers. Soc. Psychol.* 71, 110–129.

Burgoon, J. K., and Dunbar, N. E. (2006). "Dominance, power and influence," in *The SAGE Handbook of Nonverbal Communication*, eds V. Manusov and M. Patterson (Thousand Oaks, CA: Sage), 279–298.

Burgoon, J. K., and Hale, J. L. (1984). The fundamental topoi of relational communication. *Commun. Monogr.* 51, 193–214. doi: 10.1080/03637758409390195

Burgoon, J. K., and Hale, J. L. (1987). Validation and measurement of the fundamental themes of relational communication. *Commun. Monogr.* 54, 19–41. doi: 10.1080/03637758709390214

Burgoon, J. K., Manusov, V., and Guerrero, L. K. (2022). *Nonverbal Communication*, 2nd Edn. London: Routledge.

Burgoon, J. K., Manusov, V., Mineo, P., and Hale, J. L. (1985). Effects of eye gaze on hiring, credibility, attraction and relational message interpretation. *J. Nonverbal Behav.* 9, 133–146. doi: 10.1007/BF01000735

Burgoon, J. K., Wang, X., Chen, X., Pentland, S. J., and Dunbar, N. E. (2021). Nonverbal behaviors "speak" relational messages of dominance, trust, and composure. *Front. Psychol.* 12:624177. doi: 10.3389/fpsyg.2021.624177

Centorrino, S., Djemai, E., Hopfensitz, A., Milinski, M., and Seabright, P. (2015). Honest signaling in trust interactions: smiles rated as genuine induce trust and signal higher earning opportunities. *Evol. Hum. Behav.* 36, 8–16. doi: 10.1016/j.evolhumbehav.2014.08.001

Chen, X. L., Ita Levitan, S., Levine, M., Mandic, M., and Hirschberg, J. (2020). Acoustic-prosodic and lexical cues to deception and trust: deciphering how people detect lies. *Trans. Assoc. Comput. Linguistic.* 8, 199–214. doi: 10.1162/tacl_a_00311

Coker, D. A., and Burgoon, J. (1987). The nature of conversational involvement and nonverbal encoding patterns. *Hum. Commun. Res.* 13, 463–494. doi: 10.1111/j.1468-2958.1987.tb00115.x

Dillard, J. P., Solomon, D. H., and Palmer, M. T. (1999). Structuring the concept of relational communication. *Commun. Monogr.* 66, 49–65. doi: 10.1080/03637759909376462

Dorn, B., Dunbar, N. E., Burgoon, J. K., Nunamaker, J. F., Giles, M., Walls, B., et al. (2021). "A system for multi-person, multi-modal data collection in behavioral information systems," in *Detecting Trust and Deception in Group Interaction*, eds V. S. Subrahmanian, J. K. Burgoon, and N. E. Dunbar (Berlin: Springer), 57–73.

Dunbar, N. E., Dorn, B., Hansia, M., Ford, B., Giles, M., Metzger, M., et al. (2021). "Dominance in groups: how dyadic power theory can apply to group discussions," in *Detecting Trust and Deception in Group Interaction*, eds V. S. Subrahmanian, J. K. Burgoon, and N. E. Dunbar (Berlin: Springer), 75–97.

Dyer, J. H., and Chu, W. (2003). The role of trustworthiness in reducing transaction costs and improving performance: empirical evidence from the United States, Japan, and Korea. *Organ. Sci.* 14, 57–68.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). "Opensmile: the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, (New York, NY: ACM), 1459–1462.

Ezrati-Vinacour, R., and Levin, I. (2004). The relationship between anxiety and stuttering: a multidimensional approach. *J. Fluency Disord.* 29, 135–148. doi: 10.1016/j.jfludis.2004.02.003

Gao, C., Braun, S., Kiselev, I., Anumula, J., Delbruck, T., and Liu, S. C. (2019). "Real-time speech recognition for IoT purpose using a delta recurrent neural network accelerator," in *Proceedings of the 2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, (Sapporo: IEEE), 1–5.

Giles, M., Hansia, M., Metzger, M., and Dunbar, N. E. (2021). *Detecting Trust and Deception in Group Interaction*. Cham: Springer, 98–136.

Golembiewski, R. T., and McConkie, M. (1975). "The centrality of interpersonal trust in group processes," in *Theories of Group Processes*, ed. C. L. Cooper (Wiley), 131–185.

Guerrero, L. K., Andersen, P. A., and Afifi, W. A. (2017). *Close Encounters: Communication in Relationships.* Thousand Oaks, CA: Sage.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., and Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Sci. Robot.* 4:eaay7120.

Hartwig, M., and Bond, C. F. Jr. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychol. Bull.* 137, 643–659. doi: 10.1037/a0023589

Hawes, L. C. (1973). Elements of a model for communication processes. *Q. J. Speech* 59, 11–21. doi: 10.1080/00335637309383149

Joiner, T. A., Leveson, L., and Langfield-Smith, K. (2002). Technical language, advice understandability, and perceptions of expertise and trustworthiness: the case of the financial planner. *Austr. J. Manag.* 27, 25–43. doi: 10.1177/031289620202700102

Kose, O. D., and Saraclar, M. (2021). Multimodal representations for synchronized speech and real-time MRI video processing. *IEEE/ACM Trans. Audio Speech Lang. Process.* 29, 1912–1924. doi: 10.1109/taslp.2021.3084099

Krishnan, A., and Hunt, D. S. (2021). Nonverbal cues and perceptions of personality and homophily in synchronous mediated communication. *Inf. Commun. Soc.* 24, 85–101. doi: 10.1080/1369118x.2019.1635183

Krumhuber, E., Manstead, A. S., Cosker, D., Marshall, D., Rosin, P. L., and Kappas, A. (2007). Facial dynamics as indicators of trustworthiness and cooperative behavior. *Emotion* 7, 730–735. doi: 10.1037/1528-3542.7.4.730

Lam, C. (2011). Linguistic politeness in student-team emails: its impact on trust between leaders and members. *IEEE Trans. Profess. Commun.* 54, 360–375.

Likert, R. (1932). A technique for the measurement of attitudes. *Arch. Psychol.* 22:55.

McAleer, P., Todorov, A., and Belin, P. (2014). How do you say 'hello'? Personality impressions from brief novel voices. *PLoS One* 9:e90779. doi: 10.1371/journal.pone.0090779

McKnight, D. H., and Chervany, N. L. (2000). "What is trust? A conceptual analysis and an interdisciplinary model," in *Proceedings of the AMCIS Americas Conference on Information Systems*, Vol. 382, (Long Beach, CA: AMCIS), 827–833.

Moffitt, K. C., Giboney, J. S., Ehrhardt, E., Burgoon, J. K., Nunamaker, J. F., Jensen, M., et al. (2012). "Structured programming for linguistic cue extraction (SPLICE)," in *Proceedings of the HICSS-45 Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium*, (Washington, DC: Computer Society Press), 103–108.

Morgan, R. M., and Hunt, S. D. (1994). The commitment-trust theory of relationship marketing. *J. Market.* 58, 20–38. doi: 10.1089/cyber.2012.0348

Oertel, C., De Looze, C., Scherer, S., Windmann, A., Wagner, P., and Campbell, N. (2011). "Towards the automatic detection of involvement in conversation," in *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues*, eds A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud, and A. Nijholt (Berlin: Springer), 163–170. doi: 10.1016/j.csl.2015.08.003

Pentland, S. J., Spitzley, L., Chen, X., (Rebecca) Wang, X., Burgoon, J. K., and Nunamaker, J. F. (2021). "Behavioral indicators of dominance in an adversarial group negotiation game," in *Detecting Trust and Deception in Group Interaction*, eds V. S. Subrahmanian, J. K. Burgoon, and N. E. Dunbar (Berlin: Springer), 99–122. doi: 10.1007/978-3-030-54383-9_6

Scherer, K. R. (2003). Vocal communication of emotion: a review of research paradigms. *Speech Commun.* 40, 227–256. doi: 10.1016/s0167-6393(02)00084-5

Subrahmanian, V. S., Burgoon, J. K., and Dunbar, N. E. (eds) (2021). *Detecting Trust and Deception in Group Interaction*. Berlin: Springer.

Ting-Toomey, S., Yee-Jung, K. K., Shapiro, R. B., Garcia, W., Wright, T. J., and Oetzel, J. G. (2000). Ethnic/cultural identity salience and conflict styles in four US ethnic groups. *Int. J. Int. Relat.* 24, 47–81.

Zhou, L., Burgoon, J. K., Zhang, D., and Nunamaker, J. F. (2004). Language dominance in interpersonal deception in computer-mediated communication. *Comput. Hum. Behav.* 20, 381–402.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., et al. (2020). A comprehensive survey on transfer learning. *Proc. IEEE* 109, 43–76.

Conflict of Interest: JB is a principal in Discern Science International, a for-profit entity that conducts credibility analysis.

The remaining authors declare the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Computer Science

ORIGINAL RESEARCH
published: 21 March 2022
doi: 10.3389/fcomp.2022.780580

Multimodal EEG and Eye Tracking Feature Fusion Approaches for Attention Classification in Hybrid BCIs

Lisa-Marie Vortmann[1]*, *Simon Ceh*[2] *and Felix Putze*[1]

[1] *Cognitive Systems Lab, Department of Mathematics and Computer Science, University of Bremen, Bremen, Germany,*
[2] *Department of Differential Psychology, University of Graz, Graz, Austria*

Often, various modalities capture distinct aspects of particular mental states or activities. While machine learning algorithms can reliably predict numerous aspects of human cognition and behavior using a single modality, they can benefit from the combination of multiple modalities. This is why hybrid BCIs are gaining popularity. However, it is not always straightforward to combine features from a multimodal dataset. Along with the method for generating the features, one must decide when the modalities should be combined during the classification process. We compare unimodal EEG and eye tracking classification of internally and externally directed attention to multimodal approaches for early, middle, and late fusion in this study. On a binary dataset with a chance level of 0.5, late fusion of the data achieves the highest classification accuracy of 0.609–0.675 (95%-confidence interval). In general, the results indicate that for these modalities, middle or late fusion approaches are better suited than early fusion approaches. Additional validation of the observed trend will require the use of additional datasets, alternative feature generation mechanisms, decision rules, and neural network designs. We conclude with a set of premises that need to be considered when deciding on a multimodal attentional state classification approach.

Keywords: feature fusion, convolutional neural networks, attention, eye tracking, EEG, Markov Transition Fields, Gramian Angular Fields

OPEN ACCESS

Edited by:
Laura M. Ferrari,
Université Côte d'Azur, France

Reviewed by:
Maryam S. Mirian,
University of British Columbia, Canada
Siyuan Chen,
University of New South Wales, Australia

***Correspondence:**
Lisa-Marie Vortmann
vortmann@uni-bremen.de

Specialty section:
This article was submitted to Mobile and Ubiquitous Computing, a section of the journal Frontiers in Computer Science

Received: *21 September 2021*
Accepted: *21 February 2022*
Published: *21 March 2022*

Citation:
Vortmann L-M, Ceh S and Putze F (2022) Multimodal EEG and Eye Tracking Feature Fusion Approaches for Attention Classification in Hybrid BCIs. Front. Comput. Sci. 4:780580. doi: 10.3389/fcomp.2022.780580

1. INTRODUCTION

Human-machine interaction is becoming increasingly ubiquitous. In our daily lives, we want to seamlessly incorporate technology and thus rely on usability. By integrating implicit input mechanisms, the synergy between users and machines is further enhanced: These enable a system to infer information about the user without the user taking any explicit action, such as pressing a button or speaking a command, and modify their behavior accordingly.

One way of implementing implicit input mechanisms is *via* biosignal-based recognition of cognitive states. Biosignal-based recognition of cognitive states or activities in humans is a broad research field because of the manifold options for input signals, classification algorithms, and possible applications. For instance, a Brain-Computer Interface (BCI) can predict a user's attentional state from electroencephalographic (EEG) data and adapt the system's behavior using machine learning (Vortmann and Putze, 2020). Certain modalities are more suited to certain

applications and scopes than others, but for the majority of applications, more than one possible input signal can be considered. For instance, brain activity can be supported by eye gaze behavior. Such systems are commonly referred to as hybrid BCIs (Kim et al., 2015).

The fundamental premise of such multimodal approaches in the context of BCI machine learning is that the two modalities may capture distinct aspects of the user state and thus complement one another. While using a single modality can result in reliable classification accuracy, combining two or more modalities can enhance the system's recognition power and robustness, thereby improving its overall performance. D'Mello and Kory (2012) demonstrated in a review of 30 studies that multimodal classification yielded on average 8.12% improvement over the unimodal classifiers. Possible aims of the combination are to correct for temporally noisy data, resolve ambiguity, or the exploitation of correlations (Baltrušaitis et al., 2018).

In this work, we want to systematically explore the combination of EEG and eye tracking data for the classification of internally and externally directed attention. The result of such a classification could be used in a BCI to adapt the system to the user state.

1.1. Multimodal Feature Fusion

Biosignal data is heterogeneous in nature due to its inherent properties and recording mechanisms. For example, brain activity can be recorded using an EEG, which measures electrophysiological changes on the scalp and is usually recorded in microvolt, whereas eye gaze behavior is recorded by eye tracking devices that measure pupil dilation and infer gaze coordinates. During unimodal approaches, the feature extraction is either explicitly designed to generate meaningful features from the data, or the classification process implicitly learns to extract modality- and task-specific features (Kim et al., 2020). A combination of several modalities for the classification process is therefore not trivial.

The first opportunity to merge modalities is before the beginning of the classification process. Such **early fusion** approaches combine the biosignals on a feature level (Cheng et al., 2020). The joint representation of previously extracted meaningful features or preprocessed raw data presupposes that all modalities can be aligned properly for classification. This approach allows for the learning of cross-modal correlations during the classification process, but requires concatenation of the inputs and limits the extraction of modality-specific features.

Oppositely, **late fusion** approaches merge the modalities at the end of the classification process. The inputs are separately processed in individually tailored steps, typically until the prediction of individual labels. The fusion happens on the decision level based on the multiple predictions (Cheng et al., 2020). In Mangai et al. (2010), this was discussed as classifier combination because several classifiers are trained individually per modality before the results of the classifiers are combined (or one classifier is selected as overall output). The authors suggested different approaches how to choose the classifier combination, based on the available individual output formats per modality classifier. For instance, if each classifier predicts only a class label, an odd number of classifiers should be chosen to allow for (weighted) majority votes for the final output. In other cases, the classifiers could produce vectors in which the values represent the support for each label. Such certainty evaluations per modality classifier allow for a more sophisticated assessment of the final combined multimodal output. A decision rule has to define how the individual predictions are combined for the final prediction. This rules can either be set or learned using machine learning. The setting of a decision rule requires good *a priori* knowledge on the expected results, while machine learning based late fusion requires a large amount of data to enable the training of such decision rule. Especially regarding the proposed attention classification biosignal data, such large datasets are often not available and rule-based late fusion approaches should be favored. An apparent advantage of late fusion is the power of a tailored classification processes, whereas the shortcoming lies in the exploitation of modality correlations (Polikar, 2012).

One can also steer a middle course in fusing the modalities in the middle of the classification process. The idea of **middle fusion** (or halfway fusion) approaches is to first process the modalities individually but merge intermediate results as soon as possible, followed by further classification steps. In terms of neural networks, the first layers process the distinct inputs simultaneously before concatenating the layers' outputs for the following shared layers. The advantage of this fusion approach is that the modalities could first be processed tailored to their individual properties before exploiting the cross-correlations and arriving at a joint prediction.

1.2. EEG and Eye Tracking Based Mental State Detection

Hybrid BCIs have been used to detect a variety of mental states by analyzing eye movement patterns rather than relying on the user's explicit gaze behavior for direction control or target selection. As mentioned before, MI is a suitable use case for BCIs in general. Dong et al. (2015) used the natural gaze behavior of the participants to smooth the noisy predictions that resulted only from EEG motor imagery tasks. Cheng et al. (2020) explicitly compared late and early fusion of the multimodal features for their MI task. For the feature level fusion, they remarked that EEG and eye tracking data are so dissimilar, fusing them is not trivial and requires several preprocessing steps. For the decision level fusion, they used a decision rule based on the D-S evidence theory (Zhang et al., 2018). They found that feature fusion outperforms single modalities and that late fusion outperforms early fusion of eye tracking and EEG data.

In Guo et al. (2019), the authors investigate emotion recognition using a multimodal approach. They combine eye tracking and EEG data and classify the input after an early fusion using a deep neural network model that combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. For the early fusion of the modalities, they apply a Bimodal Deep AutoEncoder (BDAE) that extracts a high-level representation of features. This approach was first presented in Liu et al. (2016). Another early fusion approach for emotion recognition was examined in Lu et al. (2015). They fused 33

different features from eye movement data with 62 channel EEG signals and achieved 87.59% accuracy in classifying three emotions. Zheng et al. (2014) combined EEG signals and pupil dilation either in an early fusion approach or in a late fusion approach and found that both improved the performance of the emotion recognition model compared to unimodal approaches with a slightly higher accuracy for early fusion. Later, the authors presented a multimodal emotion recognition framework called EmotionMeter that also combines EEG and eye tracking data to recognize emotions in real-world applications. They successfully classified four different emotions with an accuracy of more than 85% using a multimodal neural network, outperforming both single modalities (Zheng et al., 2019). Another study on multimodal emotion recognition was conduced by López-Gil et al. (2016) who found that combining different signal sources on the feature level enables the detection of self-regulatory behavior more effectively than only using EEG data. Most recently, Wu et al. (2021) fused EEG and eye tracking data for emotion classification using effective deep learning for a gradient neural network. They report an 88% accuracy for the recognition of eight emotions.

The authors of Zhu et al. (2020) demonstrated that when eye movement and EEG data are combined for the detection of depression, a content-based ensemble method outperforms traditional approaches. The mental workload level is another cognitive state that can be classified using the proposed multimodal data. Debie et al. (2021) state in their review, that the combined features outperform single modalities for workload assessments. For example, Lobo et al. (2016) fused previously extracted eye tracking and EEG features on the feature level before training person-dependent and person-independent classifiers on them. They found that an almost perfect classification performance could be achieved for individual classifiers while independent classifiers only reached a lot worse accuracy.

1.3. Attentional State Classification

This study will examine different feature fusion strategies for a multimodal classification of EEG and eye tracking data to recognize internally and externally directed attention in a paradigm that manipulates internal/external attention demands. In general, attentional mechanisms are applied to filter the vast amount of available information at every moment for a better focus on relevant goals. Internally directed attention refers to a focus on information that is independent of sensory input, such as thoughts, memories, or mental arithmetic. It can occur deliberately (e.g., planning; Spreng et al., 2010) or spontaneously (e.g., mind wandering; Smallwood and Schooler, 2006). Externally directed attention instead describes a state of attentiveness to sensory input produced by the surroundings (Chun et al., 2011). Because concurrent self-evaluation of attentiveness to internal/external states while completing particular tasks would directly interfere with the direction of attention itself, a common approach is to ask participants in retrospect. Arguably, a system that would concurrently monitor the attentional state without interfering with the user may be better suited for application.

The suitability of eye tracking data for this classification task was shown by Annerer-Walcher et al. (2021) who achieved a classification accuracy of 69% for 4 s windows of raw eye tracking data. They compared gaze-specific properties and found that blinks, pupil diameter variance, and fixation disparity variance indicated differences in attentional direction. In Putze et al. (2016) and Vortmann et al. (2019a), the authors showed that such attentional differences can also be classified from EEG in different settings. They achieved 74.3% for 2 s windows and 85% for 13 s windows, respectively.

Eye tracking and EEG data have been collected simultaneously in several studies on attention (e.g., Vortmann and Putze, 2021). Kulke et al. (2016) investigated neural differences between covert and overt attention using EEG. The eye gaze was analyzed to control the correct labeling of the data. Dimigen et al. (2011) performed a co-registration of eye movement and EEG data for reading tasks and analyzed the fixation-related potentials. However, in these studies, the modalities were not combined but used for different purposes during the analysis.

To the best of our knowledge, the only paper that addresses feature fusion of EEG and eye tracking data for internally and externally directed attention in the context of attention classification is by Vortmann et al. (2019b). The authors implemented a real-time system for the attentional state classification and found that a late fusion approach with a decision rule improves the classification result of both single modalities. For 1.5 s data windows, the classification accuracy for the EEG data ranged between 0.56 and 0.81, for eye tracking data between 0.46 and 0.78 and for the late fusion approach between 0.58 and 0.86, calculated for 10 participant and a chance level of 0.5.

This work will systematically compare the unimodal approaches for EEG and eye tracking data with early, middle, and late fusion multimodal approaches for internally and externally directed attention.

2. METHODS

A dataset of 36 participants was analyzed for within-person classification accuracies of different multimodal neural networks.

2.1. Data

The data used in this study was recorded by Ceh et al. (2020)[1]. It encompasses EEG and eye tracking recordings of 36 participants (24 female, 12 male; age: $M = 24$ $SD = 2.72$; all right-handed; four had corrected-to-normal vision). The data set was chosen because the EEG and the eye tracking data were sampled with the same sampling rate. This makes the temporal alignment for the early fusion approaches easier and more accurate. The data collection was performed in a controlled laboratory setup which results in higher quality data and less confounding factors compared to more flexible setups that require, for instance, free movements (Vortmann and Putze, 2020).

[1] Publicly available at 10.17605/OSF.IO/5U6R9.

2.1.1. Task

During the recording, the participants had to perform two different tasks under two different conditions each. For all tasks, a meaningful German word of four letters was presented. For one task, the participants had to create **anagrams** of the word (i.e., "ROBE" is transformed to "BORE"). For the other task, a four-word long **sentence** had to be generated, each word starting with one of the four letters from the presented word (i.e., "ROBE" is transformed to "Robert observes eye behavior"). The employed paradigm builds on both a convergent (anagram) and divergent (sentence generation) thinking task and has been used in several studies investigating the effect of attention demands in the visual domain (Benedek et al., 2011, 2016, 2017; Ceh et al., 2020, 2021). Within the tasks, the attentional demands are manipulated using stimulus masking: in half of all trials, the stimulus is masked after a short processing period (500 ms), requiring participants to keep and manipulate the word in their minds. This enforces completion of the task relying on internally directed attention. In the other half of all trials, the stimulus word is continuously available (20 s), allowing for continuous retrieval using external sensory processing. The paradigm thus differentiates convergent and divergent thinking in a more internal vs. external attentional setting. For a detailed description of the task, see the original article.

2.1.2. Conditions

The effects of manipulating attention using these tasks were previously looked at for EEG (Benedek et al., 2011), fMRI (Benedek et al., 2014), and eye tracking (Benedek et al., 2017) data, or a combination of EEG and ET (Ceh et al., 2020), and fMRI and eye tracking (Ceh et al., 2021) data. Across these studies, the investigators found robust differences between the internal and external conditions on the level of eye behavior (e.g., increased pupil diameter during internally directed cognition; Benedek et al., 2017; Ceh et al., 2020, 2021), EEG (e.g., relatively higher alpha power over parieto-occipital regions during internally directed cognition; Benedek et al., 2011; Ceh et al., 2020), and fMRI (e.g., internally directed cognition was associated with activity in regions related to visual imagery, while externally directed cognition recruited regions implicated in visual perception; Benedek et al., 2016; Ceh et al., 2021). The observed attention effects were highly consistent across both tasks in all studies (i.e., across different modalities).

In this study, we will not differentiate between the two tasks. The classification will be based on masked (internally directed attention) and unmasked (externally directed attention) stimuli. Each participant performed 44 trials of each condition (chance level for the classification = 0.5).

2.1.3. Recordings

EEG was recorded with a BrainAmp amplifier by Brain Products GmbH with a sampling rate of 1,000 Hz using 19 active electrodes, positioned according to the 10-20 system in the following positions: Fp1, Fp2, F7, F3, Fz, F4, F8, T7, C3, Cz, C4, T8, P7, P3, Pz, P4, P8, O1, and O2. Additionally, three electrooculogram electrodes were included (left and right of the eyes, and adjacent to the radix nasi). References were placed on the left and right mastoid and the ground electrode was placed centrally on the forehead. Impedances were kept below 30 kOhm.

The eye tracking data was recorded using an EyeLink 1000 Plus eye tracker by SR Research Ltd. with a sampling rate of 1,000 Hz. For a more detailed description of the experimental setup and procedure (see Ceh et al., 2020).

2.2. Preprocessing

Simple preprocessing steps were applied to both data input sets to reduce the noise in the data. The classification will be performed per participant, with participant-dependently trained classifiers. Thus, correcting data to account for inter-individual differences is not necessary.

For the **eye tracking**, the X- and Y- coordinates and the pupil diameter of the left and the right eye were cleaned from non-existing values by dropping the respective samples. Binocular blinks (as defined by the eye tracker's built-in detection algorithm) were also excluded. The X- and Y-coordinates recorded by the eye tracker can be interpreted as the current gaze position relative to the screen.

The **EEG data** were processed using the MNE toolbox by Gramfort et al. (2013). First, the data was bandpass-filtered between 1 and 45 Hz using windowed FIR filters. An additional notch filter was applied at 50 Hz (power-line noise). Afterward, the data was re-referenced to average. Bad channels or epochs were not excluded from the data.

For both data sets, each trial was cut into four non-overlapping 3 s windows: 3–6, 7–10, 11–14, and 15–18 s after trial onset. The first seconds of each trial were not used to avoid an effect of the masking process in the data. In total, each participant's data set contained $4^*44 = 176$ data windows. No baseline correction was applied.

We generated two **feature sets** for each modality. As argued earlier, for early feature fusion approaches, the input format from both modalities must be temporally compatible so it can be combined. The data synchronization was performed on the basis of the available timestamps. Missing values were dropped for both modalities. The first feature set is the plain **preprocessed time series**, without any further computations or feature extraction steps. This raw input has been proven suitable for EEG data classification (Schirrmeister et al., 2017). To generate the second feature set, we followed an approach introduced in Wang and Oates (2015). The authors suggest transforming time-series data into **representative images** that convolutional neural networks can classify. The first algorithm for the image generation is called Markov Transition Field (MTF). MTFs represent transition probabilities between quantiles of the data. As a second algorithm, they suggest Gramian Angular Summation Fields (GASF), which visualizes the distances between polar-coordinates of the time series data. They argue that both approaches keep spatial and temporal information about the data. The application of this feature generation approach for eye tracking data during internally and externally directed attention was implemented by Vortmann et al. (2021). They were able to show that the imaging time-series approach with a convolutional neural net achieve higher classification accuracies than classical eye gaze-specific features.

TABLE 1 | Shallow FBCSP Convolutional Neural Network structure (shallow FBCSP CNN) from Schirrmeister et al. (2017), implemented using the braindecode toolbox by Schirrmeister et al. (2017).

Layer name	Type	Properties
conv_time	Conv2d	Out = 40, kernel_size = (25, 1), stride = (1, 1)
conv_spat	Conv2d	Out = 40, kernel_size = (1, 23), stride = (1, 1)
bnorm	BatchNorm2d	Out = 40, eps = 1e-05, momentum = 0.1
pool	AvgPool2d	Kernel_size = (75, 1), stride = (15, 1), padding = 0
drop	Dropout	$p = 0.5$
conv_classifier	Conv2d	Out = 2, kernel_size = (194, 1), stride = (1, 1)

TABLE 2 | Simple Convolutional Neural Network structure (simple CNN) similar to Vortmann et al. (2021), implemented using the PyTorch library by Paszke et al. (2019). fc, fully connected.

Layer name	Type	Properties
conv1	Conv2d	Out = 60, kernel_size = (5, 5), stride = (1, 1)
conv2	Conv2d	Out = 120, kernel_size = (5, 5), stride = (1, 1)
conv_dropout	Dropout2d	$p = 0.5$
fc1	Linear	In = 9,720, out = 500
fc2	Linear	In = 500, out = 120
fc3	Linear	In = 120, out = 20
fc4	Linear	In = 20, out = 2

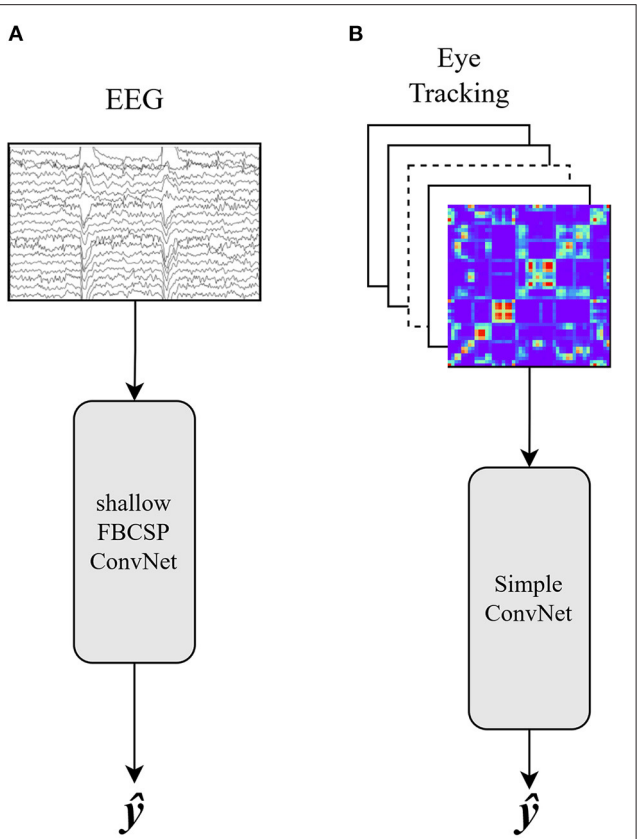


FIGURE 1 | (A) Single modality classification approach for EEG data using the shallow FBCSP Convolutional Neural Network. The preprocessed time-series is used as the input. **(B)** Single modality classification approach for eye tracking data using a Simple Convolutional Neural Network. The stacked images represent the Markov Transition Fields and Gramian Angular Fields computed on the raw data (see section 2.2). Both: $\hat{y}$ is the class prediction.

We calculated the MTF and the GASF image with 48x48 pixels for each channel in the data, resulting in 12 images for the eye tracking data: 2 images $*$ 2 eyes $*$ [x-coordinate, y-coordinate, pupil diameter] and 44 images for the EEG data: 2 images $*$ (22 EEG channels + 3 EOG channels). This results in an image matrix of 56 images per trial.

2.3. Classifier

The classification was performed in a person-dependent manner, resulting in an individual model for each participant. We used two different convolutional neural networks as classification algorithms, one for each feature set (time-series features and image features). Schirrmeister et al. (2017) introduced a shallow CNN that was inspired by Filterbank Common Spatial Pattern (FBCSP) analysis for EEG time-series. The layers of the network can be seen in **Table 1**. This **shallow FBCSP CNN** will be used to classify the time series feature set of both modalities. As optimizer, we used the AdamW optimizer (Loshchilov and Hutter, 2017), null loss, a learning rate of $0.0625 * 0.01$, and a weight decay of $0.5*0.0001$.

The second neural network that we used for the image features was the **simple CNN** adapted from Vortmann et al. (2021). **Table 2** describes the network structure in detail. This time, the Adam optimizer (Kingma and Ba, 2014), cross-entropy loss, a learning rate of 0.0001, and no weight decay were used. The label prediction the maximum of the softmax of the output layer was calculated.

In the first step, we classified the data using single modality approaches. The data were randomly split into training and testing data, using 33% for testing (stratified). We trained for a maximum of 30 epochs with a batch size of 40. Early stopping was applied if the classification accuracy on the training data was above 95% for more than five epochs to avoid overfitting.

The EEG data were classified using the time series feature set and the shallow FBCSP CNN (see **Figure 1A**). The eye tracking data were classified using the image feature set and a simple CNN (see **Figure 1B**). All evaluations are based on the network accuracy tested on the test data. Because of the equal distribution of the two conditions, the chance level for a correct window classification is 50%. The training and testing split, followed by the classification process, was repeated five times for each participant with each modality and fusion approach. As a final result for each participant, we calculated the average accuracy for the five runs.

2.4. Fusion Approaches

We compared the single modality results to four different fusion approaches. For the early feature fusion, we implemented two different versions: (1) the image feature sets of the EEG and eye tracking data are concatenated and classified by a simple

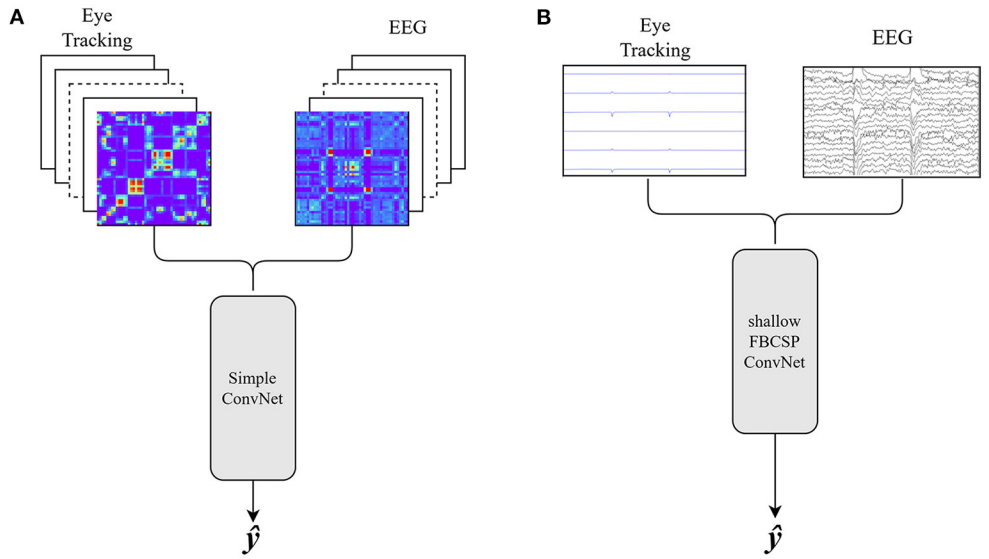


FIGURE 2 | (A) Early fusion approach using image features for EEG and eye tracking data. The stacked images are classified by a Simple Convolutional Neural Network. **(B)** Early fusion approach using time series of EEG and eye tracking data. The combined input is classified by the shallow FBCSP Convolutional Neural Network. Both: $\hat{y}$ is the class prediction.

CNN, and (2) the time series feature sets of both modalities are combined and classified using the shallow FBCSP CNN (see **Figure 2**). All parameters and training strategies were identical to the single modality classification process described in section 2.3.

In the middle fusion approach, the time-series features of the EEG data and the image features of the eye tracking data were used. As described in **Figure 3**, both feature sets were first processed simultaneously by different neural networks. A reduced version of the shallow FBCSP CNN got trained on the EEG data. The reduced model is identical to the model described in **Table 1** but the output size of the last layer (conv_classifier) was increased to 40. The eye tracking data were used to train the first layers of a simple CNN, until after the first linear layer (fc1; see **Table 2**). At this point, the outputs of both networks got concatenated, changing the input size of the second fully connected layer (fc2) before passing through the rest of the linear layers of a simple CNN.

Lastly, in the late fusion approach, the EEG and eye tracking data were classified separately as described for the single modality approaches. The prediction probabilities of both classes were used to decide on the final prediction (see **Figure 4**). We used the following decision rule: if both modalities predict the same label, use it as the final prediction. Else, if the probability of the EEG prediction $P(\hat{y}) > 0.5$, use the label predicted by the EEG classifier. Else, use the label that was predicted by the eye tracking classifier.

The decision was mutual (case 1) in 0.572 ± 0.074 of the trials. For 0.368 ± 0.071 of the trials, the EEG prediction was passed on and for 0.06 ± 0.024 the eye tracking decision was used.

3. RESULTS

All reported results are the statistics computed across all participants. We will first report the mean, standard deviation, range, and 95%-confidence interval of each approach, before testing for significant differences. All results can be seen in **Figure 5**.

The EEG-based single unimodal classification reached an average accuracy of 0.635 ± 0.095. The results ranged from 0.450 to 0.859, and the 95%-confidence interval of the classification accuracy for a new subject is [0.603, 0.668].

For the eye tracking approach, the average accuracy was 0.582 $\pm$ 0.092 within the range [0.397, 0.870]. The 95%-confidence interval was [0.551, 0.614].

When both modalities were represented by their time-series and processed with the shallow FBCSP CNN (Early Fusion—TS), the mean accuracy was 0.572 ± 0.077 (range [0.386, 0.853]). With a 95% confidence, the classification accuracies for this approach will reach between 0.545 and 0.598. The early fusion approach using image features (Early Fusion—Images) reached an average accuracy of 0.608 $\pm$ 0.083 over all participants. The range for this approach was [0.422, 0.887] and the 95%-confidence interval [0.580, 0.636].

For the middle fusion, the mean accuracy was 0.617 $\pm$ 0.101, range of [0.431, 0.870], and 95%-confidence interval of [0.583, 0.652].

Finally, the late fusion approach with the decision rule described in section 2.4 achieved the highest mean classification accuracy with 0.642 $\pm$ 0.096, a range of [0.456, 0.881] and a confidence interval between 0.609 and 0.675.

We performed the significance analysis using a paired two-tailed t-test of the accuracy on all combinations of approaches (see **Table 3**). Our main aim in this study was to identify promising approaches for the feature combination of a multimodal classifier. These results hint at which approach is worth improving, adjusting, and optimizing further. Thus, we would prefer a False Positive over a False Negative because it would make us "exclude" a promising approach for further

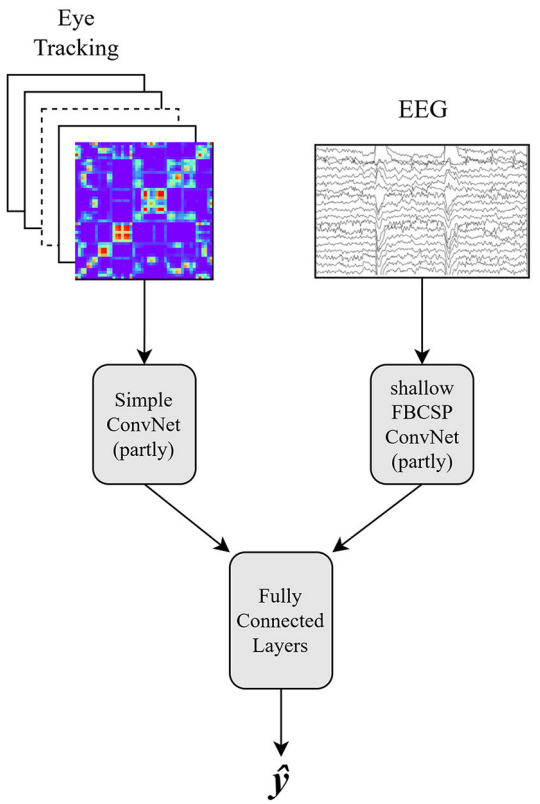


FIGURE 3 | Middle fusion approach using image features for eye tracking data in combination with the first steps of a Simple Convolutional Neural Network and the time series of EEG data with the reduced shallow FBCSP Convolutional Neural Network. The concatenated outputs are put through further linear layers. $\hat{y}$ is the class prediction.

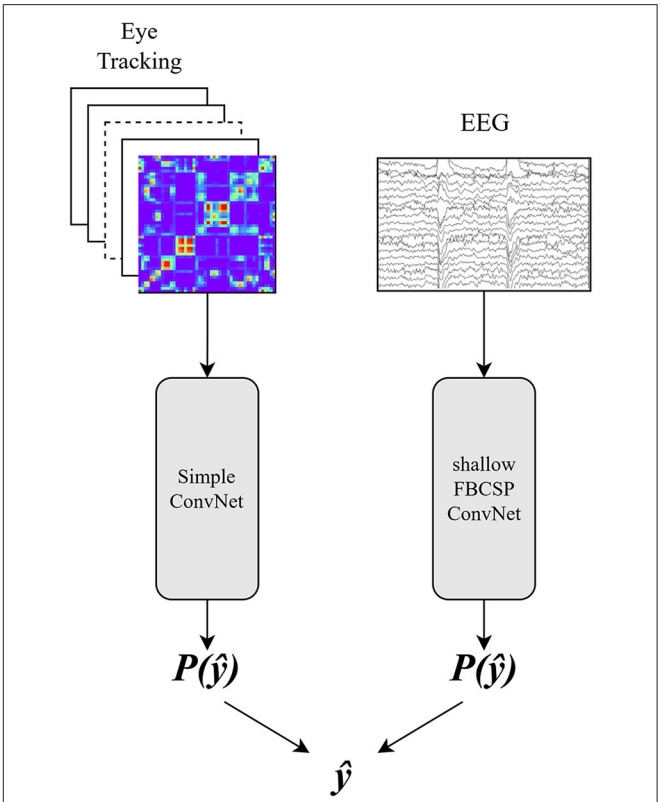


FIGURE 4 | Late fusion approach using image features for eye tracking data in combination with a Simple Convolutional Neural Network and the time series of EEG data with the shallow FBCSP Convolutional Neural Network. The probabilities of the predictions $P(\hat{y})$ are used for the final prediction $\hat{y}$.

studies on this topic. Following this philosophy, we chose a less conservative correction for multiple testing. By controlling the False Detection Rate (FDR) following Benjamini and Hochberg (1995), we find six significant differences. For the single modalities, the results for the EEG classification are not significantly better than the eye tracking results because they were identified as a false positive. Between the two early fusion approaches, the results obtained by the image feature set were significantly better than for the time-series features. No classification approach was significantly different from all other approaches, but the multimodal late fusion outperformed both unimodal classification approaches.

4. DISCUSSION

A system requires information in order to adapt more effectively to the needs of its users. The synergy may increase further, if a user does not have to explicitly state such requirements. Biosignals are a means of implicitly acquiring information, and combining multiple signals concurrently may result in a more accurate fit. Thus, we classified attention as internally or externally directed using 3 s multimodal EEG and eye tracking data in the current study. We compared different feature sets and feature fusion strategies. For the two feature sets and neural networks, we chose one combination that was previously used for EEG data (Schirrmeister et al., 2017) and one combination that was previously used for eye tracking data (Vortmann et al., 2021).

In a preliminary analysis of classification accuracies for the two single modalities, we discovered that prediction accuracies based on EEG data ($M = 0.635$) were significantly higher than those based on eye tracking data ($M = 0.582$). Regardless of the suitability of the modalities themselves, the disparities could also be explained by the disparate classification processes.

Interestingly, fusion of image features ($M = 0.608$) outperformed time series classification ($M = 0.572$) significantly for the two early fusion approaches. The image features were previously used for the eye tracking classification. As a result, we conclude that the different accuracies cannot be attributed solely to the quality of the classification approaches themselves. Rather than that, it appears as though the classification strategy and modality being used interact.

Neither of the early fusion approaches outperformed the single modalities by a significant margin. The time-series-based early fusion approach ($M = 0.572$) performed even worse than the unimodal EEG classification ($M = 0.635$). As discussed in the related work, other early fusion strategies have been used in the

TABLE 3 | *P*-values of two-tailed paired *t*-test for the comparison of the feature fusion approaches.

	EEG	ET	Early—TS	Early—images	Middle	Late
Only ET	**<0.001**					
Early fusion—TS	**= 0.005**	= 0.578				
Early fusion—images	= 0.141	= 0.068	**= 0.038**			
Middle fusion	= 0.42	= 0.091	**= 0.002**	= 0.648		
Late fusion	**=0.016**	**<0.001**	**= 0.003**	= 0.0693	= 0.268	
Average accuracy (%)	63.5	58.2	57.2	60.8	61.7	64.2

Significant differences are marked in bold. A significance threshold of $\alpha < 0.05$ is assumed. FDR correction by Benjamini and Hochberg (1995) was applied to correct for multiple testing. TS, time series.

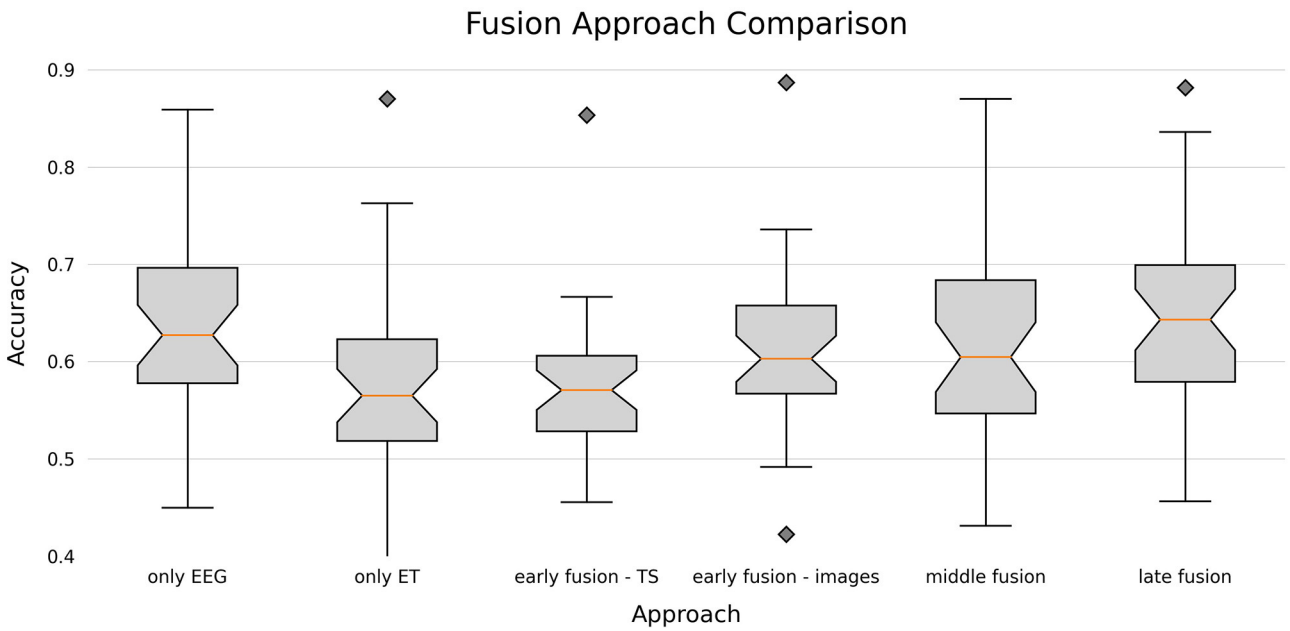


FIGURE 5 | Boxplots for the classification accuracies of the different fusion approaches visualizing the lower to upper quartile values of the data, with a line at the median. The whiskers represent the 95% range of the data. Diamonds represent outliers.

past to combine EEG and eye tracking data (Mangai et al., 2010; Liu et al., 2016; Guo et al., 2019). Different feature extraction algorithms or early statistics-based feature fusion techniques could be used in future studies to improve classification accuracy for the early fusion approaches. However, it was already noted in Polikar (2006) that early fusion is not reasonable as opposed to late fusion because of the diversity in the data. Thus, we see an advantage for middle and late fusion approaches.

As proposed in the section 1, a middle fusion could be an effective way to combine the advantages of feature-level and decision-level fusion. Individual modalities are processed independently first, resulting in classifier branches that are optimally adapted and trained for each modality. The two branches are connected in the middle, and the available data from both modalities can be used to train the rest of the network. While this approach enables correlations to be exploited, it also identifies significant unimodal data patterns that would be missed by other feature extraction approaches used in early fusion strategies. The primary difficulty with the middle fusion approach is network design. While it combines the strengths of the other two fusion strategies, it also incorporates their challenges. In a first step, suitable feature extraction and representation, as well as network structure for each modality, have to be found. These neural network branches must be designed in such a way that they allow for concatenation at a predetermined point. Finally, the neural network's subsequent layers must be appropriately designed for the merged modalities. On the one hand, complex correlations, and interactions must be discovered in order for the network to outperform a late fusion approach. On the other hand, the network's complexity must remain reasonable in comparison to the amount of data available. Otherwise, middle fusion networks will almost certainly have an excessively large number of parameters, rendering them unsuitable for a wide variety of applications.

It is difficult to generalize the results of the middle fusion, in particular: The neural network's structure is extremely adaptable, with an infinite number of possible configurations. The fully connected layers add parameters for successfully classifying multimodal data by learning correlations. The results of this study indicate that middle fusion is more promising than unimodal and early fusion approaches, but does not outperform late fusion. We assume that the network structure chosen was not optimal for maximizing the benefits of intermediate fusion. The layers were designed to resemble the individual unimodal networks and merged appropriately to maintain comparability. We hypothesize that more conservative and informed neural network engineering could significantly improve classification results. On the downside, this engineering is likely to be highly dataset and application dependent and will require a thorough understanding of the modalities' interactions.

In conclusion, our findings indicate that performing feature fusion in the middle of the classification process can slightly improve classification performance when compared to early fusion approaches. But supposedly, the neural network that intermediately combines the two modalities is subject to many adjustments and requires special engineering for each feature set combination and application.

While there was no significant difference between the middle fusion ($M = 0.617$) and the late fusion ($M = 0.642$), the late fusion approach was the only approach to significantly outperform both unimodal approaches in this data set. However, it did not outperform both early fusion approaches.

By comparison, the late fusion approach's optimization of the decision rule contains fewer parameters and is easily adaptable to new feature sets. However, the approach suggested here required expert knowledge to come up with a decision rule. For more efficient decision level fusion, statistical approaches or attention mechanisms could be applied (Mirian et al., 2011).

Improved unimodal classification pipelines would be a primary goal of improving late fusion. The primary disadvantage of the late fusion approach discussed in section 1 is the absence of correlation exploration between the modalities, which are processed independently. Thus, any information encoded in the early combination cannot be discovered using late fusion approaches that combine the modalities only at the decision level. A possible solution to this issue would be to add another "branch" of classification that predicts an output based on fused input, while maintaining the single modality classification. In our example, the decision rule would consider the EEG, ET, and a third combined prediction in addition to the two predicted labels and their probabilities.

We discovered during the training process that classification accuracy was highly dependent on the current training and test split for the same data set. Increasing the size of the data set may eliminate this effect. If more training data were available, the variance in the data would help to reduce bias and the likelihood of overfitting on the training data.

Another aspect that requires further thought is the inter-subject variability. The appropriate classification approach may depend on the participant and the quality of the data of each modality. For subjects with low individual EEG and eye tracking classification accuracies a middle or early fusion approach might increase the accuracy significantly. On the other hand, it the individual classification accuracies are already good, a late fusion might benefit from the modality specific classification.

We used a designated EEG and eye tracking co-registration study to have similar data quality for both modalities. The data was collected in a controlled laboratory environment. Applications and use cases with a more flexible setup and varying data qualities require another examination because one of the suggested approaches could be better suitable to correct for the worse quality of one modality than the others.

Overall, the differences between the approaches are not substantial enough to generally recommend the use of one over the others. We were able to show that a classification of more strongly internally vs. externally directed attention based on short data windows is possible above chance level for several approaches. We assume that the best fusion approach is highly dependent on the structure of the available multimodal data (e.g., sampling rate, data quality) and conclude that testing several approaches is necessary to find the most suitable for the data set. **Table 4** summarizes the advantages, challenges, and premises for each fusion approach.

TABLE 4 | Summary of the advantages, challenges, and premises for each fusion approach.

Fusion approach	Advantages (+), Challenges (−), and Premises (*)
Early fusion	+ Possibly finds correlations between modalities
	− Very different data structures to combine
	− Must use similar feature structures for all modalities
	* The same sampling rates for the data
	* Or preprocessing to adapt the data to each other
	* Best used when high chance of important modality interactions
Middle fusion	+ Tailored initial modality specific layers
	+ Possibly finds correlations between modalities
	+ Can work with different feature structures
	− Advanced NN engineering
	* Enough data for complex NN structure
	* Preliminary individual engineering of individual modalities was very different
	* Possibly important modality interactions
Late fusion	+ Tailored modality specific network design and features
	+ Missing data from one modality can be easily compensated
	− Finding a suitable decision rule or algorithm
	* Either good insight to find decision rule
	* Or enough data to train decision using ML
	* Best used when low chance of important modality interaction

4.1. Future Work

The current results may inspire further, more fine-grained comparisons even within the groups of early fusion approaches,

middle fusion networks, and late fusion decisions. On top of the presented suggestions on improving the current approaches, classification accuracy might increase if pre-trained models or transfer learning were applied For future work, other comparable data sets will be used to enlarge the data available for the training. The generalizability of the presented results should also be tested with further unrelated data sets. This study exclusively analyzed the data person-dependently. In the future, person-independence should be evaluated. The classification of unseen participants would include training the model on a pooled dataset of other participants, for example, in a leave-one-out approach. While the increased size of the training dataset might improve the accuracy of the classifier, the differences between participants might increase the variance in the dataset. Previous results have shown that the person-independent classification of EEG data is difficult and person-specific models are still the norm (Vortmann and Putze, 2021), whereas attempts to classify the eye tracking data of unseen participants for different attentional states were promising on larger datasets (Vortmann et al., 2021). However, the problem of generalizability was already discussed by Annerer-Walcher et al. (2021) who state that for internally and externally directed attention eye tracking data does not generalize well over participants. Our results have shown that a multimodal classifier outperforms unimodal classifiers for within-person training and testing and the next step will be to explore whether these improvements also hold for person-independent classification. For the real-time application of such a classifier in a BCI, the possibility to classify unseen participants without the need for person-dependent training data would highly increase the range of applications and the usability.

DATA AVAILABILITY STATEMENT

Materials and data are provided on the Open Science Framework (OSF, https://osf.io/5u6r9/).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Graz, Austria. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

L-MV performed the analysis and prepared the manuscript. SC and FP contributed to the final version of the manuscript. FP supervised the project. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Austrian Science Fund (FWF): P29801.

REFERENCES

Annerer-Walcher, S., Ceh, S. M., Putze, F., Kampen, M., Körner, C., and Benedek, M. (2021). How reliably do eye parameters indicate internal versus external attentional focus? *Cogn. Sci*. 45, e12977. doi: 10.1111/cogs.12977

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2018). Multimodal machine learning: a survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell*. 41, 423–443. doi: 10.1109/TPAMI.2018.2798607

Benedek, M., Bergner, S., Könen, T., Fink, A., and Neubauer, A. C. (2011). EEG alpha synchronization is related to top-down processing in convergent and divergent thinking. *Neuropsychologia* 49, 3505–3511. doi: 10.1016/j.neuropsychologia.2011.09.004

Benedek, M., Jauk, E., Beaty, R. E., Fink, A., Koschutnig, K., and Neubauer, A. C. (2016). Brain mechanisms associated with internally directed attention and self-generated thought. *Sci. Rep*. 6, 1–8. doi: 10.1038/srep22959

Benedek, M., Schickel, R. J., Jauk, E., Fink, A., and Neubauer, A. C. (2014). Alpha power increases in right parietal cortex reflects focused internal attention. *Neuropsychologia* 56, 393–400. doi: 10.1016/j.neuropsychologia.2014.02.010

Benedek, M., Stoiser, R., Walcher, S., and Körner, C. (2017). Eye behavior associated with internally versus externally directed cognition. *Front. Psychol*. 8:1092. doi: 10.3389/fpsyg.2017.01092

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Ceh, S. M., Annerer-Walcher, S., Körner, C., Rominger, C., Kober, S. E., Fink, A., et al. (2020). Neurophysiological indicators of internal attention: an electroencephalography-eye-tracking coregistration study. *Brain Behav*. 10, e01790. doi: 10.1002/brb3.1790

Ceh, S. M., Annerer-Walcher, S., Koschutnig, K., Körner, C., Fink, A., and Benedek, M. (2021). Neurophysiological indicators of internal attention: an fMRI-eye-tracking coregistration study. *Cortex* 143, 29–46. doi: 10.1016/j.cortex.2021.07.005

Cheng, S., Wang, J., Zhang, L., and Wei, Q. (2020). Motion imagery-BCI based on EEG and eye movement data fusion. *IEEE Trans. Neural Syst. Rehabil. Eng*. 28, 2783–2793. doi: 10.1109/TNSRE.2020.3048422

Chun, M. M., Golomb, J. D., and Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annu. Rev. Psychol*. 62, 73–101. doi: 10.1146/annurev.psych.093008.100427

Debie, E., Fernandez Rojas, R., Fidock, J., Barlow, M., Kasmarik, K., Anavatti, S., et al. (2021). Multimodal fusion for objective assessment of cognitive workload: a review. *IEEE Trans. Cybern*. 51, 1542–1555. doi: 10.1109/TCYB.2019.2939399

Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., and Kliegl, R. (2011). Coregistration of eye movements and eeg in natural reading: analyses and review. *J. Exp. Psychol*. 140, 552. doi: 10.1037/a0023885

D'Mello, S., and Kory, J. (2012). "Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction, ICMI '12* (New York, NY: Association for Computing Machinery), 31–38. doi: 10.1145/2388676.2388686

Dong, X., Wang, H., Chen, Z., and Shi, B. E. (2015). "Hybrid brain computer interface via Bayesian integration of EEG and eye gaze," in *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)* (Montpellier), 150–153. doi: 10.1109/NER.2015.7146582

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., et al. (2013). MEG and EEG data analysis with MNE-Python. *Front. Neurosci*. 7, 267. doi: 10.3389/fnins.2013.00267

Guo, J.-J., Zhou, R., Zhao, L.-M., and Lu, B.-L. (2019). "Multimodal emotion recognition from eye image, eye movement and EEG using deep neural networks," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Berlin), 3071–3074. doi: 10.1109/EMBC.2019.8856563

Kim, M., Kim, B. H., and Jo, S. (2015). Quantitative evaluation of a low-cost noninvasive hybrid interface based on EEG and eye movement. *IEEE Trans. Neural Syst. Rehabil. Eng.* 23, 159–168. doi: 10.1109/TNSRE.2014.2365834

Kim, M., Lee, S., and Kim, J. (2020). "Combining multiple implicit-explicit interactions for regression analysis," in *2020 IEEE International Conference on Big Data (Big Data)* (Atlanta, GA), 74–83. doi: 10.1109/BigData50022.2020.9378402

Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kulke, L. V., Atkinson, J., and Braddick, O. (2016). Neural differences between covert and overt attention studied using EEG with simultaneous remote eye tracking. *Front. Hum. Neurosci.* 10, 592. doi: 10.3389/fnhum.2016.00592

Liu, W., Zheng, W.-L., and Lu, B.-L. (2016). "Emotion recognition using multimodal deep learning," in *International Conference on Neural Information Processing* (Kyoto: Springer), 521–529. doi: 10.1007/978-3-319-46672-9_58

Lobo, J. L., Ser, J. D., De Simone, F., Presta, R., Collina, S., and Moravek, Z. (2016). "Cognitive workload classification using eye-tracking and EEG data," in *Proceedings of the International Conference on Human-Computer Interaction in Aerospace, HCI-Aero '16* (New York, NY: Association for Computing Machinery). doi: 10.1145/2950112.2964585

López-Gil, J.-M., Virgili-Gomá, J., Gil, R., Guilera, T., Batalla, I., Soler-González, J., et al. (2016). Method for improving EEG based emotion recognition by combining it with synchronized biometric and eye tracking technologies in a non-invasive and low cost way. *Front. Comput. Neurosci.* 10, 85. doi: 10.3389/fncom.2016.00119

Loshchilov, I., and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, Y., Zheng, W.-L., Li, B., and Lu, B.-L. (2015). "Combining eye movements and EEG to enhance emotion recognition," in *IJCAI, Vol. 15* (Buenos Aires), 1170–1176.

Mangai, U. G., Samanta, S., Das, S., and Chowdhury, P. R. (2010). A survey of decision fusion and feature fusion strategies for pattern classification. *IETE Techn. Rev.* 27, 293–307. doi: 10.4103/0256-4602.64604

Mirian, M. S., Ahmadabadi, M. N., Araabi, B. N., and Siegwart, R. R. (2011). Learning active fusion of multiple experts' decisions: an attention-based approach. *Neural Comput.* 23, 558–591. doi: 10.1162/NECO_a_00079

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "Pytorch: an imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems, Vol. 32*, eds H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Vancouver: Curran Associates, Inc.), 8024–8035.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits Syst. Mag.* 6, 21–45. doi: 10.1109/MCAS.2006.1688199

Polikar, R. (2012). *Ensemble Learning*. Boston, MA: Springer US, 1–34. doi: 10.1007/978-1-4419-9326-7_1

Putze, F., Scherer, M., and Schultz, T. (2016). "Starring into the void? Classifying internal vs. external attention from EEG," in *Proceedings of the 9th Nordic Conference on Human-Computer Interaction* (Gothenburg), 1–4. doi: 10.1145/2971485.2971555

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., et al. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.* 38, 5391–5420. doi: 10.1002/hbm.23730

Smallwood, J., and Schooler, J. W. (2006). The restless mind. *Psychol. Bull.* 132, 946. doi: 10.1037/0033-2909.132.6.946

Spreng, R. N., Stevens, W. D., Chamberlain, J. P., Gilmore, A. W., and Schacter, D. L. (2010). Default network activity, coupled with the frontoparietal control network, supports goal-directed cognition. *NeuroImage* 53, 303–317. doi: 10.1016/j.neuroimage.2010.06.016

Vortmann, L.-M., Knychalla, J., Walcher, S., Benedek, M., and Putze, F. (2021). Imaging time series of eye tracking data to classify attentional states. *Front. Neurosci.* 15, 625. doi: 10.3389/fnins.2021.664490

Vortmann, L.-M., Kroll, F., and Putze, F. (2019a). EEG-based classification of internally-and externally-directed attention in an augmented reality paradigm. *Front. Hum. Neurosci.* 13, 348. doi: 10.3389/fnhum.2019.00348

Vortmann, L.-M., and Putze, F. (2020). "Attention-aware brain computer interface to avoid distractions in augmented reality," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu), 1–8. doi: 10.1145/3334480.3382889

Vortmann, L.-M., and Putze, F. (2021). Exploration of person-independent bcis for internal and external attention-detection in augmented reality. *Proc. ACM Interact. Mobile Wear. Ubiquit. Technol.* 5, 1–27. doi: 10.1145/3463507

Vortmann, L.-M., Schult, M., Benedek, M., Walcher, S., and Putze, F. (2019b). "Real-time multimodal classification of internal and external attention," in *Adjunct of the 2019 International Conference on Multimodal Interaction* (Suzhou), 1–7. doi: 10.1145/3351529.3360658

Wang, Z., and Oates, T. (2015). "Imaging time-series to improve classification and imputation," in *Twenty-Fourth International Joint Conference on Artificial Intelligence* (Buenos Aires: AAAI Press), 3939–3945. doi: 10.5555/2832747.2832798

Wu, Q., Dey, N., Shi, F., Crespo, R. G., and Sherratt, R. S. (2021). Emotion classification on eye-tracking and electroencephalograph fused signals employing deep gradient neural networks. *Appl. Soft Comput.* 110, 107752. doi: 10.1016/j.asoc.2021.107752

Zhang, W., Ji, X., Yang, Y., Chen, J., Gao, Z., and Qiu, X. (2018). "Data fusion method based on improved ds evidence theory," in *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)* (Shanghai), 760–766. doi: 10.1109/BigComp.2018.00145

Zheng, W.-L., Dong, B.-N., and Lu, B.-L. (2014). "Multimodal emotion recognition using EEG and eye tracking data," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Chicago, IL), 5040–5043.

Zheng, W.-L., Liu, W., Lu, Y., Lu, B.-L., and Cichocki, A. (2019). Emotionmeter: a multimodal framework for recognizing human emotions. *IEEE Trans. Cybern.* 49, 1110–1122. doi: 10.1109/TCYB.2018.2797176

Zhu, J., Wang, Z., Gong, T., Zeng, S., Li, X., Hu, B., et al. (2020). An improved classification model for depression detection using EEG and eye tracking data. *IEEE Trans. NanoBiosci.* 19, 527–537. doi: 10.1109/TNB.2020.2990690

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Psychology

ORIGINAL RESEARCH
published: 25 April 2022
doi: 10.3389/fpsyg.2022.882446

Prediction of Disorientation by Accelerometric and Gait Features in Young and Older Adults Navigating in a Virtually Enriched Environment

Stefan J. Teipel[1,2], Chimezie O. Amaefule[1], Stefan Lüdtke[3,4], Doreen Görß[2], Sofia Faraza[2], Sven Bruhn[5] and Thomas Kirste[3]*

[1] Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE) Rostock/Greifswald, Rostock, Germany, [2] Department of Psychosomatic Medicine, University Medicine Rostock, Rostock, Germany, [3] Mobile Multimedia Information Systems, Institute for Visual and Analytic Computing, University of Rostock, Rostock, Germany, [4] Institute for Enterprise Systems, University of Mannheim, Mannheim, Germany, [5] Institute for Sports Science, University of Rostock, Rostock, Germany

OPEN ACCESS

Edited by:
Siyuan Chen,
University of New South Wales, Australia

Reviewed by:
Antonella Lopez,
University of Bari Aldo Moro, Italy
Alessandro Oronzo Caffò,
University of Bari Aldo Moro, Italy

***Correspondence:**
Stefan J. Teipel
stefan.teipel@med.uni-rostock.de

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Psychology

Received: *23 February 2022*
Accepted: *22 March 2022*
Published: *25 April 2022*

Citation:
Teipel SJ, Amaefule CO, Lüdtke S, Görß D, Faraza S, Bruhn S and Kirste T (2022) Prediction of Disorientation by Accelerometric and Gait Features in Young and Older Adults Navigating in a Virtually Enriched Environment. Front. Psychol. 13:882446. doi: 10.3389/fpsyg.2022.882446

Objective: To determine whether gait and accelerometric features can predict disorientation events in young and older adults.

Methods: Cognitively healthy younger (18–40 years, n = 25) and older (60–85 years, n = 28) participants navigated on a treadmill through a virtual representation of the city of Rostock featured within the Gait Real-Time Analysis Interactive Lab (GRAIL) system. We conducted Bayesian Poisson regression to determine the association of navigation performance with domain-specific cognitive functions. We determined associations of gait and accelerometric features with disorientation events in real-time data using Bayesian generalized mixed effect models. The accuracy of gait and accelerometric features to predict disorientation events was determined using cross-validated support vector machines (SVM) and Hidden Markov models (HMM).

Results: Bayesian analysis revealed strong evidence for the effect of gait and accelerometric features on disorientation. The evidence supported a relationship between executive functions but not visuospatial abilities and perspective taking with navigation performance. Despite these effects, the cross-validated percentage of correctly assigned instances of disorientation was only 72% in the SVM and 63% in the HMM analysis using gait and accelerometric features as predictors.

Conclusion: Disorientation is reflected in spatiotemporal gait features and the accelerometric signal as a potentially more easily accessible surrogate for gait features. At the same time, such measurements probably need to be enriched with other parameters to be sufficiently accurate for individual prediction of disorientation events.

Keywords: navigation, virtual reality, aging, visuo-spatial abilities, executive function, gait, actimetry

INTRODUCTION

Aging is associated with a decline in walking ability (Baudendistel et al., 2021) and cognitive performance (Iachini et al., 2009). These changes become particularly evident in dual-task conditions. For example, older people have difficulties walking and navigating in a new environment (Lithfous et al., 2013; Lester et al., 2017), resulting in reduced wayfinding abilities. These changes are even more pronounced during the transition from healthy aging to cognitive impairment and dementia (Gazova et al., 2012; Cohen and Verghese, 2019; Costa et al., 2020). They represent a high burden on older people and lead to fear of getting lost, social withdrawal, and a subsequent decrease in physical mobility (Panel on Prevention of Falls in Older Persons, American Geriatrics Society and British Geriatrics Society, 2011).

At the same time, wayfinding problems are amenable to technical assistance. Navigation systems are already part of our everyday environment; they support drivers and pedestrians, for example. For older people and people with cognitive impairments, in particular, it is important that assistance systems do not replace remaining cognitive abilities, but rather make use of them. Previous work has shown that habitual use of navigation aids may decrease spatial memory performance even in cognitively healthy people (Dahmani and Bohbot, 2020). Current technology development is therefore aimed at situation-aware navigation assistance that supports the user only when necessary (Teipel et al., 2016). Such systems require accurate detection of navigation behavior, especially real-time detection of episodes of disorientation before the user is lost (Yordanova et al., 2017).

Previous studies used experiments in virtual reality (VR) environments to assess spatial orientation (Zakzanis et al., 2009; Kizony et al., 2017; Tascon et al., 2018; Costa et al., 2020; Paliokas et al., 2020). VR approaches are highly controlled but lack the dual-task characteristic of combining spatial navigation with walking. One previous study found that navigational performance results were comparable between a VR and a real-world navigational test in young and older cognitively normal adults and people with dementia (Cushman et al., 2008), but VR testing alone obviously does not allow assessment of gait and motion features during spatial navigation. On the other hand, several studies used wearable sensors to assess the gait and movement characteristics of cognitively normal older people and people with dementia in real-life situations (Becu et al., 2020; Mc Ardle et al., 2021; Pawlaczyk et al., 2021; Weizman et al., 2021). Some of these real-world studies were primarily aimed at exploring different components of spatial orientation in normal human behavior and the underlying neural basis but did not aim to map the full range of navigational behavior in everyday situations (Wei et al., 2020). Other studies mainly focused on the early detection of dementia symptoms using gait characteristics in real-world environments (Mc Ardle et al., 2021; Mulas et al., 2021; Weizman et al., 2021) or under dual task conditions (Oh, 2021).

In a previous study, we had assessed whether accelerometric features from wearable sensor devices were useful to identify episodes of disorientation even before an individual has deviated from the intended route (Schaat et al., 2019). We found that accelerometry-detected episodes of disorientation with an area under the receiver operating characteristics (ROC) curve (AUC) of 75% and 79% correctly allocated disorientation episodes in people with mild cognitive impairment (MCI) or dementia moving through an urban environment (Schaat et al., 2019). This level of accuracy suggested that there were relevant features in the accelerometric signal to detect disorientation at the group level. At the same time, the accuracy was not high enough for individual situation detection. In addition, people with dementia or MCI experienced a relatively small number of disorientation episodes, which limited the training of an accurate model based on positive events (Schaat et al., 2019).

Here, we transferred our previous approach to the better-controlled environment within the Gait Real-Time Analysis Interactive Lab (GRAIL) system. The GRAIL consists of a physical treadmill combined with a large hemisphere screen (Amaefule et al., 2020). In our experiment, the GRAIL screen featured a virtual representation of the city center of Rostock, resembling the environment of the previous real-world experiment (Schaat et al., 2019). Participants were asked to navigate through this environment while walking on the treadmill. In a previous pilot study, we showed that this set-up was feasible for use with older participants, including people with cognitive decline, and allowed us to record a comprehensive set of predictive features, including accelerometry, gait features, and physiological signals (Amaefule et al., 2020). In addition, we were able to induce disorientation episodes by removing landmarks from the virtual environment to provide more instances for model training. The key role of landmarks for spatial orientation in virtual environments has previously been shown (Caffo et al., 2018). In this study, we presented the results of this approach in young and older adults without manifest cognitive impairment. As a primary aim, we wanted to determine whether a combination of accelerometry and gait characteristics was accurate enough to immediately detect episodes of disorientation. We hypothesized that the accerelometric and gait features may yield sufficient accuracy for individual detection of disorientation episodes in real time. Especially, we expected a level of accuracy above 80% for the binary outcome of oriented *vs.* disoriented. As a secondary aim, we determined whether the number of disorientation events per participant was associated with cognitive scores and aggregated accelerometric and gait characteristics. The results of this study will be relevant to the design of experiments with individuals with manifest cognitive decline and also to the design of future real-world experiments targeting situation-aware navigation aids.

MATERIALS AND METHODS

Subjects

For the ongoing GRAIL study, we recruited three groups of participants: mobile, physically and cognitively healthy younger (18–40 years) and older (60–85 years) participants, and physically healthy persons with diagnosed MCI or mild dementia due to AD

(Age: 60–85 years, MMSE: 15–27) according to NIA-AA criteria (Albert et al., 2011; McKhann et al., 2011).

Patients and healthy older adults were recruited from the memory clinic of the Rostock University Medical Center, while the healthy young adults were recruited from within the University of Rostock student community. Exclusion criteria for all groups were other neurological conditions besides MCI or dementia in the patient group, inability to understand task instructions and questionnaire items, deaf-muteness, and blindness.

Due to the COVID pandemic restrictions, recruitment of patients with MCI and dementia was not possible for a longer time interval so only four patients had been recruited during the planned run-time of the project. Therefore, for the current analysis, we used only the data of a subset of 28 older and 25 young cognitively healthy participants that had complete data sets and behavioral annotation.

This study has been reviewed and approved by the Ethics committee of the Rostock University Medical Center (Approval No. A 2019-0062).

Experimental Set-Up

The experimental set-up has been described before (Amaefule et al., 2020). In brief, the participants were guided along a path in the virtual environment. Afterward, they were set back to the starting point and asked to walk the same path again, this time unguided. Navigation was possible by walking more to the left or right on the treadmill; this rotated the participant's position in the virtual environment to the left or right. The navigation route consisted of 14 major decision points (DP) which were primarily locations at which the participant had to decide to either continue in a particular direction, make a turn, or identify the goal position. For half of the healthy young or older subjects (the experimental group), phases of disorientation were induced by changing landmarks or decision points in the VR environment. These changes included (a) moving a landmark from one intersection to the next intersection, (b) adding a decision point, that is, an intersection, (c) blocking a road, and (d) moving the goal indicator to a different location. Overall, five locations were manipulated in the experimental group as follows: DP4 – a red pillar was moved from DP7 to DP4; DP9 – the road was blocked; DP11 – a new path was introduced; DP13 – the color of the pillar was changed to red; DP14 – the goal location was moved a little further away to DP14a. No changes to the environment were conducted in the control group.

Before the experiment, the participants were familiarized with the depicted city center by briefly showing them a map, such that problems in wayfinding would be due to disorientation instead of exploration in an unknown environment. We recorded spatiotemporal and kinematic gait parameters through the GRAIL system. In addition, we recorded accelerometric signals from three wearable sensors on the left wrist, right ankle, and chest, respectively, that each contained a three-axes accelerometer and three-axes gyroscope sampled with 64 Hz. Additionally, the chest sensor recorded an electrocardiogram (ECG, 1,024 Hz), and the wrist sensor recorded electrodermal activity (EDA, 32 Hz).

The experiments were video-recorded for subsequent offline annotation of behavior.

Randomization of the young and older participants into the experimental or control group was carried out using the program Research Randomizer, accessible at https://www.randomizer.org.

Behavior Annotation

An offline annotation procedure was applied to the video data recorded during the orientation task, for assessing the observable orientation behavior of the participants using the ELAN 5.8 tool (Wittenburg et al., 2006). As a coding scheme, we used an adequate adaption of the coding scheme provided by Yordanova et al. (2017). The same scheme had been used in one field study before (Schaat et al., 2019). This coding scheme also covers aspects of orientation behavior, which were beyond the scope of wayfinding in our VR set-up (e.g., behaviors associated with attention to traffic). For this reason, we adapted the coding scheme to capture only those behaviors that are obtainable within our virtual reality set-up.

Specifically, to identify instances of disorientation, we annotated when participants showed wandering behavior (i.e., non–goal-directed walk), communication behavior (i.e., asking for help when disoriented), topological orientation (i.e., trying to orient themselves based on the surrounding environment), or spatial orientation (i.e., trying to orient themselves based on landmarks). In addition, different types of errors that are associated with disoriented behavior were annotated (i.e., initiation, realization, sequence, and completion errors). The annotations were being evaluated based on the level of agreement between two annotators independently rating the data of five individuals, resulting in a Cohen's kappa of 0.87.

For the current analysis, the different types of disorientation behaviors were collapsed into a single feature of disorientation to provide a binary outcome of oriented *vs.* non-oriented state at a given time interval.

Neuropsychological Assessment

Neuropsychological assessment was only conducted on the older participants and the MCI or dementia patients. The assessment included the CERAD neuropsychological battery (Morris et al., 1989), the Rey–Osterrieth Complex Figure Test (Rey, 1941; Osterrieth, 1944), and the Perspective Taking/Spatial Orientation Test (PTSOT) (Hegarty and Waller, 2004). Cognitive domain composite scores assessing visual memory, executive functions, visuospatial constructional ability, and spatial orientation were computed by transforming raw scores of single tests to z-scores (Coley et al., 2016; Voss et al., 2018). Each of these domain scores were calculated as the mean score of specific tests, after transformation to z-scores. The visual memory composite included the delayed figural recall scores from the CERAD and the Rey Complex Figure Test after 3 min; the visuospatial composite included the direct figure copy scores from CERAD and the Rey Complex Figure Test. For executive function, we used the ratio of Trail Making Test B to A, and for the domain

of spatial orientation, we included cognitive scores from the Perspective Taking/Spatial Orientation Test.

Predictors

We included spatiotemporal and kinematic gait parameters from the GRAIL system, as well as the mean accelerometric signal from ankle, chest, and wrist-worn sensors and variability of these measures. To reduce the dimensionality of the models for the association analysis, we selected *a priori* features of interest. These included the ankle, wrist, and chest-worn mean accelerometric signal as well as the mean values of the spatiotemporal gait characteristics of walking speed, step length, stride time, step width, stance time, and swing time (Beauchet et al., 2017). The explorative multivariate models for real-time detection were allowed to select across all spatiotemporal and kinematic gait features (Lohman et al., 2011) first and second moments (mean and variance), the accelerometric signal means and variances at the time point of behavior assessment as well as the time-lagged features one, two, or three time intervals before the rated behavior (lagged features). **Supplementary Table 1** provides an overview of the feature sets defined for the different analyses.

Gait and Accelerometric Data Preprocessing

Accelerometric data, gait parameters, as well as video annotations were synchronized by an event-based mechanism (participants performed a distinctive movement at the beginning of the recording, which could be easily located in all sensors). The data were resampled at 100 Hz using cubic spline interpolation. We then aggregated the data in non-overlapping segments of length 10 s. Specifically, for the accelerometric data, we computed the mean, variance, skewness, and kurtosis of the magnitude of each of the three sensor positions, resulting in 12 features per segment. For the spatio-temporal gait parameters (walking speed, step length, stride time, stance time, swing time, and step width), the mean and coefficient of variation (CV) were computed for each segment. The CV was calculated for each gait parameter as the ratio of the standard deviation to the mean multiplied by 100.

We assigned a binary disorientation label to each 10-s segment based on the video annotation using the following rule: Whenever a navigation error or disoriented behavior was noted at any time during the segment, the segment was labeled as "disoriented." Conversely, if neither a navigation error nor a disoriented behavior was noted during the segment, the segment was labeled "not disoriented."

Statistical Analysis

Unless otherwise noted, all statistical analyses were performed using R statistical software, version 4.1.2, accessed *via* R Studio version 2021.09. Analyses were conducted in a Bayesian framework to allow estimation of model plausibility and determining effect sizes with credibility intervals. Demographic characteristics were compared between experimental groups using the Bayesian *t*-test or the Chi-square test as appropriate using Jeffreys's Amazing Statistics Program (JASP) 0.16 with default priors.

Subsequently, we conducted two groups of analyses:

The *first group of analyses* (A1) used the disorientation data aggregated across the entire observation period per participant. We selected two readouts for disorientation: the number of disorientation per subject during the navigation experiment (henceforth called disorientation counts) and the percentage of the length of the disorientation episodes relative to the overall length of the experiment per subject (henceforth called disorientation percentage).

First, we determined the regression of aggregated disorientation data on cognitive scores (only in old people) and aggregated accelerometric and gait features (in young and old people). We used generalized linear models with disorientation counts and percentage, respectively, as dependent variables, and cognitive scores and aggregated accelerometric and gait features as independent variables, respectively, controlling for age, gender, and experimental condition. The dependent variable (count data) was not normally distributed, therefore we fitted a Poisson regression model using the R library "brms." We compared the fit of the Poisson with the Gaussian regression model using leave-one-out cross-validation for Bayesian models with the R library "loo."

The *second group of analyses* (A2) used the binary variable of oriented (0) *vs*. disoriented (1) during each of the 10-s intervals as the dependent variable in all individuals. To enrich for disorientation episodes, we only considered time intervals during decision points (see **Supplementary Table 2** for the proportion of disorientation events per decision point).

First (A2.1), we used the Bayesian mixed-effects logistic regression models with accelerometric or gait features at each of the 10-s intervals as independent variables, controlling for age, gender, and experimental condition as fixed effects covariates, and with a random intercept for patients as random effect variable (observations nested within patients). These models were calculated using the R library "brms.

Second (A2.2), we determined, whether single accelerometric or gait features that had shown an effect in the previous analysis had a relevant predictive accuracy for episodes of disorientation. We used the area under the ROC curve to estimate a single feature's ability to predict disorientation at a time interval. ROC analysis was done using the library "ROCnReg" in R allowing for Bayesian estimates of credibility intervals for the areas under the ROC curves.

Third (A2.3), we used a multivariate approach to find a combination of accelerometric or gait features that may contribute to relevant accuracy in the detection of disorientation episodes. In this study, we used as the primary model a support vector machine (SVM), implemented using the R library "e1071." Before SVM training, we used feature selection based on the correlation coefficient of every single predictive feature with the dependent binary variable "oriented" *vs*. "disoriented." Only features with an absolute value of the correlation coefficient larger than 0.12 were entered into the SVM training. After visual inspection of the data revealed no linear separation between groups, we decided to use a radial kernel whose parameters cost and gamma function were determined using a 10-fold cross-validation using the function tune in library "e1071." To account

for the binding of the data within patients, we determined the accuracy of the SVM models using patients as folds. Within each patient, 80% of each patient's data were used as training data and the remaining 20% as test data. Accuracy was determined as the percentage of correctly classified time intervals where the predicted states of orientation or disorientation agreed with the observed states of orientation or disorientation relative to all observations per patient.

Finally (A2.4), we used a Gaussian Hidden Markov Model (HMM) respecting the temporal nature of the data. Using the HMM approach, we generated states (constraining the model to two possible states only) from the observed response variables, and subsequently compared the distribution of the generated states with the distribution of the observed states. This analysis was conducted using library "depmixS4" in R. Taking into account the origin of the time-series data, we split the analysis according to participants. We estimated transition matrices and means and standard deviations of the response variables from the data and used these estimates to fit the states' model per participant.

RESULTS

Demographic characteristics of our sample can be found in **Table 1**. Bayes factor analysis suggested no evidence in favor of a difference in sex distribution and education years across the groups and was in favor of no difference in age between experimental and control conditions within the young and older groups, respectively. By design, young and older groups differed in age. Participants in the experimental condition were presented with altered landmarks to induce disorientation, whereas participants in the control condition were not.

Aggregated Data

The average number of disorientation events, mean ankle-worn accelerometric signal, and walking speed per age group and the experimental condition is plotted in **Figure 1**. We found extreme evidence in favor of a difference between older control and experimental cases and between older experimental and young control cases, and moderate evidence in favor of a difference between older experimental and young experimental cases. Evidence for differences within the young age group and between the older control and the young control groups was not conclusive. For ankle-worn accelerometry and walking speed, there was mainly an age effect and a less-pronounced effect of experimental condition (see **Table 1** for details).

Leave-one-out-cross-validation of the *Watanabe-Akaike information criterion* (WAIC) (Vehtari et al., 2017) confirmed that the Poisson regression was superior to the Gaussian regression model fit [WAIC difference in favor of Poisson = –21.6 (SE = 8.9)] when using condition, age, and gender as the only predictors for the base model.

The number of disorientation events across the experiment were associated with **executive function** (smaller number of disorientation counts with higher executive function), but not with visuospatial constructional ability, visual memory, or perspective-taking/spatial orientation. **Ankle-worn sensor overall level of activity** was associated with counts of disorientation (more activity, less disorientation), but not wrist or chest-worn sensors.

When considering gait features, slower **walking speed** and lower **step length** were associated with a higher number of disorientation events.

Across all models, **experimental condition** and higher **age** were associated with a higher number of disorientation events, whereas gender was unrelated to disorientation events.

Detailed results can be found in **Table 2**. When repeating these analyses with the percentage of disorientation events per patient's time of experiment as an outcome, the results were essentially unchanged (data not shown). The only difference was that in addition to the previous effects, a higher wrist-worn accelerometric signal was associated with a higher percentage of disorientation events (main effect = 5.50, 95% credibility interval 2.25–8.67) as well.

Real-Time Data

For **accelerometric features**, we found the main effect of lower ankle, wrist, and chest-worn sensors' levels of activity with more disorientation events. In addition, we found interactions of ankle- and wrist-worn sensors' levels of activity with the experimental condition, showing more pronounced negative associations in the control than in the experimental condition (see **Figure 2** for an example of ankle-worn sensor activity). In addition, experimental condition, but not age or gender, was associated with more disorientation events. See **Table 2** for details.

For **gait features**, all *a priori* selected gait features showed a main effect on disorientation events. Lower walking speed and step length and width as well as longer stride, swing, and stance times were associated with more disorientation events. In addition, we found interactions of walking speed, step length, step width, and swing time with an experimental condition, showing more pronounced negative associations in the control than the experimental condition for walking speed, step length, and step width, and a more positive association for swing time (see **Figure 3** for an example of walking speed). In addition, experimental condition, but not age or gender, was associated with more disorientation events. See **Table 3** for details.

Accuracy of Disorientation Event Detection

We used the Bayesian ROC curve analysis to estimate the accuracy of single markers that had shown an association with orientation in the previous mixed-effect models. For ankle-worn accelerometric signal, the area under the ROC curve was 0.60 (95% credibility interval 0.588–0.615). For the remaining accelerometric features and the gait features, AUC values were below 0.60. These number indicate a detectable, but clinically irrelevant effect of single markers on accuracy levels.

Subsequently, we implemented a multivariate cross-validated support vector machine to determine the accuracy of a (non-linear combination of markers). Feature selection was done using absolute correlation coefficients > 0.12 between candidate

TABLE 1 | Demographic, orientation, and gait characteristics.

	Young controls	Young experimental	Older controls	Older experimental
N (f/m)[1]	4/6	8/7	9/5	9/5
Age[2] (years) (SD)	24.2 (2.7)	24.7 (4.3)	69.5 (4.0)	72.0 (5.3)
Education[3] (years) (SD)	13.3 (0.9)	14.0 (1.5)	13.9 (2.9)	15.0 (2.5)
Mean number disorientation[4] (SD)	0.30 (0.95)	2.20 (2.57)	0.57 (1.09)	5.79 (3.22)
Mean accelerometry[5] ankle (SD)	1.49 (0.11)	1.37 (0.01)	1.34 (0.10)	1.28 (0.05)
Mean walking speed[6] (SD)	1.44 (0.16)	1.23 (0.17)	1.06 (0.21)	0.93 (0.13)

[1]*Bayes factor in favor of no difference between groups, BF_{10} = 0.157.*
[2]*Bayes factor in favor of no difference, BF_{10} = 0.736, between older experimental and control cases, and in favor of no difference, BF_{10} = 0.390, between young experimental and control cases.*
[3]*Bayes factor in favor of no difference between groups, BF_{10} = 0.341.*
[4]*Bayes factor in favor of a difference between older controls and older experimentals, older experimentals and both young controls and young experimentals (BF_{10} > 14.7).*
[5]*Bayes factor in favor of a difference between older controls and young experimentals, older experimentals and young controls, and young experimentals and young controls (BF_{10} > 9.0).*
[6]*Bayes factor in favor of a difference between older controls and young controls, older experimentals and young experimentals and young controls, and young experimentals and young controls (BF_{10} > 9.0).*

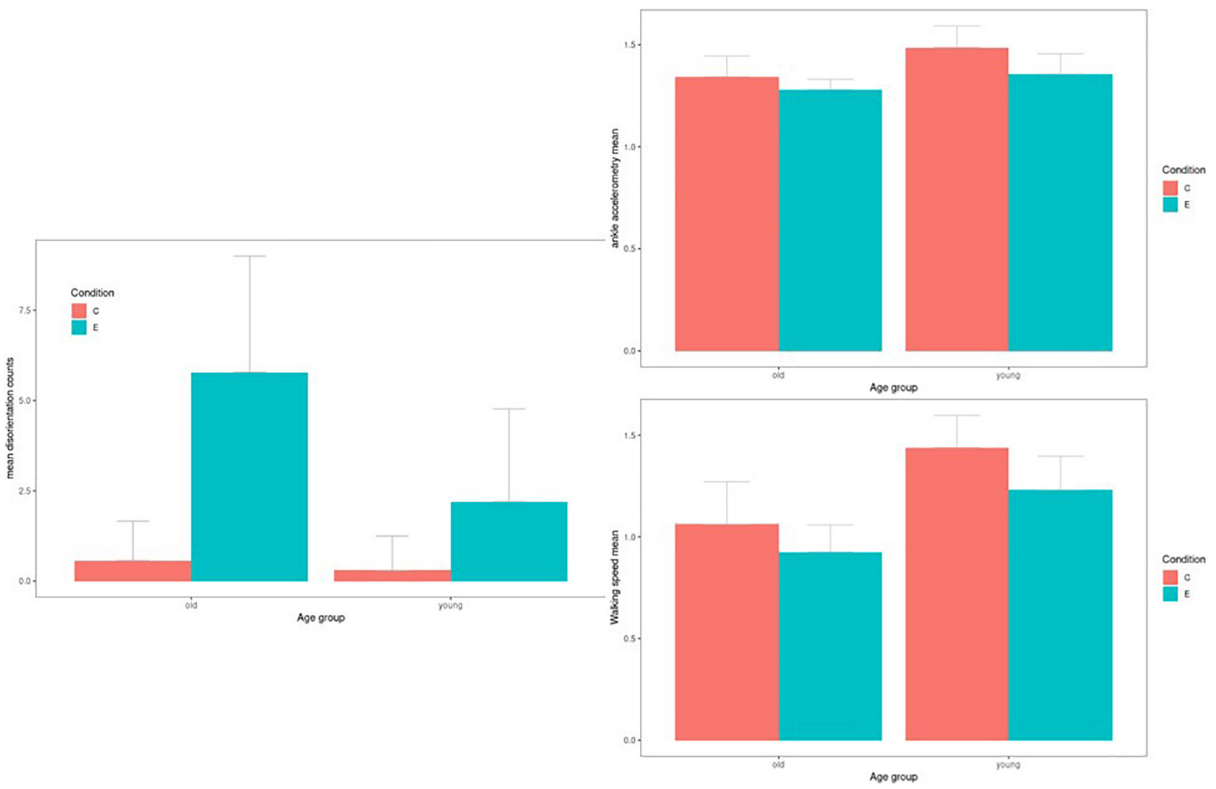


FIGURE 1 | Aggregated disorientation events, accelerometry, and walking speed by age group by condition. Disorientation events (upper row), ankle worn accelerometric signal (middle row), and walking speed mean (lower row) according to age group and condition [control (C) *vs*. experimental (E)]. Bars show mean and 95% credibility intervals.

features and orientation status across all data. We chose a radial kernel as the plotting of data did not indicate a linear separation (see **Figure 4**), with a cost parameter of 10 and a gamma parameter of 1, based on the initial grid search using the whole data set. Subsequently, we determined group discrimination within each patient fold applied to a random selection of 80% of the data as a training sample and the remaining 20% of data as a test sample. The mean accuracy of correctly allocated instances of orientation/disorientation was 72% (SD 11%) across the cross-validated patient folds.

Using a generative Hidden Markov model implemented in library "depmixS4" in R reached an average accuracy of correctly allocated instances of orientation/disorientation of only 64% (SD 14%) when comparing the binary states of oriented/disoriented as generated from the observed variables ankle-worn accelerometric signal and walking speed mean and variance as compared with

TABLE 2 | Number of disorientation events by cognitive, accelerometric, and gait features.

Cognitive scores				
Independent variables	**Main effect cognitive score**	**Condition**	**Age (years)**	**Gender**
Visuospatial	0.01 (–0.25 to 0.29)	**2.26 (1.56** to **3.05)**	**0.04 (0** to **0.08)**	–0.04 (–0.5 to 0.4)
Executive function	**–0.2 (–0.4** to **0.01)**	**2.2 (1.49** to **3.01)**	**0.06 (0.01** to **0.11)**	–0.09 (–0.53 to 0.34)
Visual memory	–0.15 (–1.22 to 1)	**2.29 (1.57** to **3.11)**	**0.04 (0** to **008)**	–0.03 (–0.47 to 0.37)
PTSOT	–0.12 (–0.82 to 0.48)	**2.26 (1.52** to **3.07)**	0.02 (–0.03 to 0.07)	0.1 (–0.47 to 0.67)
Accelerometric features				
Independent variables	**Main effect accelerometry**	**Condition**	**Age (years)**	**Gender**
Ankle mean (g)	**–4.62 (–7.66** to **–1.73)**	**1.97 (1.34** to **2.63)**	**0.01 (0.01** to **0.02)**	0.01 (–0.37 to 0.38
Wrist mean (g)	3.02 (–1.79 to 7.52)	**2.25 (1.68** to **2.93)**	**0.02 (0.01** to **0.03)**	–0.22 (–0.64 to 0.2)
Chest mean (g)	1.01 (–11.18 to 13.17)	**2.23 (1.63** to **2.92)**	**0.02 (0.01** to **0.03)**	–0.08 (–0.47 to 0.31)
Gait features				
Independent variables	**Main effect gait**	**Condition**	**Age (years)**	**Gender**
Walking speed (m/s)	**–2.23 (–3.46** to **–0.99)**	**1.97 (1.38** to **2.63)**	**0.01 (0** to **0.02)**	0.08 (–0.3 to 0.44)
Step length (m)	**–2.85 (–5.45** to **–0.31)**	**2.04 (1.44** to **2.75)**	**0.01 (0** to **0.02)**	0 (–0.38 to 0.37)
Stride time (s)	–0.06 (–1.08 to 0.87)	**2.24 (1.61** to **2.96)**	**0.02 (0.01** to **0.03)**	–0.09 (–0.48 to 0.31)
Step width (m)	0.93 (–4.56 to 6.32)	**2.22 (1.63** to **2.9)**	**0.02 (0.01** to **0.03)**	–0.12 (–0.56 to 0.3)
Stance time (s)	0.43 (–0.73 to 1.56)	**2.33 (1.71** to **3.12)**	**0.02 (0.01** to **0.03)**	–0.04 (–0.43 to 0.33)
Swing time (s)	–2.78 (–5.91 to 0.11)	**2.11 (1.5** to **2.8)**	**0.02 (0.01** to **0.03)**	–0.15 (–0.52 to 0.22)

Gender = factor level effects for male vs. female sex.
Cognitive variables represent domain scores derived as the mean score of specific tests, after transformation to z-scores.
Values in bold indicate effects where the 95% credibility interval excludes 0.
g = acceleration constant g (1 g = 9.81 m/s^2).
m = meter.
s = seconds.

the observed disorientation instances. As can be seen from **Figure 5**, the Hidden Markov model produced substantially fewer disorientation states than had been observed (**Figure 5A**), and accuracy decreased with a higher number of observed disorientation states per patient (**Figure 5B**), with a correlation coefficient of –0.51.

DISCUSSION

Here, we studied the association of accelerometric and gait features with episodes of disorientation in cognitively normal young and older adults in a hybrid experiment. We found that decreased accelerometric signal from ankle-worn sensors as well as decrease in walking speed and step length were associated with a higher number of aggregated disorientation events. Similarly, decreases in accelerometric signal and changes in a range of spatiotemporal gait features were associated with a higher number of episodes of disorientation in real time. At the same time, the prediction accuracy of single accelerometric and gait features for episodes of disorientation in real time was below 60%. However, even when combining the most strongly associated features in a multivariate non-linear support vector machine, reached only 72% accuracy for correctly allocated instances of orientation/disorientation. This level of accuracy would not be sufficient for individual detection of disorientation episodes and situation-aware assistance. Thus, we were able to confirm the expected association of accelerometric and gait characteristics with disorientation in cognitively unimpaired individuals, but we did not find sufficient accuracy for individual prediction.

Our study was able to replicate the age-related decline in spatiotemporal gait features that has been reported in a large number of studies, systematically reviewed in Herssens et al. (2018) and Osoba et al. (2019). Spatial orientation requires visuospatial abilities and higher-order cognitive processes, such as egocentric and allocentric representations, cognitive mapping, spatial strategies, encoding, and processing of spatial information (Lithfous et al., 2013; Meneghetti et al., 2014; Muffato et al., 2016). In our study, we focused on the domains of visual memory, visuoconstructional ability, executive function, and spatial orientation. Our results demonstrated a relationship between executive function and aggregated orientation in older adults; the number of disorientation events was lower in individuals with higher executive function. In this study, we had used the ratio of Trail Making Test B to A as a measure of executive function, assessing motor speed and visual speed (Arbuthnott and Frank, 2000; Sanchez-Cubillo et al., 2009). The Trail Making Test ratio serves as an index of executive control function because it can provide an independent measure of cognitive flexibility (Bezdicek et al., 2017). Moreover, it has also been associated with frontal executive function (Arbuthnott and Frank, 2000). An association of executive functions and effective

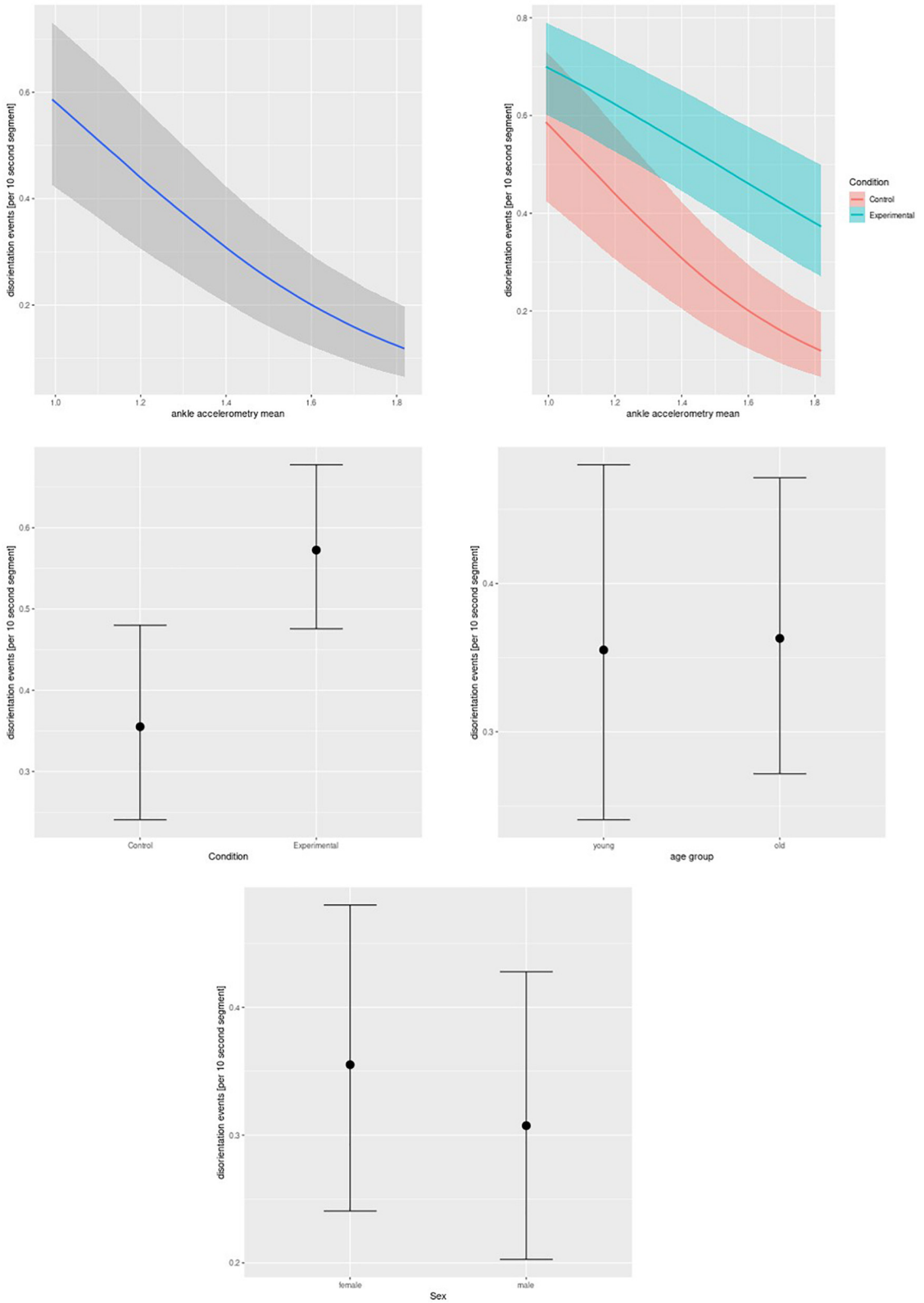


FIGURE 2 | Real-time data, ankle worn accelerometry. Bayesian mixed-effect logistic regression of disorientation events on ankle worn accelerometric signal (main effect, upper left and interaction effect with condition, upper right), condition (experimental or control, middle left), age group (middle right), and gender (lower row). The graphs feature mean effects and 95% credibility intervals.

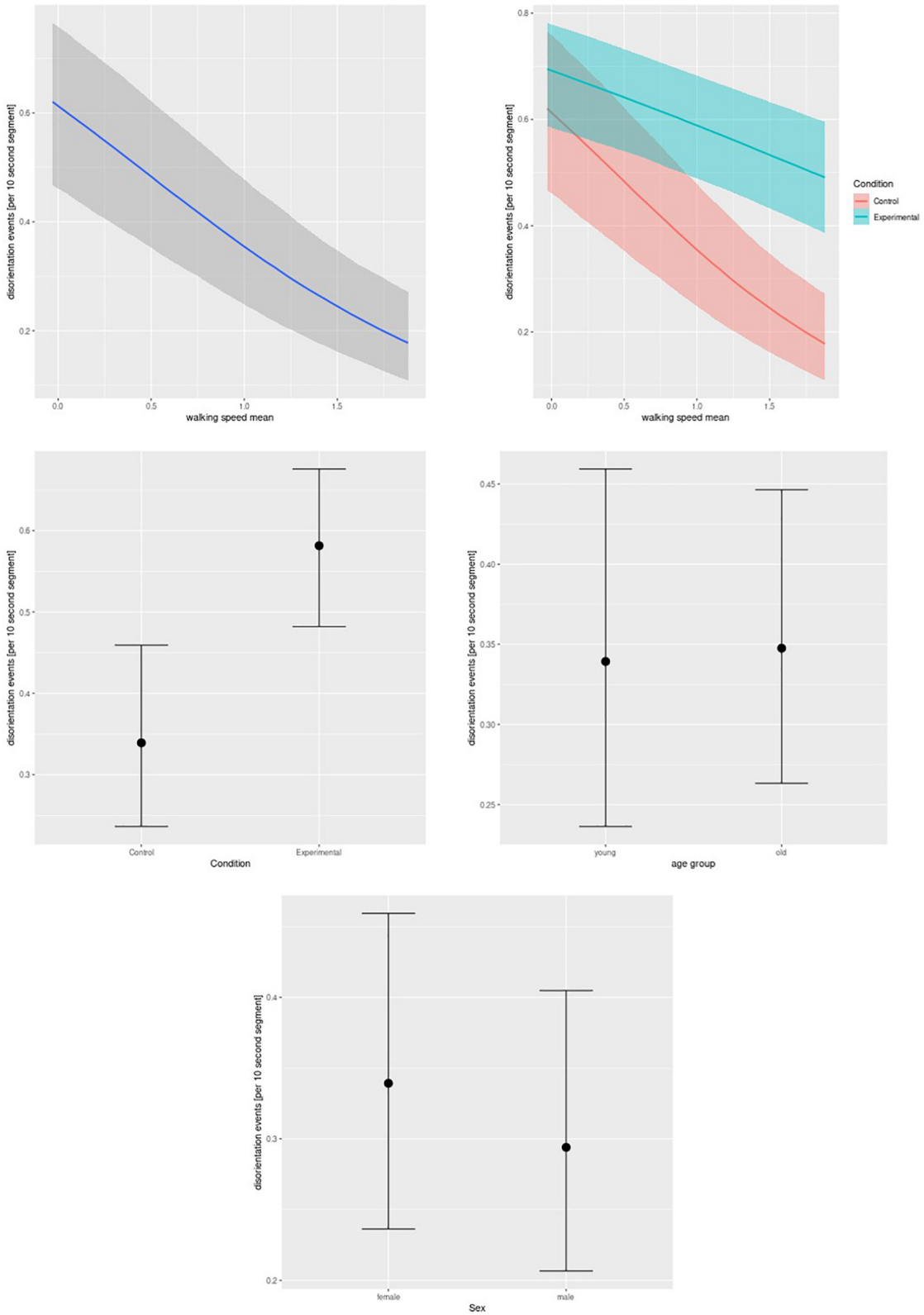


FIGURE 3 | Real-time data, walking speed. Bayesian mixed-effect logistic regression of disorientation events on mean walking speed (main effect, upper left and interaction effect with condition, upper right), condition (experimental or control, middle left), age group (middle right), and gender (lower row). The graphs feature mean effects and 95% credibility intervals.

TABLE 3 | Incidence of disorientation events and accelerometric and gait features in real time.

Accelerometric features					
Independent variables	**Main effect**	**Accelerometry by Condition**	**Condition**	**Age group**	**Gender**
Ankle mean	**–0.48 (–0.64** to **–0.32)**	**0.2 (0.02** to **0.38)**	**0.93 (0.44** to **1.39)**	0.04 (–0.45 to 0.49)	–0.19 (–0.65 to 0.26)
Wrist mean	**–0.67 (–0.91** to **–0.43)**	**0.27 (0.01** to **0.54)**	**1.21 (0.78** to **1.65)**	–0.16 (–0.61 to 0.29)	–0.03 (–0.47 to 0.4)
Chest mean	**–0.45 (–0.68** to **–0.24)**	0.18 (–0.08 to 0.43)	**0.99 (0.53** to **1.46)**	–0.23 (–0.76 to 0.28)	–0.21 (–0.66 to 0.24)
Gait features					
Independent variables	**Main effect**	**Gait by Condition**	**Condition**	**Age group**	**Gender**
Walking speed	**–0.47 (–0.62** to **–0.32)**	**0.27 (0.1** to **0.44)**	**1 (0.56** to **1.44)**	0.04 (–0.44 to 0.5)	–0.2 (–0.64 to 0.22)
Step length	**–2.69 (–3.5** to **–1.85)**	**2.08 (1.13** to **2.99)**	**1.07 (0.62** to **1.52)**	0.07 (–0.35 to 0.52)	–0.21 (–0.67 to 0.22)
Stride time	**0.19 (0.11** to **0.29)**	–0.03 (–0.21 to 0.21)	**1.17 (0.72** to **1.61)**	0.27 (–0.16 to 0.7)	–0.29 (–0.7 to 0.12)
Step width	**–0.71 (–0.95** to **–0.5)**	**0.79 (0.55** to **1.04)**	**1.36 (0.92** to **1.84)**	0.21 (–0.24 to 0.67)	–0.32 (–0.78 to 0.14)
Stance time	**0.14 (0.06** to **0.22)**	0.05 (–0.14 to 0.28)	**1.18 (0.73** to **1.61)**	0.25 (–0.16 to 0.67)	–0.32 (–0.73 to 0.12)
Swing time	**0.34 (0.19** to **0.52)**	**–0.27 (–0.5** to **–0.04)**	**1.17 (0.74** to **1.61)**	0.27 (–0.17 to 0.7)	–0.3 (–0.74 to 0.14)

Age group = old vs. young.
Gender = factor level effects for male vs. female sex.
The accelerometric and gait variables were z-score transformed before being entered into the models.
Values in bold indicate effects where the 95% credibility interval excludes 0.

spatial navigation has been previously reported (Wei et al., 2020; Laczo et al., 2021). Based on our results, we assume that higher executive functions play an important role in tasks requiring the use of effective wayfinding strategies. Effects on visuospatial abilities were absent, whereas effects on visual memory were not conclusive. We had expected an association between these domains, since they have been implicated in navigation efficiency and environment learning (Meneghetti et al., 2014; Wei et al., 2020). Previous studies have demonstrated an age-related decline in navigation skills, due to difficulties in environment route learning and spatial recall of relationships between landmarks and directions at decisions points (Zhong and Moffat, 2016; Ramanoel et al., 2020). The absence of an effect, therefore, was unexpected. A post hoc explanation would relate to previous observations that paper–pencil testing of spatial abilities found a poor correlation with real-world navigation performance (Nadolne and Stringer, 2001; Taillade et al., 2015), which has been used as an argument for the creation of novel ecologically valid test instruments (Nadolne and Stringer, 2001).

Furthermore, we had expected an association of orientation with the Perspective Taking/Spatial Orientation Test, since previous work suggested alterations of egocentric topographic orientation in older adults (Caffo et al., 2020). Two of the 28 participants, however, were not able to perform the task at all and several participants had difficulties when performing the Perspective Taking/Spatial Orientation Test. As we saw in practice, it was challenging for our participants to understand the task instructions and they might have felt overstrained. Difficulties regarding the understanding of instructions on similar tasks have been previously reported in young adults (Hegarty and Waller, 2004). Although the Perspective Taking/Spatial Orientation Test by design seemed well suited to test a trait of orientation ability and it has been widely used in spatial cognition literature (Friedman et al., 2020), it was not easy to use, at least in our hands, even for cognitively normal older people. The test has only been used in a few previous studies with older people (Zancada-Menendez et al., 2016) who on average were 8–10 years younger than our older group of participants.

A relationship between gait characteristics and disorientation has already been demonstrated in conditions such as delirium and dementia (Arjunan et al., 2019; Evensen et al., 2019; Oh, 2021; Weizman et al., 2021). In contrast, the detection of disorientation events using gait and accelerometry features has

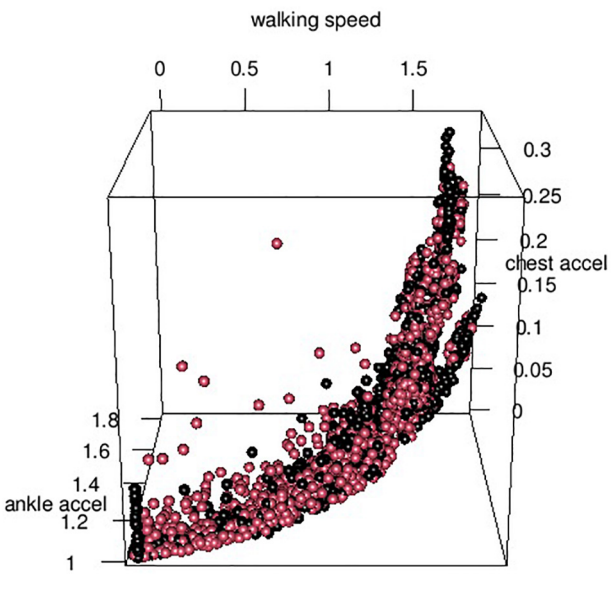


FIGURE 4 | Distribution of orientation status across features. Three-dimensional representation of the distribution of orientation status (oriented – black beads, disoriented – red beads) across ankle and chest-worn accelerometric signal and walking speed mean.

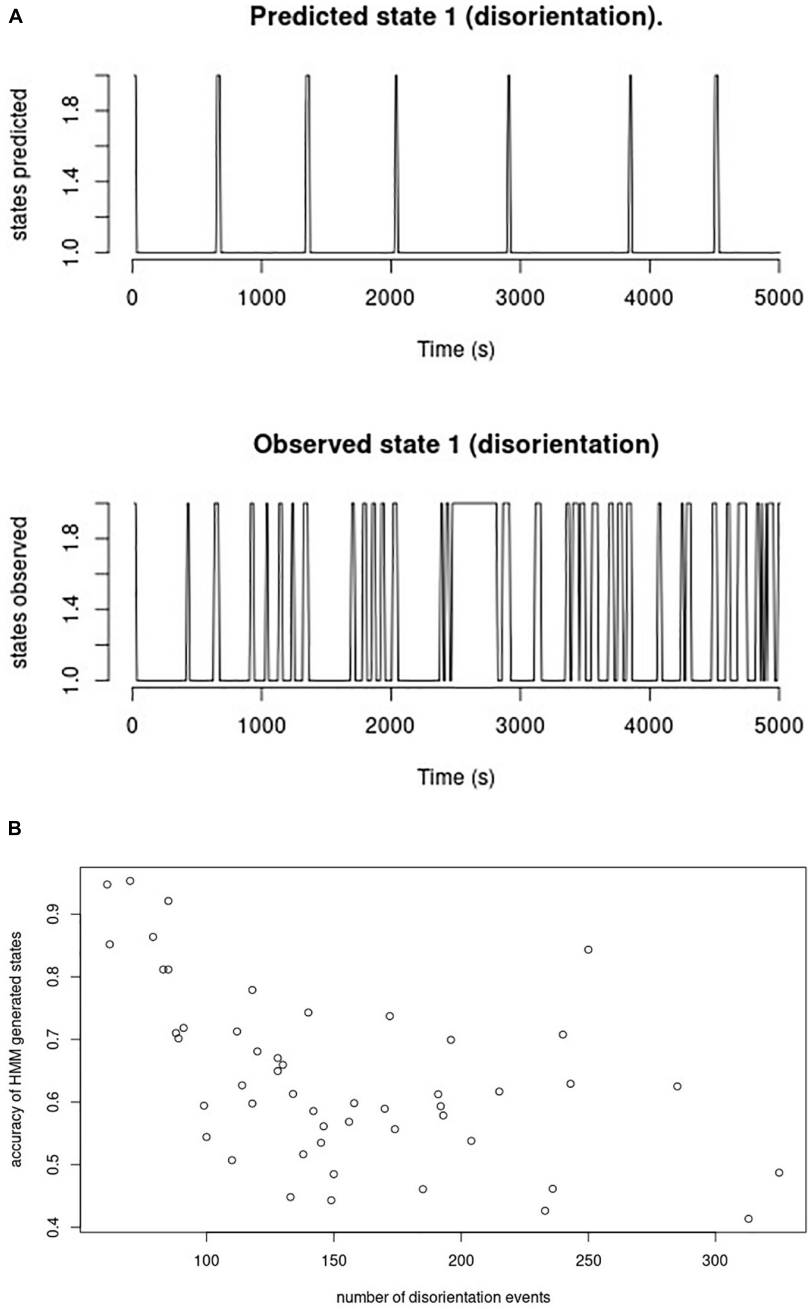


FIGURE 5 | Hidden Markov model generated states and observed orientation states. **(A)** Time series of states within 5,000 s. The upper row plots the orientation states generated from the Hidden Markov model during the first 500 time segments (= 5,000 s, pooled across participants) with 1 = oriented, 2 = disoriented; the lower row plots the observed orientation states from the same time segments. **(B)** Association between number of disorientation events and accuracy of HMM generated states. This graph plots the accuracy of the HMM generated states relative to the observed states per participant (y-axis) vs. the number of disorientation events per participant (x-axis).

been little explored. In a similar set-up to our study, one previous study reported gait features for a group of 17 young and 17 older participants navigating on a treadmill through a virtual shopping mall (Kafri et al., 2021). However, detection of disorientation was not an outcome parameter in this earlier study. In the current study, we found that reduced ankle-worn accelerometry signal was associated with more disorientation events in both aggregated and real-time data. The reduction of walking speed and step length and the increase in stance, swing, and stride time was associated with more disorientation events. This is consistent with the reduction in overall signal from the ankle-worn sensors and suggests that the acceleration signal

may be useful as a surrogate measure for less easily measured gait characteristics, but with the caveat that none of the gait characteristics examined achieved a useful level of predictive accuracy for disorientation events.

Even when combining features in a non-linear support vector machine, the accuracy level in our hybrid set-up was below the accuracy level which we had achieved in a real-world experiment with people with MCI or dementia. In this study, the accelerometric features had achieved an AUC of 75% and 79% of correctly allocated instances of orientation/disorientation (Schaat et al., 2019). In the previous experiment, we struggled with the low occurrence of disorientation episodes relative to the total time of the experiment, which made training the models difficult and led to unbalanced sensitivity and specificity estimates. In this study, we wanted to improve this situation in a much more controlled environment. Based on this setting, we were able to focus on the time series at the decision points only and induce disorientation even in young individuals. Indeed, this approach was successful with a proportion of 41% of intervals being annotated as disorientated in the total time series, and 49% at the decision points compared with less than 10% in the previous real-world setting (Schaat et al., 2019). Although we achieved a higher proportion of disorientation events our models performed less accurately. There are several post hoc explanations for this unexpected result which also relate to the limitations of our study.

The limitations of our study include the following points: First, the measurement of orientation states was based on offline video annotation which carries some imprecision. However, inter-rater reliability was very good (Cohen's kappa > 0.8), and even using lagged features, allowing sensor values of a time frame of 30 s before the actual rating of disorientation to be included in the prediction models, did not alter the results. Second, the difference in set-up where walking on a treadmill and walking on a street pose different requirements on cognitive and motion abilities so that the resulting gait and movement features may not directly be comparable. A previous study reported a slower gait with shorter, less variable strides during treadmill walking compared with walking outdoors on the sideway in young and older adults (Schmitt et al., 2021). Thus, walking on a motorized treadmill may reduce the variability of gait characteristics compared with walking outdoors, thereby also reducing disorientation-induced changes in gait characteristics. Third, in this study, we had studied cognitively unimpaired individuals who may show less pronounced changes in walking behavior during episodes of disorientation than individuals with MCI or dementia who were lost in a real-world setting (Schaat et al., 2019). Fourth, from the Hidden Markov model, it became obvious that the model produced less instances of disorientation than were observed, that is, only approximately 64%. In comparison, the previous model for the real-world data had produced a high number of false alarms, that is, more instances of disorientation than had been observed (Schaat et al., 2019). This may suggest that grouping disorientation events into only two states (oriented *vs.* not oriented) was too simplistic for the present data. There may be different subtypes of disorientation states, each associated with different behavioral characteristics. For example, externally triggered disorientation events might represent a different category of disorientation states than spontaneously occurring disorientation events; however, the two states were not distinguished in our models. Finally, the sample size was relatively small in our study. Consequently, our study was only powered to detect moderate-to-large effects. The effort required to complete the experiment was high for each participant. So we had even considered to use a cross-over design where each participant would undergo both conditions, experimental and control, in a randomized, balanced design. We decided against this option because already the experiment with only one condition was exhausting for some of the older participants.

In summary, in a prospective analysis of young and older cognitively healthy adults in a hybrid environment featuring a treadmill-based navigation through a virtual environment, we found an association between executive function, ankle-worn accelerometric signal, and spatiotemporal gait features with an aggregated number of disorientation events across age groups and experimental conditions. This was replicated by an association of accelerometric signal and spatiotemporal gait features with disorientation events in the real-time data analysis. Despite these consistent associations, the predictive accuracy of single or combined acceleration and gait features was insufficient for individual detection of disorientation events in real time. The lessons from this analysis are that age-related and experimentally induced disorientation is reflected in spatiotemporal gait features and also in the accelerometric signal as a potentially more easily accessible surrogate for gait features. At the same time, such measurements probably need to be enriched with other parameters to be sufficiently accurate for individual prediction of disorientation events. In future directions, further experiments may test whether such predictions can be more accurate for people with dementia. For this group of individuals, based on our preliminary experience with a small number of patients, external induction of disorientation events is not necessary, as they already showed pronounced disorientation under undisturbed control conditions. Finally, the set-up of our experiment may be useful not only to monitor but even to train navigation abilities under dual-task conditions with high transfer potential to real-world environment.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics committee of the Rostock University Medical Center (Approval No. A 2019-0062). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

ST was involved in all stages of the work, contributing to the study design, research question, performed analyses and interpretation

of the data, and drafted and revised the manuscript. CA and SL contributed to the acquisition of the data, provided feedback and revised the manuscript. DG provided feedback and revised the manuscript. SF contributed to acquisition of the neuropsychological data, provided feedback, and revised the manuscript. SB provided feedback and revised the manuscript. TK contributed significantly to the conception and design of the study, provided feedback, and revised the manuscript. All authors read and approved the manuscript.

FUNDING

The GRAIL was funded by the German Research Community (DFG, INST 264/137-1 FUGG).

ACKNOWLEDGMENTS

We thank Martin Gube (Institute for Sports Science, University of Rostock, Rostock, Germany) for assistance with the laboratory work. Part of the presented material was from the doctoral theses of Charlotte Hinz, Anne Klostermann, and Isabell Kampa. We also thank Charlotte Hinz, Anne Klostermann, and Isabell Kampa for their contribution to the acquisition of the data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.882446/full#supplementary-material

REFERENCES

Albert, M. S., Dekosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., et al. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* 7, 270–279. doi: 10.1016/j.jalz.2011.03.008

Amaefule, C. O., Ludtke, S., Kirste, T., and Teipel, S. J. (2020). Effect of spatial disorientation in a virtual environment on gait and vital features in patients with dementia: pilot single-blind randomized control trial. *JMIR Serious Games* 8:e18455. doi: 10.2196/18455

Arbuthnott, K., and Frank, J. (2000). Trail making test, part B as a measure of executive control: validation using a set-switching paradigm. *J. Clin. Exp. Neuropsychol.* 22, 518–528. doi: 10.1076/1380-3395(200008)22:4;1-0;FT518

Arjunan, A., Peel, N. M., and Hubbard, R. E. (2019). Gait speed and frailty status in relation to adverse outcomes in geriatric rehabilitation. *Arch. Phys. Med. Rehabil.* 100, 859–864. doi: 10.1016/j.apmr.2018.08.187

Baudendistel, S. T., Schmitt, A. C., Stone, A. E., Raffegeau, T. E., Roper, J. A., and Hass, C. J. (2021). Faster or longer steps: maintaining fast walking in older adults at risk for mobility disability. *Gait Posture* 89, 86–91. doi: 10.1016/j.gaitpost.2021.07.002

Beauchet, O., Allali, G., Sekhon, H., Verghese, J., Guilain, S., Steinmetz, J. P., et al. (2017). Guidelines for assessment of gait and reference values for spatiotemporal gait parameters in older adults: the biomathics and canadian gait consortiums initiative. *Front. Hum. Neurosci.* 11:353. doi: 10.3389/fnhum.2017.00353

Becu, M., Sheynikhovich, D., Tatur, G., Agathos, C. P., Bologna, L. L., Sahel, J. A., et al. (2020). Age-related preference for geometric spatial cues during real-world navigation. *Nat. Hum. Behav.* 4, 88–99. doi: 10.1038/s41562-019-0718-z

Bezdicek, O., Stepankova, H., Axelrod, B. N., Nikolai, T., Sulc, Z., Jech, R., et al. (2017). Clinimetric validity of the Trail Making Test Czech version in Parkinson's disease and normative data for older adults. *Clin. Neuropsychol.* 31, 42–60. doi: 10.1080/13854046.2017.1324045

Caffo, A. O., Lopez, A., Spano, G., Serino, S., Cipresso, P., Stasolla, F., et al. (2018). Spatial reorientation decline in aging: the combination of geometry and landmarks. *Aging Ment. Health* 22, 1372–1383. doi: 10.1080/13607863.2017.1354973

Caffo, A. O., Lopez, A., Spano, G., Stasolla, F., Serino, S., Cipresso, P., et al. (2020). The differential effect of normal and pathological aging on egocentric and allocentric spatial memory in navigational and reaching space. *Neurol. Sci.* 41, 1741–1749. doi: 10.1007/s10072-020-04261-4

Cohen, J. A., and Verghese, J. (2019). Gait and dementia. *Handb. Clin. Neurol.* 167, 419–427.

Coley, N., Gallini, A., Ousset, P. J., Vellas, B., Andrieu, S., and Guidage Study, G. (2016). Evaluating the clinical relevance of a cognitive composite outcome measure: an analysis of 1414 participants from the 5-year GuidAge Alzheimer's prevention trial. *Alzheimers Dement.* 12, 1216–1225. doi: 10.1016/j.jalz.2016.06.002

Costa, R., Pompeu, J. E., Viveiro, L. A. P., and Brucki, S. M. D. (2020). Spatial orientation tasks show moderate to high accuracy for the diagnosis of mild cognitive impairment: a systematic literature review. *Arq. Neuropsiquiatr.* 78, 713–723. doi: 10.1590/0004-282X20200043

Cushman, L. A., Stein, K., and Duffy, C. J. (2008). Detecting navigational deficits in cognitive aging and Alzheimer disease using virtual reality. *Neurology* 71, 888–895. doi: 10.1212/01.wnl.0000326262.67613.fe

Dahmani, L., and Bohbot, V. D. (2020). Habitual use of GPS negatively impacts spatial memory during self-guided navigation. *Sci. Rep.* 10:6310. doi: 10.1038/s41598-020-62877-0

Evensen, S., Bourke, A. K., Lydersen, S., Sletvold, O., Saltvedt, I., Wyller, T. B., et al. (2019). Motor activity across delirium motor subtypes in geriatric patients assessed using body-worn sensors: a Norwegian cross-sectional study. *BMJ Open* 9:e026401. doi: 10.1136/bmjopen-2018-026401

Friedman, A., Kohler, B., Gunalp, P., Boone, A. P., and Hegarty, M. (2020). A computerized spatial orientation test. *Behav. Res. Methods* 52, 799–812. doi: 10.3758/s13428-019-01277-3

Gazova, I., Vlcek, K., Laczo, J., Nedelska, Z., Hyncicova, E., Mokrisova, I., et al. (2012). Spatial navigation-a unique window into physiological and pathological aging. *Front. Aging Neurosci.* 4:16. doi: 10.3389/fnagi.2012.00016

Hegarty, M., and Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence* 32, 175–191. doi: 10.1016/j.intell.2003.12.001

Herssens, N., Verbecque, E., Hallemans, A., Vereeck, L., Van Rompaey, V., and Saeys, W. (2018). Do spatiotemporal parameters and gait variability differ across the lifespan of healthy adults? A systematic review. *Gait Posture* 64, 181–190. doi: 10.1016/j.gaitpost.2018.06.012

Iachini, I., Iavarone, A., Senese, V. P., Ruotolo, F., and Ruggiero, G. (2009). Visuospatial memory in healthy elderly, AD and MCI: a review. *Curr. Aging Sci.* 2, 43–59. doi: 10.2174/1874609810902010043

Kafri, M., Weiss, P. L., Zeilig, G., Bondi, M., Baum-Cohen, I., and Kizony, R. (2021). Performance in complex life situations: effects of age, cognition, and walking speed in virtual versus real life environments. *J. Neuroeng. Rehabil.* 18:30. doi: 10.1186/s12984-021-00830-6

Kizony, R., Zeilig, G., Krasovsky, T., Bondi, M., Weiss, P. L., Kodesh, E., et al. (2017). Using virtual reality simulation to study navigation in a complex environment as a functional-cognitive task; a pilot study. *J. Vestib. Res.* 27, 39–47. doi: 10.3233/VES-170605

Laczo, M., Wiener, J. M., Kalinova, J., Matuskova, V., Vyhnalek, M., Hort, J., et al. (2021). Spatial navigation and visuospatial strategies in typical and atypical aging. *Brain Sci.* 11:1421. doi: 10.3390/brainsci11111421

Lester, A. W., Moffat, S. D., Wiener, J. M., Barnes, C. A., and Wolbers, T. (2017). The aging navigational system. *Neuron* 95, 1019–1035. doi: 10.1016/j.neuron.2017.06.037

Lithfous, S., Dufour, A., and Despres, O. (2013). Spatial navigation in normal aging and the prodromal stage of Alzheimer's disease: insights from imaging and behavioral studies. *Ageing Res. Rev.* 12, 201–213. doi: 10.1016/j.arr.2012.04.007

Lohman, E. B. III, Balan Sackiriyas, K. S., and Swen, R. W. (2011). A comparison of the spatiotemporal parameters, kinematics, and biomechanics between shod, unshod, and minimally supported running as compared to walking. *Phys. Ther. Sport* 12, 151–163. doi: 10.1016/j.ptsp.2011.09.004

Mc Ardle, R., Del Din, S., Donaghy, P., Galna, B., Thomas, A. J., and Rochester, L. (2021). The impact of environment on gait assessment: considerations from real-world gait analysis in dementia subtypes. *Sensors (Basel)* 21:813. doi: 10.3390/s21030813

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R. Jr., Kawas, C. H., et al. (2011). The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* 7, 263–269. doi: 10.1016/j.jalz.2011.03.005

Meneghetti, C., Ronconi, L., Pazzaglia, F., and De Beni, R. (2014). Spatial mental representations derived from spatial descriptions: the predicting and mediating roles of spatial preferences, strategies, and abilities. *Br. J. Psychol.* 105, 295–315. doi: 10.1111/bjop.12038

Morris, J. C., Heyman, A., Mohs, R. C., Hughes, J. P., Van Belle, G., Fillenbaum, G., et al. (1989). The Consortium to Establish a Registry for Alzheimer's disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology* 39, 1159–1165. doi: 10.1212/wnl.39.9.1159

Muffato, V., Meneghetti, C., and De Beni, R. (2016). Not all is lost in older adults' route learning: the role of visuo-spatial abilities and type of task. *J. Environ. Psychol.* 47, 230–241. doi: 10.1016/j.jenvp.2016.07.003

Mulas, I., Putzu, V., Asoni, G., Viale, D., Mameli, I., and Pau, M. (2021). Clinical assessment of gait and functional mobility in Italian healthy and cognitively impaired older persons using wearable inertial sensors. *Aging Clin. Exp. Res.* 33, 1853–1864. doi: 10.1007/s40520-020-01715-9

Nadolne, M. J., and Stringer, A. Y. (2001). Ecologic validity in neuropsychological assessment: prediction of wayfinding. *J. Int. Neuropsychol. Soc.* 7, 675–682. doi: 10.1017/s1355617701766039

Oh, C. (2021). Single-task or dual-task? gait assessment as a potential diagnostic tool for Alzheimer's dementia. *J. Alzheimers Dis.* 84, 1183–1192. doi: 10.3233/JAD-210690

Osoba, M. Y., Rao, A. K., Agrawal, S. K., and Lalwani, A. K. (2019). Balance and gait in the elderly: a contemporary review. *Laryngoscope Investig. Otolaryngol.* 4, 143–153. doi: 10.1002/lio2.252

Osterrieth, P. A. (1944). Le test de copie d'une figure complexe; contribution à l'étude de la perception et de la mémoire [Test of copying a complex figure; contribution to the study of perception and memory]. *Arch. Psychol.* 30, 206–356.

Paliokas, I., Kalamaras, E., Votis, K., Doumpoulakis, S., Lakka, E., Kotsani, M., et al. (2020). Using a virtual reality serious game to assess the performance of older adults with frailty. *Adv. Exp. Med. Biol.* 1196, 127–139. doi: 10.1007/978-3-030-32637-1_13

Panel on Prevention of Falls in Older Persons, American Geriatrics Society and British Geriatrics Society (2011). Summary of the Updated American Geriatrics Society/British Geriatrics Society clinical practice guideline for prevention of falls in older persons. *J. Am. Geriatr. Soc.* 59, 148–157. doi: 10.1111/j.1532-5415.2010.03234.x

Pawlaczyk, N., Szmytke, M., Meina, M., Lewandowska, M., Stepniak, J., Balaj, B., et al. (2021). Gait analysis under spatial navigation task in elderly people-a pilot study. *Sensors (Basel)* 21:270. doi: 10.3390/s21010270

Ramanoel, S., Durteste, M., Becu, M., Habas, C., and Arleo, A. (2020). Differential brain activity in regions linked to visuospatial processing during landmark-based navigation in young and healthy older adults. *Front. Hum. Neurosci.* 14:552111. doi: 10.3389/fnhum.2020.552111

Rey, A. (1941). L'examen psychologique dans les cas d'encephopathie traumatique (The psychological examination of cases of traumatic encephalopathy). *Arch. Psychol.* 28, 286–340.

Sanchez-Cubillo, I., Perianez, J. A., Adrover-Roig, D., Rodriguez-Sanchez, J. M., Rios-Lago, M., Tirapu, J., et al. (2009). Construct validity of the Trail Making Test: role of task-switching, working memory, inhibition/interference control, and visuomotor abilities. *J. Int. Neuropsychol. Soc.* 15, 438–450. doi: 10.1017/S1355617709090626

Schaat, S., Koldrack, P., Yordanova, K., Kirste, T., and Teipel, S. (2019). Real-time detection of spatial disorientation in persons with mild cognitive impairment and dementia. *Gerontology* 66, 85–94. doi: 10.1159/000500971

Schmitt, A. C., Baudendistel, S. T., Lipat, A. L., White, T. A., Raffegeau, T. E., and Hass, C. J. (2021). Walking indoors, outdoors, and on a treadmill: gait differences in healthy young and older adults. *Gait Posture* 90, 468–474. doi: 10.1016/j.gaitpost.2021.09.197

Taillade, M., N'kaoua, B., and Sauzeon, H. (2015). Age-related differences and cognitive correlates of self-reported and direct navigation performance: the effect of real and virtual test conditions manipulation. *Front. Psychol.* 6:2034. doi: 10.3389/fpsyg.2015.02034

Tascon, L., Castillo, J., Leon, I., and Cimadevilla, J. M. (2018). Walking and non-walking space in an equivalent virtual reality task: sexual dimorphism and aging decline of spatial abilities. *Behav. Brain Res.* 347, 201–208. doi: 10.1016/j.bbr.2018.03.022

Teipel, S., Babiloni, C., Hoey, J., Kaye, J., Kirste, T., and Burmeister, O. K. (2016). Information and communication technology solutions for outdoor navigation in dementia. *Alzheimers Dement.* 12, 695–707. doi: 10.1016/j.jalz.2015.11.003

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* 27, 1413–1432. doi: 10.1007/s11222-016-9696-4

Voss, T., Li, J., Cummings, J., Farlow, M., Assaid, C., Froman, S., et al. (2018). Randomized, controlled, proof-of-concept trial of MK-7622 in Alzheimer's disease. *Alzheimers Dement. (N Y)* 4, 173–181. doi: 10.1016/j.trci.2018.03.004

Wei, E. X., Anson, E. R., Resnick, S. M., and Agrawal, Y. (2020). Psychometric tests and spatial navigation: data from the baltimore longitudinal study of aging. *Front. Neurol.* 11:484. doi: 10.3389/fneur.2020.00484

Weizman, Y., Tirosh, O., Beh, J., Fuss, F. K., and Pedell, S. (2021). Gait assessment using wearable sensor-based devices in people living with dementia: a systematic review. *Int. J. Environ. Res. Public Health* 18:12735. doi: 10.3390/ijerph182312735

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). "Elan: a professional framework for multimodality research," in *Proceedings of the 5th International Conference on Language Resources and Evaluation*. (Genoa).

Yordanova, K., Koldrack, P., Heine, C., Henkel, R., Martin, M., Teipel, S., et al. (2017). Situation model for situation-aware assistance of dementia patients in outdoor mobility. *J. Alzheimers Dis.* 60, 1461–1476. doi: 10.3233/JAD-170105

Zakzanis, K. K., Quintin, G., Graham, S. J., and Mraz, R. (2009). Age and dementia related differences in spatial navigation within an immersive virtual environment. *Med. Sci. Monit.* 15, CR140–CR150.

Zancada-Menendez, C., Sampedro-Piquero, P., Lopez, L., and Mcnamara, T. P. (2016). Age and gender differences in spatial perspective taking. *Aging Clin. Exp. Res.* 28, 289–296. doi: 10.1007/s40520-015-0399-z

Zhong, J. Y., and Moffat, S. D. (2016). Age-related differences in associative learning of landmarks and heading directions in a virtual navigation task. *Front. Aging Neurosci.* 8:122. doi: 10.3389/fnagi.2016.00122

Conflict of Interest: ST participated in scientific advisory boards of Roche Pharma AG, Biogen, GRIFOLS, EISAI, and MSD and received lecture fees from Roche and MSD.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Computer Science

REVIEW
published: 01 July 2022
doi: 10.3389/fcomp.2022.869123

A Review on the Role of Affective Stimuli in Event-Related Frontal Alpha Asymmetry

Priya Sabu[1,2], *Ivo V. Stuldreher*[1], *Daisuke Kaneko*[3] *and Anne-Marie Brouwer*[1]*

[1] The Netherlands Organisation for Applied Scientific Research (TNO), Department Human Performance, Soesterberg, Netherlands, [2] Mechanical, Maritime and Materials Engineering (3mE), Delft University, Delft, Netherlands, [3] Kikkoman Europe R&D Laboratory B.V., Wageningen, Netherlands

Frontal alpha asymmetry refers to the difference between the right and left alpha activity over the frontal brain region. Increased activity in the left hemisphere has been linked to approach motivation and increased activity in the right hemisphere has been linked to avoidance or withdrawal. However, research on alpha asymmetry is diverse and has shown mixed results, which may partly be explained by the potency of the used stimuli to emotionally and motivationally engage participants. This review gives an overview of the types of affective stimuli utilized with the aim to identify which stimuli elicit a strong approach-avoidance effect in an affective context. We hope this contributes to better understanding of what is reflected by alpha asymmetry, and in what circumstances it may be an informative marker of emotional state. We systematically searched the literature for studies exploring event-related frontal alpha asymmetry in affective contexts. The search resulted in 61 papers, which were categorized in five stimulus categories that were expected to differ in their potency to engage participants: images & sounds, videos, real cues, games and other tasks. Studies were viewed with respect to the potency of the stimuli to evoke significant approach-avoidance effects on their own and in interaction with participant characteristics or condition. As expected, passively perceived stimuli that are multimodal or realistic, seem more potent to elicit alpha asymmetry than unimodal stimuli. Games, and other stimuli with a strong task-based component were expected to be relatively engaging but approach-avoidance effects did not seem to be much clearer than the studies using perception of videos and real cues. While multiple factors besides stimulus characteristics determine alpha asymmetry, and we did not identify a type of affective stimulus that induces alpha asymmetry highly consistently, our results indicate that strongly engaging, salient and/or personally relevant stimuli are important to induce an approach-avoidance effect.

Keywords: alpha asymmetry, EEG, approach-avoidance, emotion, motivation, computational psychophysiology, affective computing, mental state monitoring

OPEN ACCESS

Edited by:
Siyuan Chen,
University of New South Wales, Australia

Reviewed by:
Peter König,
Osnabrück University, Germany
Giulia Cartocci,
Sapienza University of Rome, Italy
Katharina Paul,
University of Hamburg, Germany

***Correspondence:**
Anne-Marie Brouwer
anne-marie.brouwer@tno.nl

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *03 February 2022*
Accepted: *30 May 2022*
Published: *01 July 2022*

Citation:
Sabu P, Stuldreher IV, Kaneko D and Brouwer A-M (2022) A Review on the Role of Affective Stimuli in Event-Related Frontal Alpha Asymmetry. Front. Comput. Sci. 4:869123. doi: 10.3389/fcomp.2022.869123

INTRODUCTION

When examining the emotional experience of individuals with a certain product, task or situation, they are commonly asked about it. For instance, in food research, usage of explicit, verbal questionnaires is by far the most common way to assess consumers' emotional experience (Lagast et al., 2017; Kaneko et al., 2018). However, explicit, verbal measures have their shortcomings.

Firstly, social desirability and self-presentational concerns can influence self-reported measures (Gawronski and de Houwer, 2014). Dell et al. (2012) found that respondents were about 2.5 times more likely to favor a technology believed to be developed by the interviewer than an exactly identical alternative. Furthermore, questionnaires usually reflect summative emotions post-interaction (Lottridge et al., 2012). Explicit measures are not well-suited for continuous monitoring to understand how emotional experience changes over time, such as during the interaction with a product. Continuous self-reporting is demanding and adds another task, and affects the emotional experience itself. To overcome such limitations, researchers have been arguing for the use of implicit measures (Gawronski and de Houwer, 2014), such as those inferred from spontaneous behavior or physiological signals. These allow for more objective measures that are not affected by response biases and continuous observation of the individual's emotional or affective state (Reuderink et al., 2013).

The circumplex model of affect characterizes emotions by valence and arousal (Russell, 1980). Valence refers to pleasantness, i.e., the degree of positive or negative affect, whereas arousal refers to the energetic component of the emotion (alertness). Research has consistently linked skin conductance to arousal (Christopoulos et al., 2019; Bartolomé-Tomás et al., 2020). Also, other types of physiological responses have been found to generally map better on arousal rather than valence (Mauss and Robinson, 2009). Valence has been found to be more difficult to assess using physiological measures. In this regard, asymmetric frontal cortical activation is of particular interest for implicitly measuring emotional processes (Coan and Allen, 2004; Harmon-Jones et al., 2010; Diaz and Bell, 2012). Early research has reported high incidence of negative affect in individuals with unilateral left hemispheric brain damage (Alford, 1933; Goldstein, 2004). These patients showed increased negative responses, fear and pessimism about the future. On the other hand, patients with unilateral right hemisphere damage displayed euphoric reactions (Denny-brown et al., 1952), such as inappropriate presentation of positive affect and laughing (Scherer and Ekman, 1984). In the late 70s, patterns of emotion processing have been associated with differences in the EEG alpha band (8–12 Hz) between the left and right frontal cortex, and was termed frontal alpha asymmetry (Tucker et al., 1981; Ahern and Schwartz, 1985; Davidson et al., 1985). Note that alpha power is inversely related to brain activity, such that low alpha activity is taken as an indication of high regional brain activation (Cook et al., 1998; Allen et al., 2004a).

Initial research focused on an affective explanation of frontal alpha asymmetry responses to stimuli. Larger relative left hemispheric activation was argued to be associated with positively valenced stimuli and increased right hemispheric activation with negatively valenced stimuli (Briesemeister et al., 2013). Next to this valence model, the approach-avoidance, or approach-withdrawal, model was explored. In this model, activity in the right frontal cortex has been related to avoidance motivation, a tendency to withdraw from a certain stimulus, and activity in the left frontal cortex with approach motivation toward a stimulus (Davidson et al., 1990; Davidson and Irwin, 1999; Coan and Allen, 2003; Davidson, 2004; Alves et al., 2008; Harmon-Jones et al., 2010; Diaz and Bell, 2012). Since approach motivation is often associated with positive valence and avoidance with negative valence, the expected cortical activity patterns of these two theories overlap in many cases (Reuderink et al., 2013). Studies that specifically disentangled valence and approach-avoidance motivation were in line with the approach-avoidance model (Carver and Harmon-Jones, 2009; Berkman and Lieberman, 2010). The defining difference was found in the hemispheric activation pattern in response to anger (Reuderink et al., 2013). Anger as a negatively valenced emotion was found to be lateralized in the left hemisphere just like happiness instead of the right hemisphere as would be expected based on valence motivation (Davidson, 1984). Further support for the approach-avoidance model was found in transcranial magnetic stimulation experiments (Rutherford and Lindell, 2011).

Frontal alpha asymmetry as a tool to monitor motivational processes related to emotion would be desirable in a variety of application fields, such as marketing (including evaluating public service announcements, e.g., Inguscio et al., 2021), product design (e.g., cosmetics—Gabriel et al., 2021), human-computer interfaces, gaming and the diagnosis of affective disorders (Briesemeister et al., 2013). Another upcoming application and research area where frontal alpha asymmetry is highly relevant, is neuroesthetics (Babiloni et al., 2015; Cartocci et al., 2018, 2021; Daly et al., 2019). However, it is important to realize that frontal alpha asymmetry is not specific for motivational processes, but is also moderated by e.g., unilateral hand contractions (Harmon-Jones et al., 2010) and seating position (Baldwin and Penaranda, 2012). Variations in such factors between studies may underlie diverse results in recent literature, together with differences in data recording (e.g., noise, number of participants, recording length), processing and analysis methods (Smith et al., 2017). Additionally, researchers have used a wide variety of stimuli that were hypothesized to induce frontal alpha asymmetry and found mixed results. This review focuses on the factor of affective stimuli potentially affecting the approach-avoidance effect as measured by alpha asymmetry in the context of emotion. We expect that stimuli may crucially affect frontal alpha asymmetry through their potential to emotionally and motivationally engage the recorded individuals. Since frontal alpha asymmetry describes an approach- avoidance effect, affective stimuli that are strongly motivating in either of the directions are expected to produce clear results. Although it is extremely difficult to quantify this a priori (Brouwer et al., 2015b), we think some general expectations can be formulated for stimulus categories that are prevalent in alpha asymmetry emotional research.

We expect affective stimuli to induce strong approach-avoidance effects when they are engaging and realistic. In that sense, real stimuli that are part of an engaging task would be most effective. We expect that stimuli that are only perceived are less potent than active tasks. Within the "perception" category, we expect images and sounds (i.e., sensory unimodal stimuli that represent a certain object or situation) to be less potent than videos (bi-modal), followed by real cues (multimodal and realistic; the actual object or situation itself). Within the "action" category, we expect that games may be particularly engaging

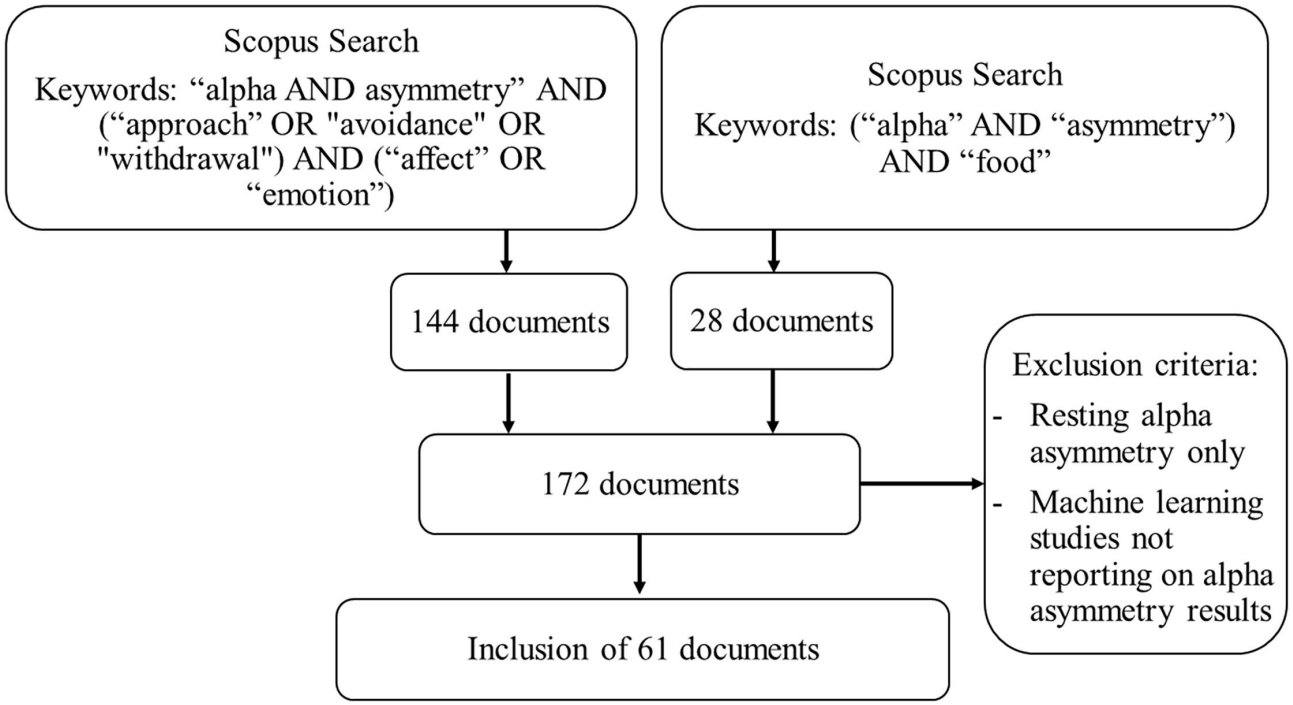


FIGURE 1 | Flow chart of inclusion process for the review.

tasks and therefore elicit strong approach-avoidance effects. Finally, clearer effects of affective stimuli on alpha asymmetry are expected if the stimuli are particularly relevant for the participants under study (e.g., food is likely to produce stronger approach motivation for individuals who have not eaten for a long time compared to individuals who have).

To date there is no review focused on the stimuli that can evoke an approach-avoidance effect measured by frontal alpha asymmetry. Hence, as of yet it is unclear which types of affective stimuli elicit a strong approach or avoidance effect. Exploring this will help to understand better what is reflected by frontal alpha asymmetry, under which circumstances frontal alpha asymmetry can be expected to be an informative marker of emotion and what causes the diversity in literature in order to unify conflicting results.

METHODS

Literature was searched on Scopus using the keywords "alpha AND asymmetry" AND ("approach" OR "avoidance" OR "withdrawal") AND ("affect" OR "emotion") and yielded 144 documents. Additionally, given our special interest in this measure from the perspective of studying food related emotion (Kaneko et al., 2018; Modica et al., 2018; Songsamoe et al., 2019), a search on Scopus using the terms ("alpha" AND "asymmetry") AND "food" was conducted as well, resulting in 28 more papers. Out of the resulting 172 documents only those that had measured frontal alpha asymmetry related to an event or a stimulus (i.e., not resting alpha asymmetry only) were included. Furthermore, studies using a machine learning approach without separately reporting on the exact alpha asymmetry results were excluded. This resulted in the inclusion of 61 papers. **Figure 1** visualizes the search and selection procedure.

The 61 selected studies were divided into five stimulus categories that were expected to systematically differ in their effectiveness to engage the subjects: 1. Images & sounds, 2. Videos, 3. Real cues, 4. Games, 5. Other tasks (Imagery; Modifying facial expression; Speech, reading and writing). While most studies involve some task, studies in the category "Games" and "Other tasks" specifically designed tasks to elicit a certain emotional state: performing the task serves as the main stimulus, and in case of games, the resulting or expected outcome in addition to performing the task.

Papers are summarized and evaluated per stimulus type. A summarized description of all 61 studies can be found in **Table 1**. Studies were rated based on whether the stimulus alone induced an alpha asymmetry approach-avoidance effect (one before last column in **Table 1**) and if applicable, whether alpha asymmetry approach-avoidance effects were found for, or in interaction with certain conditions or participant subgroups (last column in **Table 1**). Effects are indicated by "++" for a significant effect, "+" in case of a trend and "0" for no effect. **Supplementary Table 1** contains information on the context or goal of the 61 studies and more details about the stimuli and results. Furthermore, since cortical hemispheric specialization of emotion may differ between left- and right-handed individuals (Harmon-Jones et al., 2008; Walsh et al., 2017), handedness is indicated in the "Participants" column of **Supplementary Table 1**

TABLE 1 | Overview of studies arranged by stimulus types with (the hypothesized) approach-avoidance effect indicated by ++ (significant), + (trend) and 0 (none) of the stimulus alone, and/or other effects involving the stimulus. Note that many studies were set up for studying the 'other' effect (e.g. interaction with person characteristics or interaction between stimuli and other condition).

Study	Stimuli	Participants	Stimulus effect	Other effect
Images & sounds (n = 18): images (n = 16), sounds (n = 2)				
Deng et al. (2021)	Viewing pictures (neutral, positive, negative and drug-related contents) before and after drug abstaining training.	40 male drug abstainers: training group (n=20) and control group (n = 20)	0	++
Grassini et al. (2020)	Images depicting snakes, spiders, butterflies, and birds.	34 students (28 female)	0	
Gayathiri et al. (2020)	Neutral and high valence - high arousal pictures from the International Affective Picture System (IAPS).	15 adults suffering from major depressive disorder (7 male)	0	+
Adolph et al. (2017)	Negative, neutral, and positive emotional pictures.	43 students (28 female)	0	+
Schöne et al. (2016)	Erotic pictures, pictures of dressed attractive women and control pictures (and pictures of extreme sport and daily activities).	17 male students	++	
Gable and Poole (2014)	Anger pictures and neutral pictures.	32 students (15 female)	0	+
Uusberg et al. (2014)	Affective pictures ranging from very pleasant to unpleasant.	70 students (28 male)	0	
Ischebeck et al. (2014)	Neutral, aversive, and pictures related to OCD (obsessive-compulsive disorder).	20 patients (9 male) with (OCD) and 20 matched healthy controls (8 male)	0	+
Poole and Gable (2014)	Approach-positive, approach-negative, and withdrawal-negative pictures from the internet and IAPS.	48 students (36 female)	0	
Huster et al. (2009)	36 pictures from the IAPS in restricted randomized order- three pictures of the same affective category presented successively.	28 students (13 female)	++	
Rabe et al. (2008)	Four pictures (from IAPS) for 1 min each of category neutral, positive, negative, and trauma-related.	Participants with (subsyndromal) posttraumatic stress disorder receiving cognitive behavioral therapy (n = 17, 15 females) before and after therapy, wait-list controls (n = 18, 10 females)	0	++
Wiedemann et al. (1999)	Neutral, panic-relevant, anxiety-relevant but panic-irrelevant, or anxiety-irrelevant but emotionally relevant pictures, and performance of a motor task.	Patients with panic disorder (n = 23, 3 male) and controls (n = 25, 6 male)	0	++
Gable and Harmon-Jones (2008)	Pictures of dessert or neutral pictures of objects.	26 female students	0	++
Winter et al. (2016)	Food images.	58 female participants recorded twice: once fed, once fasted	0	+
Crabbe et al. (2007)	Unpleasant, neutral and pleasant IAPS pictures, before and after rest and exercise conditions.	34 young, fit and active volunteers (13 female)	0	0
Cartocci et al. (2018)	Six neutral images from IAPS, followed by ten ineffective, effective, and awarded anti-smoking Public Service Announcements.	3 heavy smokers, 11 light smokers, 15 non-smokers	0	++
Chen et al. (2015)	Scary and soothing sound stimuli.	18 students (16 male)	0	++
Papousek et al. (2018)	Three sound recordings: anger/aggression, sadness/desperation, neutral.	62 students (30 male)	0	++
Videos (n = 15)				
Olszewska-Guizzo et al. (2021)	Nine fixed-frame videos, filmed before the pandemic: busy downtown and residential green.	25 adult Singaporeans (14 female)	0	+
Joaquim et al. (2020)	Emotion-eliciting commercials: neutral, tenderness, amusement, sadness, disgust, anger and fear.	25 male and female subjects	+	
Zhao et al. (2018)	Emotion-eliciting film excerpts: tenderness, anger, and neutral.	37 students (17 males)	++	
Cartocci et al. (2017)	Spots and images of awarded, effective and ineffective antismoking public service announcements.	7 non-smokers, 9 light-smokers, 6 heavy-smokers	++	
Papousek et al. (2014)	Film comprising scenes of real injury and death.	148 female university students	++	
Prause et al. (2014)	Neutral and a sexually motivating film.	65 participants (22 females)	++	
Vecchiato et al. (2014)	TV commercials.	24 subjects (12 female) who liked or disliked the commercials	0	++

(Continued)

TABLE 1 | Continued

Study	Stimuli	Participants	Stimulus effect	Other effect
Hosseini et al. (2007)	Movie clips to induce relaxation, happiness, anxiety and sadness.	40 female students (extroverts, introverts, neurotics and emotionally stables)	0	++
Aftanas and Varlamov (2004)	Emotional film clips (neutral, relaxation, joy, anger, sexual arousal, disgust, fear, sadness, stress stimulation).	Non-alexithymic (n =27, 7 male) and alexithymic (n = 17, 14 male) participants	0	++
Hakim et al. (2021)	Video commercials of six food products.	33 (13 male) subjects	0	
Walsh et al. (2017)	Videos of food concerns (safety, hygiene and spoilage) and matched control videos.	40 students (31 female)	++	
Hajal et al. (2017)	Videos of own infants expressing distress.	26 mothers of 5- to 8-month-olds	++	+
McGeown and Davis (2018)	Chips to eat and video of confederate eating.	93 female students		++
Missana and Grossmann (2015)	Dynamic happy and fearful body expressions in two second clips.	20 4-month-old (10 female) and 20 8-month-old (10 female) infants	0	++
Lee et al. (2017)	Visual music as an emotional stimulus.	16 participants	++	
Real cues (n = 11)				
Kline et al. (2000)	Pleasant (vanilla), unpleasant (valerian), and neutral (water) odors.	58 women, aged between 58 and 70	++	
Kaneko et al. (2019)	Tasting different types of normal drinks, and diluted vinegar.	70 healthy participants (19 men)	0	
Lagast et al. (2020)	Tasting universally accepted (sucrose) and non-accepted (caffeine) solution, a personally selected accepted and non- accepted drink, and water.	32 participants	0	
Sargent et al. (2020)	Two machines to prepare hot beverage.	26 participants (14 females)	++	
Brouwer et al. (2017)	Cooking and tasting chicken or mealworms stir-fry dishes.	41 participants (19 female)	++	+
Olszewska-Guizzo et al. (2020)	Six landscape scenes (urban green and urban downtown).	22 adults (13 female)	+	
Knott et al. (2008)	Induction of neutral mood (holding a pen) or depressive mood (holding lighted cigarette over an ashtray without bringing to mouth).	11 (5 male) regular and 11 (6 male) light smokers	++	+
Modica et al. (2018)	Visual, visuo-tactile and exploration of food (daily food and comfort food; major and private label; foreign and local product).	Experiment 1: n = 19; experiment 2: n = 13 (5 males)	++	
Bolinger et al. (2020)	Positive prompts: express love, play peek-a-boo, sing; negative prompts: pretend infant has a rash, crawled to an electrical outlet.	25 infant (-parent dyads), 12 females	+	
Uusberg et al. (2015)	Degrees of social contact, varied by different gaze directions of a "live" model.	40 students (13 male)	0	++
Pönkänen and Hietanen (2012)	Neutral and smiling young females with a direct and an averted gaze, presented "live" through a liquid crystal shutter.	22 female undergraduates	0	0
Games (n = 5)				
Rodrigues et al. (2018)	Move around in a virtual T maze via joystick, with monster trial (negative event) and sheep trial (positive event).	30 participants (12 male)	++	++
Shankman et al. (2007)	Bogus computerized slot machine paradigm with three reels of numbers and fruit and two different payoff situations: reward and no incentive.	70 individuals with current MDD (29% male), 37 control participants (34% male)	0	+
Harmon-Jones et al. (2008)	Cues indicating that an easy, medium, or hard anagram would be presented and whether correct solution would result in receiving money or avoiding losing money.	Individuals with bipolar spectrum diagnosis (n = 41, 61% female) and individuals with no major affective psychopathology (n = 53, 49% female)	0	++
Miller and Tomarken (2001)	Delayed reaction time task including manipulations of incentive, expectancy, and response.	60 students (30 male)	++	
Sobotka et al. (1992)	Reward and punishment to responses to up or downward pointing arrows using finger press or finger lift response.	15 students (7 male)	++	

(Continued)

TABLE 1 | Continued

Study	Stimuli	Participants	Stimulus effect	Other effect
Other tasks (n = 12): Imagery (n = 4), modifying facial expression (n = 2), speech, reading and writing (n = 6)				
Mennella et al. (2015)	Imagery task including pleasant, neutral, and unpleasant narratives.	Dysphoric (n = 23) and non-dysphoric (n = 24) individuals	0	++
Wacker et al. (2008)	Emotional imagery of three scenarios of approach-avoidance conflict.	93 young men either high or low in trait behavioral inhibition system (BIS)	0	++
Wacker et al. (2003)	Emotional imagery.	109 male soccer players	++	
Papousek et al. (2017)	Reappraisal Inventiveness Test with anger-eliciting vignettes.	78 female university students	0	++
Stewart et al. (2014)	Producing approach (angry and happy) and withdrawal (afraid and sad) facial expressions.	Individuals with (n = 143) and without (n = 163) lifetime major depressive disorder	0	++
Coan et al. (2001)	Producing facial configurations denoting anger, disgust, fear, joy, and sadness.	36 students (10 male)	0	++
Pérez-Edgar et al. (2013)	Dot-probe paradigm with face pairs depicting angry, happy and neutral expressions, and stressful speech condition.	45 students (23 male)	0	++
Wang et al. (2015)	Public speech combined with reappraisal writing, irrelevant writing, or non-writing.	92 students	0	++
Li et al. (2016)	Writing task describing an anger-eliciting event, where participants were irritated by people with higher or lower social power.	29 students (13 male)	++	
Rejer and Jankowski (2017)	Internet advertisements during a text-reading task.	6 subjects (5 male)	+	
Brouwer et al. (2015a)	Reading a novel with emotional and non- emotional sections.	71 participants (35 female)	++	
Brooker et al. (2016)	Three emotion-eliciting episodes (conversation with experimenter, with stranger, stranger reading a script).	89 longitudinal twin sample (54% male)	++	

for all studies that report it. Most studies use right-handed participants and those that reported to have included left-handed persons stated that the results did not change by doing so.

RESULTS

Figure 2 presents the percentage of studies showing a significant effect of stimuli alone and in interaction with other conditions or participant subgroups, separately for each of the five stimulus categories. In the next sections, studies are discussed per stimulus category.

Images and Sounds (n = 18)

Most studies using sensory unimodal stimuli used visual (n = 16) rather than auditory (n = 2) stimuli. Studies utilizing unimodal stimuli appeared to show a significant effect only when the stimulus was particularly relevant for the participant group. This pattern can be seen in several clinical and substance-related studies.

Rabe et al. (2008) described that patients with posttraumatic stress disorder from motor vehicle accidents had increased right-sided activation during exposure to trauma-related pictures compared to neutral pictures. Cognitive behavioral therapy led to a significant reduction of right anterior activation for the group receiving therapy (n = 17) compared to wait-list controls (n = 18) in response to the trauma stimulus. Likewise, Wiedemann et al. (1999) conducted a study where patients with panic disorder (n = 23) compared to healthy controls (n = 25) were confronted with neutral (mushroom), panic-relevant (emergency situation), anxiety-relevant but panic-irrelevant (spider), or anxiety-irrelevant but emotionally relevant pictures (erotic image). They found a significant decrease of right compared to left frontal alpha power in response to the emergency picture category for the group with panic disorder but not for the healthy control group. Gayathiri et al. (2020) reported elevated right hemispheric activity, indicating avoidance, when individuals with major depressive disorder (n = 15) viewed images of high valence and arousal relative to neutral ones. Contrary to these studies, Ischebeck et al. (2014) did not find differences between twenty patients with obsessive compulsive disorder and twenty matched controls during viewing neutral, aversive and OCD-related images.

In Deng et al. (2021), drug abstainers' (n = 40) responses to drug-related images were compared to positive, negative, and neutral pictures in the context of evaluating the effect of a training on emotion regulation. While there was no main effect of picture type, improved alpha asymmetry scores for negative and drug-related pictures were found for the training group pre-and post-training. Cartocci et al. (2018) found higher frontal alpha asymmetry for heavy smokers compared to light smokers and non-smokers when viewing effective public service announcement (PSA) pictures. In their earlier study frontal alpha

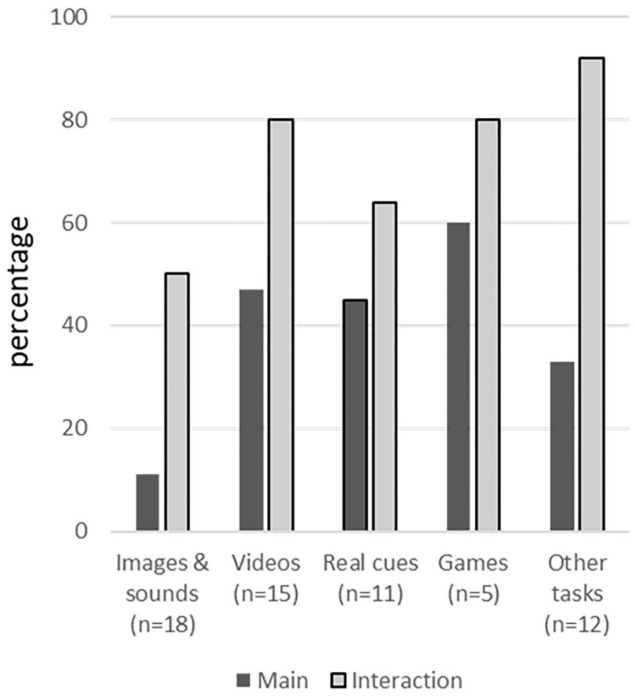


FIGURE 2 | Percentage of studies showing a significant effect of stimuli alone and in interaction with other conditions or participant subgroups, separately for each stimulus category.

asymmetry for different PSA images did not differ (Cartocci et al., 2017), which might be attributed to several differences between the studies, such as a lower number of participants in the earlier study ($n = 22$ vs. $n = 39$).

The previously described pattern of responses only in groups for whom the stimuli are relevant is likely also important for non-clinical samples. Schöne et al. (2016) asked seventeen male students to view erotic pictures of high salience as well as depictions of dressed attractive women and found significant results of picture category. Winter et al. (2016) used food images to assess the effect of hedonic hunger and restrained eating on frontal alpha asymmetry with 58 female participants. They found that higher restraint scores were associated with increased right frontal asymmetry and higher hedonic hunger was associated with increased left frontal asymmetry. Additionally, they found that overweight compared to normal weight individuals displayed greater left asymmetry. However, for the condition of fasted and fed state no differences emerged. Gable and Harmon-Jones (2008) did report for 26 female students that while dessert pictures alone did not evoke significant asymmetric activation, more time since eaten and dessert liking related to increased left frontal asymmetry for dessert pictures.

Our search resulted in two studies using auditory stimuli only. Papousek et al. (2018) ($n = 62$ students) explored inter-individual differences in frontal alpha asymmetry to other people's affect using sound recordings of three categories: anger (shouting), sadness (crying) and neutral (trivial everyday sounds) as a reference condition. Results show that individuals with higher compared to lower level of antagonism (assessed by a Personality Inventory) had less relative right frontal activation (approach) in response to the anger stimulus, whereas subjects with higher levels of detachment displayed greater relative right hemisphere activation (withdrawal) to the crying stimulus. Similarly, in Chen et al. (2015) a sample of 18 students listened to scary and soothing sounds. Subjects who showed a greater withdrawal response to scary sounds displayed a decreased pleasant state, and participants with higher approach motivation showed an increased pleasant state.

Videos ($n = 14$)

As expected, studies using video stimuli showed more often strong approach-avoidance effects than studies using unimodal stimuli (images & sounds), both as effect of the stimuli alone and in interaction with participant characteristics and conditions.

Unlike the anti-smoking image stimuli as described above in section Images and Sounds (Cartocci et al., 2017), anti-smoking video announcements induced effects in alpha asymmetry. Video announcements that had independently been classified as "awarded" induced an increased approach-avoidance effect compared to independently classified "ineffective" and "effective" ones. Another study on this topic (Cartocci et al., 2019) found that smokers showed stronger alpha-asymmetry avoidance than non-smokers in response to anti-smoking videos, highlighting again the importance of the interaction between stimuli and participant characteristics in approach-avoidance effects.

Not all studies using advertising videos have shown positive results. Joaquim et al. (2020) have reported only a trend in correlation between the asymmetry index for low alpha frequency band and negative emotions elicited by commercials viewed by 25 subjects. In a study by Hakim et al. (2021) 33 subjects watched skits from a comedy series followed by commercials of food products and later completed a choice task consisting of six products altogether. Results show that frontal alpha asymmetry as recorded during commercial viewing did not significantly differ for neither closely nor distantly ranked products. Vecchiato et al. (2014) showed six TV commercials (duration around 30 s) to 15 volunteers. Participants were divided into "LIKE" and "DISLIKE" group according to their pleasantness response rating. As expected, the "LIKE" group displayed increased left hemisphere activity compared to the "DISLIKE" group.

Walsh et al. (2017) recruited 40 students and showed them breakfast meal videos of 40 s in duration. The clips contained emotion-eliciting events with hygiene, safety, and spoilage concerns and almost identical controls without such concerns. For the spoilage videos they found greater right hemisphere activation indicating avoidance response when compared to its matched control. For the hygiene and safety videos they did not find significant differences. In another food related study, McGeown and Davis (2018) recorded the brain activity of 93 female students while watching a confederate consuming potato chips, followed by conducting a visual-probe task with non-food and food items of high craving ratings. Overweight participants (based on BMI) compared to leaner counterparts showed

increased left frontal alpha asymmetry during the confederate video and greater attentional bias toward food pictures.

Most video studies used videos to induce basic types of emotion. Zhao et al. (2018) presented three film clips (duration of around 80 s each) to 37 students to elicit tenderness, anger and neutrality. They found greater left frontal activation during the tenderness film clip. The anger eliciting film clip led to expected greater right frontal activation. Papousek et al. (2014) displayed a film with a duration of ∼10 min consisting of scenes of severely injured, mourning and dying people, to 148 female students. The expected effect of a right-sided shift of dorsolateral prefrontal asymmetry was found. In Prause et al. (2014) 65 students viewed a neutral film (10 min duration) followed by a sexual film (3 min duration). Increased alpha power was found in the left hemisphere (i.e., approach) during sexual compared to neutral films. Furthermore, self-reported mental sexual arousal and alpha asymmetry were positively correlated. In a study by Hajal et al. (2017) 26 mothers of 5- to 8-month-olds watched a 15-min video composed of 10 s clips of their own infants expressing distress. They found an association between greater right frontal asymmetry shift (from baseline to infant distress video) and higher self-reported sadness.

A large proportion of studies specifically examined the interaction between stimulus and groups of participants with certain characteristics. Hosseini et al. (2007) showed four video clips (duration of 3 min each) to induce relaxation, happiness, anxiety and sadness. Their sample consisted of 40 female students equally divided into four groups: extroverted, introverted, neurotic and emotionally stable. They found that right frontal asymmetry was associated with negative affect for the introvert and emotionally stable groups. Aftanas and Varlamov (2004) showed 10 film neutral and emotional clips each of 1.5–4.5 min duration to individuals with alexithymia ($n = 17$), a personality trait characterized by difficulties in emotional self-regulation, and non-alexithymic ($n = 27$) participants. In all cases subjects with alexithymia showed greater reactivity of the right hemisphere to the emotional clips relative to neutral, suggestive of increased avoidance motivation. Olszewska-Guizzo et al. (2021) found no significant effect for frontal alpha asymmetry for video type (nature exposure and busy public spaces), but a significant decrease of frontal alpha asymmetry as recorded following a national lockdown with a Stay-at-Home order compared to before the pandemic ($n = 22$). Missana and Grossmann (2015) studied a sample of 20 4-month-old and 20 8-month-old infants, and found that only the older infant group showed increased left-sided frontal alpha asymmetry in response to point-lighted display of happy body expressions and higher right-sided activation in response to fearful body expressions.

Real Cues ($n = 11$)

We expected that in general, real cues should produce stronger approach-avoidance effects than videos. However, the proportion of studies finding significant effects is similar.

Kaneko et al. (2019) and Lagast et al. (2020), who explored the effect of different types of drinks on frontal alpha asymmetry in, respectively, 70 and 32 participants, observed no significant effects. However, odors as researched by Kline et al. (2000) recording EEG in 58 women have led to increased relative left frontal activation for the pleasant stimulus (vanilla) when compared to unpleasant (valerian) and neutral (water).

Two neuromarketing studies utilizing real cues reported significant results for frontal alpha asymmetry. Modica et al. (2018) compared different categories of food items: daily and comfort food, major and private brands, and foreign and local products in two experiments ($n = 19$ and $n = 13$). They found increased tendency for approaching comfort compared to daily food, and foreign compared to local products during visual exploration and visual and tactile exploration phases. In addition, the private label compared to major brand also showed higher approach in the visual and tactile exploration phases. Similarly, Sargent et al. (2020) compared two machines to prepare hot beverages, one from a market leader and the other from a follower machine in an office setting ($n = 26$). It was shown that the market leader machine's user interface was preferred, indicated by self-reports and supported by significant valence measured by frontal alpha asymmetry and arousal extracted from electrodermal activity measures. Another study using a real food-related stimulus, was conducted by Brouwer et al. (2017), where 41 participants cooked and tasted two stir fry dishes. For one the main ingredient was chicken (hypothesized to induce approach) and for the other mealworms (hypothesized to evoke avoidance). The expected effect of food condition was found in frontal alpha asymmetry throughout the entire cooking and tasting session, significantly during the frying interval.

In a substance study, Knott et al. (2008), exposed 11 regular and 11 light smokers to a neutral and a cigarette-cue (holding a pen and holding a lighted cigarette above an ashtray respectively), while EEG was recorded. Results show that particularly regular female smokers exhibited withdrawal-related negative affect to holding the cigarette compared to holding the pen.

Three studies in our selection used real social interaction cues. In a study with 25 infant-parent dyades, Bolinger et al. (2020) used positive (e.g., parent played peek-a-boo with the infant) and negative prompts (e.g., parent pretended that the infant has rash on his/her face) and found significantly increased right-sided frontal alpha asymmetry (reflecting avoidance or withdrawal) for the negative prompts. No effects were observed for positive and neutral stimuli. Uusberg et al. (2015) and Pönkänen and Hietanen (2012) explored how eye-contact is related to frontal alpha asymmetry. In Uusberg et al. (2015) ($n = 40$) the degree of social contact was varied by gaze direction and as expected, neuroticism was related to stronger right-sided activation in response to direct gaze. In Pönkänen and Hietanen (2012) ($n = 22$) the expected left-sided asymmetry in response to direct gaze was not observed.

Finally, Olszewska-Guizzo et al. (2020) passively exposed 22 adults to pre-selected real landscape scenes, consisting of six park scenes and three busy urban spaces. They found a non-significant trend in the expected direction with higher approach motivation for park compared to urban spaces.

Games ($n = 5$)

Games were expected to be the most potent inducers of approach-avoidance effects. Indeed, this category seems to result

the in the largest proportion of significant results for main stimulus effects, but we should note the modest number of studies in this category ($n = 5$).

Rodrigues et al. (2018) asked 30 participants to move freely around in a virtual T-maze using a joystick. The maze contained monsters and sheep (emotionally negative and positive trials, respectively). The results aligned with the approach-avoidance model, with more left frontal alpha activation during the positive event condition and increased right frontal alpha activation in the negative condition. In Miller and Tomarken (2001), 60 participants underwent a delayed reaction time task with manipulations of the incentive, expectancy, and response. They found that variations in monetary incentives led to the expected changes in alpha asymmetry, i.e., more relative left frontal activation during reward conditions, and shifts to right frontal activation during punishment conditions. Similarly, Sobotka et al. (1992) manipulated reward and punishment in a sample of 15 students. Reward trials were associated with higher activation in the left frontal hemisphere and during punishment trials higher right-sided activation was found.

Two studies in the games category recorded from clinical samples. Shankman et al. (2007) used a slot machine game with reward and no incentive outcomes. Participants included 70 individuals with major depression and 37 controls. No differences in hemispheric asymmetry for the two outcome conditions were observed, and no overall difference between the depressed and non-depressed group. However, they found a trend between age of depression onset and hypothesized approach during reward trials. Participants with early depression onset seemed to exhibit less left frontal activity (less approach) during reward conditions compared to participants with late-onset depression and the control group. Harmon-Jones et al. (2008) explored frontal cortical responses of 41 individuals with bipolar disorders and 53 controls. For this they used anagrams of different difficulty levels (easy, medium and hard) and valence (win money or avoid losing money). They found that as expected, individuals with bipolar disorder showed greater left frontal activation in preparation for the hard-win task compared to controls. Furthermore, while non-bipolar subjects showed a decrease in left frontal activation from medium to hard win trials, those on the bipolar disorder spectrum did not.

Other Tasks ($n = 12$)

Tasks in this category entailed imagery ($n = 4$), modifying facial expression ($n = 2$) and speech, reading and writing tasks ($n = 6$). Overall, "other tasks" stimuli seemed quite potent in eliciting effects in interaction with participant group or condition, but relatively few main effects were reported.

Four studies used a variety of emotional imagery tasks, and all reported significant results. Mennella et al. (2015) measured EEG of a clinical sample of 23 dysphoric and 24 non-dysphoric individuals during pleasant, neutral and unpleasant narratives. They found reduced left relative to right activity irrespective of emotional condition in the dysphoric group compared to the control group, but no main effect of the different emotional tasks. Wacker et al. (2008) found significant approach-avoidance effects using emotional imagery scripts of three approach-avoidance conflict scenarios and a sample of 93 men with either high or low behavioral inhibition system (BIS) sensitivity. Their results showed that only the group high in trait BIS sensitivity had a significant change toward right-sided activation for the imagery compared to the pre-stimulus phase. In addition, Wacker et al. (2003) induced vivid imagery with relevant soccer scripts in a sample of 109 active, male soccer players. They found significant changes in the alpha band toward left frontal activation for the group with anger-inducing scripts and toward right frontal activation for the control and fear-withdrawal stimuli. Papousek et al. (2017) used a type of imagery task, where female university students ($n = 78$) looked at anger-eliciting vignettes supplemented by matching photographs and were instructed to imagine the depicted situation happening to them. Subsequently, they wrote down possible ways to appraise the situation to diminish anger. In a comparison task, they were asked to generate novel ideas to use a conventional, emotionally neutral object. Participants with greater capacity to generate reappraisal showed greater left-sided activity in the pre-frontal cortex. No difference was found between the two types of emotional task.

Two studies aimed to induce different emotions using facial expression tasks. Both reported significant effects. In Coan et al. (2001) students' ($n = 36$) facial configurations of anger, disgust, fear, joy and sadness matched the expected frontal activation patterns, i.e., less left frontal activity in withdrawal states compared to approach and control states. In Stewart et al. (2014) a participant group with major depressive disorder ($n = 143$) showed less left frontal activity during approach and withdrawal conditions than a control group ($n = 163$).

Six studies used speech, reading and writing tasks. Pérez-Edgar et al. (2013) presented face pairs depicting angry, happy and neutral expressions in a dot-probe paradigm, followed by speech preparation to 45 students. Relative EEG asymmetry was calculated between the speech preparation and baseline. Increased right frontal alpha activation was associated with avoidance of happy, and attentional bias toward angry faces in the dot-probe task. Brooker et al. (2016) conducted a longitudinal twin study ($n = 89$) with three emotion eliciting episodes: conversation with the experimenter, with a stranger and listening to a stranger reading a script. They found that children showed increased asymmetry scores, consistent with approach, during conversing with a stranger and experimenter compared to the stranger script episodes.

In Wang et al. (2015) 92 students were informed that they had to give a speech to elicit anxiety, and they were asked to imagine the speech scenario or think of previous embarrassing experiences. This was followed by a possible writing task depending on the group: reappraisal writing, irrelevant writing and no writing. Afterwards they were asked to re-imagine embarrassing speech scenarios. Compared to the irrelevant writing group, the reappraisal writing group had lower frontal alpha asymmetry scores during the writing manipulation period and higher "approach" frontal alpha asymmetry scores following re-exposure to stress. Li et al. (2016) also used a writing task. Participants ($n = 29$ students) were instructed to think of a situation when they were irritated by people with higher or lower social power. As expected, they found a significant association

between high social power and increased left frontal alpha asymmetry compared to the low social power condition.

In Rejer and Jankowski (2017) six subjects performed a reading task, which was interrupted by internet advertisements. This caused changes in frontal alpha asymmetry though the direction of change differed between subjects. In Brouwer et al. (2015a) 71 participants performed a reading task of a novel where emotional and non-emotional sections were pre-defined. Higher frontal alpha asymmetry was found for high compared to low emotional sections.

DISCUSSION

The aim of this review was to investigate what types of affective stimuli are effective in inducing an approach-avoidance response in frontal alpha asymmetry, in the hope that this will contribute to better understanding and application of alpha asymmetry. We reviewed findings in the affective alpha asymmetry literature following five types of commonly used stimuli that were expected to differ in their effectiveness to engage the subjects: (1) Images & sounds, (2) Videos, (3) Real cues, (4) Games and (5) Other tasks. The first three of these categories represent studies where participants' task mostly consisted of passively perceiving the stimuli, going from unimodal and less realistic, to multimodal and more realistic, where we expected this to be associated with an increasing level of affective engagement and therewith, potency to induce approach-avoidance effects. Tasks were expected to be more motivationally engaging overall, in particular games.

As expected, unimodal images and sounds appeared to be the least potent to induce clear effects—significant effects were almost only reported when the stimulus was particularly relevant for the participant group. Also as expected, studies using video stimuli showed strong approach-avoidance effects more often than studies using images and sounds, both as effect of the stimuli alone and in interaction with participant characteristics and conditions. The proportion of studies finding significant effects using real cues did not seem larger, but was approximately similar, to studies using videos. As expected, the proportion of significant results for main stimulus effects was largest for games, but we should note the modest number of studies in this category, and we conclude they are in the same order as videos and real cues. "Other tasks" stimuli seemed quite potent in eliciting effects in interaction with participant group or condition, but relatively few main effects were reported. Many studies that did not report an effect of stimulus alone reported stimulus effects in association with participant characteristics or other conditions. This makes sense in that the motivational aspect of stimuli is never completely determined by a stimulus itself, but affective approach-avoidance responses arise as an interplay between stimuli and an individual who has certain characteristics and finds him/herself in a certain situation. This aligns with ideas of Coan and colleagues and the capability model, stating that motivational tendency in an individual should be studied within a clear motivational context (Coan et al., 2006). Below, we discuss our results in more detail.

In general, viewing static images may be expected to be not very emotionally and motivationally engaging. The findings of this review revealed that picture presentation could induce approach-avoidance effects if the images were particularly emotionally relevant for the sample group, for instance anxiety-relevant pictures shown to patients with panic disorder (Wiedemann et al., 1999). Thus, for a general sample group, affective images alone might be insufficient to create motivational engagement while stimulus-relevant personal characteristics can potentiate frontal alpha asymmetry (Harmon-Jones et al., 2006; Gable and Harmon-Jones, 2008; Uusberg et al., 2014; Rejer and Jankowski, 2017). Consistent with this, significant correlations have been found between frontal alpha asymmetry and differences in emotive tendencies (e.g., dessert liking) or personality traits (Wacker et al., 2008; for examples see Gable and Harmon-Jones, 2008; Uusberg et al., 2015; Winter et al., 2016). As one of the exceptions, Schöne et al. (2016) showed that presentation of erotic pictures to male students lead to clear alpha asymmetry results, even in a brief (3 s) picture presentation task. They argue that in this case, pictures are the actual desired object themselves, and therefore create a relatively strong approach motivation in contrast to pictures that are a depiction of something that is desirable, such as food. Huster et al. (2009) aimed to improve motivational engagement for pictures by successively displaying three pictures of the same affective category, and found a main effect. Showing pictures of the same category successively also allowed for computation of frontal alpha asymmetry over a longer time period, which may have increased the robustness of the measure (Huster et al., 2009). Note that this points to another overall difference between studies that use images and other stimuli besides expected engagement—the generally short interval per stimulus that used to determine alpha asymmetry may be another factor explaining weak alpha asymmetry results for images.

From the engagement perspective, and consistent with the reasoning by Schöne et al. (2016) as mentioned above, we expected real cues to be particularly effective as they are not just a depiction of something creating a tendency to approach or avoid, but can be the genuine objects to approach or avoid. Indeed, experiments using food, odors and cigarettes found significant effects for frontal alpha asymmetry. However, those employing landscapes and tasting drinks did not. In these studies, noise caused by movement could have prevented clear results. Because body movement causes noise in EEG signals, stimuli employing movement can be expected to be less effective in producing an alpha asymmetry approach-avoidance effect. In Olszewska-Guizzo et al. (2020) participants went from one scene to the other, leading to long time intervals between recordings and hence noisy comparisons between conditions. Furthermore, in Kaneko et al. (2019) participants took sips from cups themselves, which led to noise through movement. On the other hand, Lagast et al. (2020) minimized such movements by using plastic tubes but were still not able to find a significant approach-avoidance effect. Also, results of studies in other stimulus categories did not suggest that in general, modest amounts of movement prohibit finding alpha asymmetry effects.

Out of scope for the current review that focussed on the role of affective stimuli, but also relevant for the approach-avoidance alpha asymmetry effect are data recording, processing, and analysis (for an extensive review, see Smith et al., 2017). A few essential points of consideration are the EEG recording length (Towers and Allen, 2009), selection of the electrode reference (Hagemann and Naumann, 2001; Hagemann, 2004; Stewart et al., 2010) and the reliability of the EEG measurement (Hagemann et al., 2002; Allen et al., 2004a,b). With novel wearable EEG monitoring devices and processing techniques, recordings in less controlled environments are becoming more reliable (e.g., see Aricò et al., 2018; Pion-Tonachini et al., 2019), but controlled experiments and lab-grade equipment will have some advantage on signal quality. Furthermore, aspects of the design besides choice of stimulus such as the number and duration of trials and baselines, analysis (e.g., exact definition of the alpha band and methods for artifact removal) are not standardized and can lead to big differences.

This brings us to the limitations of this literature review. One is that experiments are very diverse and thus difficult to compare. We focused on the overall effect of affective stimulus category. For almost every stimulus category, studies were identified that reported no effect of stimuli on alpha asymmetry at all; but glancing through these studies did not bring to light one obvious factor underlying these null results.

Second, even though keywords were clear, it was noted that not all relevant papers were captured through the search. We do not claim that we here provide an exhaustive overview, and our results should be taken as indicative. Still, we believe that the inclusion of 61 papers results in a representative review of the literature.

Thirdly, we should note that while our choice of stimulus categories was not arbitrary, other choices and definitions of stimulus categories would have been possible as well and could have influenced the conclusions. Also, our categories were not exactly exclusive and sometimes overlapping, e.g., the cooking and tasting experiment by Brouwer et al. (2017) could be arguably belonging to tasks rather than real cues. In such cases, the stimulus' affective content led to the final categorization decision. We hope that our summarizing **Supplementary Table** facilitates potential follow-up research, viewing the results from possible other perspectives.

Furthermore, most of the papers reviewed here reported significant alpha asymmetry approach-avoidance results, or trends in that direction. Papers that reported null findings possibly did not include the keywords used in our search. An example is Walden et al. (2015), where frontal theta activity was studied as a function of approach-avoidance affective autobiographical memory recall. They mention in a footnote that no effect on alpha-asymmetry was observed. In addition, many of such findings were probably withheld from publication in the first place, commonly known as publication bias. Not reporting null-findings is a general problem that could lead to another research group investigating the same line of thought, leading to null findings again, ultimately wasting resources, distorting literature and damaging the integrity of knowledge (Joober et al., 2012). Furthermore, negative outcomes are valuable for science since they force critical reflection, validation of current thinking and direct new approaches (Matosin et al., 2014). Therefore, researchers should be more encouraged and journals more open to publish manuscripts reporting negative results. Taking into account the likely underreporting of null findings, and the finding that roughly 50% of studies reporting a solid effect of stimulus only for four of the five categories, where this percentage was even considerably lower for the images & sounds category, we can conclude that alpha asymmetry approach-avoidance is not an easy to find phenomenon, especially not when tested in general populations without further manipulation of context to increase stimulus relevance.

Despite of the aforementioned limitations, the exploration of frontal alpha asymmetry as an indicator of affective approach-avoidance can benefit marketing, human-computer interfaces and the diagnosis of affective disorders. Frontal alpha asymmetry may provide a more objective and continuous measure of mental state than traditional methods that are influenced by social factors and may affect the mental state itself. This review confirmed that overall, strongly engaging, salient and/or personally relevant stimuli are important to induce an approach-avoidance effect and that the selection of stimuli accounts for part of the diversity in alpha asymmetry research. More work is required to gain a better understanding of other factors influencing frontal alpha asymmetry as a marker of emotion.

AUTHOR CONTRIBUTIONS

PS, IS, DK, and A-MB: conceptualization. PS: literature search, summarizing literature, and writing first draft. All authors contributed to revising the article and approved the submitted version.

FUNDING

This research was funded by the Kikkoman Europe R&D Laboratory B.V. The authors declare that this study received funding from Kikkoman Europe R&D Laboratory B.V. Other than that one of the co-authors (Daisuke Kaneko) was employed by Kikkoman Europe R&D Laboratory B.V., the funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

ACKNOWLEDGMENTS

We gratefully acknowledge comments of Dimitra Dodou on a previous version of this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2022.869123/full#supplementary-material

REFERENCES

Adolph, D., von Glischinski, M., Wannemüller, A., and Margraf, J. (2017). The influence of frontal alpha-asymmetry on the processing of approach- and withdrawal-related stimuli—A multichannel psychophysiology study. *Psychophysiology* 54, 1295–1310. doi: 10.1111/psyp.12878

Aftanas, L., and Varlamov, A. (2004). Associations of alexithymia with anterior and posterior activation asymmetries during evoked emotions: EEG evidence of right hemisphere "electrocortical effort." *Int. J. Neurosci.* 114, 1443–1462. doi: 10.1080/00207450490509230

Ahern, G. L., and Schwartz, G. E. (1985). Differential lateralization for positive and negative emotion in the human brain: EEG spectral analysis. *Neuropsychologia* 23, 745–755. doi: 10.1016/0028-3932(85)90081-8

Alford, L. B. (1933). Localization of consciousness and emotion. *Am. J. Psychiatry* 89, 789–799. doi: 10.1176/ajp.89.4.789

Allen, J. J. B., Coan, J. A., and Nazarian, M. (2004a). Issues and assumptions on the road from raw signals to metrics of frontal EEG asymmetry in emotion. *Biol. Psychol.* 67, 183–218. doi: 10.1016/j.biopsycho.2004.03.007

Allen, J. J. B., Urry, H. L., Hitt, S. K., and Coan, J. A. (2004b). The stability of resting frontal electroencephalographic asymmetry in depression. *Psychophysiology* 41, 269–280. doi: 10.1111/j.1469-8986.2003.00149.x

Alves, N. T., Fukusima, S. S., and Aznar-Casanova, J. A. (2008). Models of brain asymmetry in emotional processing. *Psychol. Neurosci.* 1, 63–66. doi: 10.3922/j.psns.2008.1.010

Aricò, P., Borghini, G., Di Flumeri, G., Sciaraffa, N., and Babiloni, F. (2018). Passive BCI beyond the lab: current trends and future directions. *Physiol. Meas.* 39, 08TR02. doi: 10.1088/1361-6579/aad57e

Babiloni, F., Rossi, D., Cherubino, P., Trettel, A., Picconi, D., Maglione, A. G., et al. (2015). "The first impression is what matters: a neuroaesthetic study of the cerebral perception and appreciation of paintings by Titian," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. (Titian: Institute of Electrical and Electronics Engineers Inc.), 7990–7993.

Baldwin, C. L., and Penaranda, B. N. (2012). Adaptive training using an artificial neural network and EEG metrics for within- and cross-task workload classification. *Neuroimage* 59, 48–56. doi: 10.1016/j.neuroimage.2011.07.047

Bartolomé-Tomás, A., Sánchez-Reolid, R., Fernández-Sotos, A., Latorre, J. M., and Fernández-Caballero, A. (2020). Arousal detection in elderly people from electrodermal activity using musical stimuli. *Sensors* 20, 4788. doi: 10.3390/s20174788

Berkman, E. T., and Lieberman, M. D. (2010). Approaching the bad and avoiding the good: Lateral prefrontal cortical asymmetry distinguishes between action and valence. *J. Cogn. Neurosci.* 22, 1970–1979. doi: 10.1162/jocn.2009.21317

Bolinger, E., Ngo, H. V., Kock, V., Wassen, D. T., Matuz, T., et al. (2020). Affective cortical asymmetry at the early developmental emergence of emotional expression. *eNeuro* 7, 1–10. doi: 10.1523/ENEURO.0042-20.2020

Briesemeister, B. B., Tamm, S., Heine, A., and Jacobs, A. M. (2013). Approach the good, withdraw from the bad—A review on frontal alpha asymmetry measures in applied psychological research. *Psychology* 4, 261–267. doi: 10.4236/psych.2013.43A039

Brooker, R. J., Davidson, R. J., and Goldsmith, H. H. (2016). Maternal negative affect during infancy is linked to disrupted patterns of diurnal cortisol and alpha asymmetry across contexts during childhood. *J. Exp. Child Psychol.* 142, 274–290. doi: 10.1016/j.jecp.2015.08.011

Brouwer, A. M., Hogervorst, M., Reuderink, B., van der Werf, Y., and van Erp, J. (2015a). Physiological signals distinguish between reading emotional and non-emotional sections in a novel. *Brain Comp. Interf.* 2, 76–89. doi: 10.1080/2326263X.2015.1100037

Brouwer, A. M., Hogervorst, M. A., Grootjen, M., van Erp, J. B. F., and Zandstra, E. H. (2017). Neurophysiological responses during cooking food associated with different emotions. *Food Qual. Prefer.* 62, 307–316. doi: 10.1016/j.foodqual.2017.03.005

Brouwer, A. M., Zander, T. O., van Erp, J. B. F., Korteling, J. E., and Bronkhorst, A. W. (2015b). Using neurophysiological signals that reflect cognitive or affective state: six recommendations to avoid common pitfalls. *Front. Neurosci.* 9, 136. doi: 10.3389/fnins.2015.00136

Cartocci, G., Caratù, M., Modica, E., Maglione, A. G., Rossi, D., Cherubino, P., et al. (2017). Electroencephalographic, heart rate, and galvanic skin response assessment for an advertising perception study: application to antismoking public service announcements. *J. Vis. Exp.* 126, 55872. doi: 10.3791/55872

Cartocci, G., Modica, E., Rossi, D., Cherubino, P., Maglione, A. G., Colosimo, A., et al. (2018). neurophysiological measures of the perception of antismoking public service announcements among young population. *Front. Hum. Neurosci.* 12, 231. doi: 10.3389/fnhum.2018.00231

Cartocci, G., Modica, E., Rossi, D., Inguscio, B., Arico, P., Levy, A. C. M., et al. (2019). Antismoking campaigns? Perception and gender differences: a comparison among EEG indices. *Comput. Intell. Neurosci.* 2019, 7348795. doi: 10.1155/2019/7348795

Cartocci, G., Rossi, D., Modica, E., Maglione, A. G., Martinez Levy, A. C., Cherubino, P., et al. (2021). Neurodante: poetry mentally engages more experts but moves more non-experts, and for both the cerebral approach tendency goes hand in hand with the cerebral effort. *Brain Sci.* 11, 1–25. doi: 10.3390/brainsci11030281

Carver, C. S., and Harmon-Jones, E. (2009). Anger is an approach-related affect: evidence and implications. *Psychol. Bull.* 135, 183–204. doi: 10.1037/a0013965

Chen, X., Takahashi, I., Okita, Y., Hirata, H., and Sugiura, T. (2015). Psychological response to sound stimuli evaluated by EEG: joint consideration of AAE model and comfort vector model. *J. Psychophysiol.* 29, 112–118. doi: 10.1027/0269-8803/a000142

Christopoulos, G. I., Uy, M. A., and Yap, W. J. (2019). The body and the brain: measuring skin conductance responses to understand the emotional experience. *Organ. Res. Methods* 22, 394–420. doi: 10.1177/1094428116681073

Coan, J. A., and Allen, J. J. B. (2003). Frontal EEG asymmetry and the behavioral activation and inhibition systems. *Psychophysiology* 40, 106–114. doi: 10.1111/1469-8986.00011

Coan, J. A., and Allen, J. J. B. (2004). Frontal EEG asymmetry as a moderator and mediator of emotion. *Biol. Psychol.* 67, 7–50. doi: 10.1016/j.biopsycho.2004.03.002

Coan, J. A., Allen, J. J. B., and Harmon-Jones, E. (2001). Voluntary facial expression and hemispheric asymmetry over the frontal cortex. *Psychophysiology* 38, 912–925. doi: 10.1111/1469-8986.3860912

Coan, J. A., Allen, J. J. B., and McKnight, P. E. (2006). A capability model of individual differences in frontal EEG asymmetry. *Biol. Psychol.* 72, 198–207. doi: 10.1016/j.biopsycho.2005.10.003

Cook, I. A., O'Hara, R., Uijtdehaage, S. H. J., Mandelkern, M., and Leuchter, A. F. (1998). Assessing the accuracy of topographic EEG mapping for determining local brain function. *Electroencephalogr. Clin. Neurophysiol.* 107, 408–414. doi: 10.1016/S0013-4694(98)00092-3

Crabbe, J. B., Smith, J. C., and Dishman, R. K. (2007). Emotional & electroencephalographic responses during affective picture viewing after exercise. *Physiol. Behav.* 90, 394–404. doi: 10.1016/j.physbeh.2006.10.001

Daly, I., Williams, D., Hwang, F., Kirke, A., Miranda, E. R., and Nasuto, S. J. (2019). Electroencephalography reflects the activity of sub-cortical brain regions during approach-withdrawal behaviour while listening to music. *Sci. Rep.* 9, 1–22. doi: 10.1038/s41598-019-45105-2

Davidson, R. J. (1984). "Affect, cognition, and hemispheric specialization," in *Emotion, Cognition, and Behavior*, eds. C. R. Izard, J. Kagan, and R. B. Zajonc (New York, NY: Cambridge University Press), 320–365.

Davidson, R. J. (2004). What does the prefrontal cortex "do" in affect: perspectives on frontal EEG asymmetry research. *Biol. Psychol.* 67, 219–234. doi: 10.1016/j.biopsycho.2004.03.008

Davidson, R. J., Ekman, P., Saron, C. D., Senulis, J. A., and Friesen, W. V. (1990). Approach-withdrawal and cerebral asymmetry: emotional expression and brain physiology I. *J. Pers. Soc. Psychol.* 58, 330–341. doi: 10.1037/0022-3514.58.2.330

Davidson, R. J., and Irwin, W. (1999). The functional neuroanatomy of emotion and affective style. *Trends Cogn. Sci.* 3, 11–21. doi: 10.1016/S1364-6613(98)01265-0

Davidson, R. J., Schaffer, C. E., and Saron, C. (1985). Effects of lateralized presentations of faces on self-reports of emotion and EEG asymmetry in depressed and non-depressed subjects. *Psychophysiology* 22, 353–364. doi: 10.1111/j.1469-8986.1985.tb01615.x

Dell, N., Vaidyanathan, V., Medhi, I., Cutrell, E., and Thies, W. (2012). ""yours is better!" Participant response bias in HCI," in *Conference on Human Factors in Computing Systems - Proceedings*, eds J. A. Konstan, E. H. Chi, and K. Hook (New York, NY, USA: ACM), 1321–1330.

Deng, Y., Hou, L., Chen, X., and Zhou, R. (2021). Working memory training improves emotion regulation in drug abstainers: evidence from frontal alpha asymmetry. *Neurosci. Lett.* 742, 135513. doi: 10.1016/j.neulet.2020.135513

Denny-brown, D., Meyer, J. S., and Horenstein, S. (1952). The significance of perceptual rivalry resulting from parietal lesion. *Brain* 75, 432–471. doi: 10.1093/brain/75.4.432

Diaz, A., and Bell, M. A. (2012). Frontal EEG asymmetry and fear reactivity in different contexts at 10 months. *Dev. Psychobiol.* 54, 536–545. doi: 10.1002/dev.20612

Gable, P., and Harmon-Jones, E. (2008). Relative left frontal activation to appetitive stimuli: considering the role of individual differences. *Psychophysiology* 45, 275–278. doi: 10.1111/j.1469-8986.2007.00627.x

Gable, P. A., and Poole, B. D. (2014). Influence of trait behavioral inhibition and behavioral approach motivation systems on the LPP and frontal asymmetry to anger pictures. *Soc. Cogn. Affect. Neurosci.* 9, 182–190. doi: 10.1093/scan/nss130

Gabriel, D., Merat, E., Jeudy, A., Cambos, S., Chabin, T., Giustiniani, J., et al. (2021). Emotional effects induced by the application of a cosmetic product: a real-time electrophysiological evaluation. *Appl. Sci.* 11, 4766. doi: 10.3390/app11114766

Gawronski, B., and de Houwer, J. (2014). "Implicit measures in social and personality psychology," in *Handbook of Research Methods in Social and Personality Psychology*, eds H. Reis and C. Judd (Cambridge: Cambridge University Press), 283–310.

Gayathiri, R. R., Bhuvana Devi, M., Kavya, G. A., Veezhinathan, M., and Geethanjali, B. (2020). "EEG based visualization and analysis of emotional processing in major depressive disorder," in *2020 6th International Conference on Advanced Computing and Communication Systems, ICACCS 2020*. (Institute of Electrical and Electronics Engineers Inc.), 336–341.

Goldstein, K. (2004). *The Organism: A Holistic Approach to Biology Derived From Pathological Data in Man*. American Book Publishing

Grassini, S., Sikka, P., Revonsuo, A., and Koivisto, M. (2020). Subjective ratings of fear are associated with frontal late positive potential asymmetry, but not with early brain activity over the occipital and centro-parietal cortices. Psychophysiology. 57, e13665. doi: 10.1111/psyp.13665

Hagemann, D. (2004). Individual differences in anterior EEG asymmetry: methodological problems and solutions. *Biol. Psychol.* 67, 157–182. doi: 10.1016/j.biopsycho.2004.03.006

Hagemann, D., and Naumann, E. (2001). The effects of ocular artifacts on (lateralized) broadband power in the EEG. *Clin. Neurophysiol.* 112, 215–231. doi: 10.1016/S1388-2457(00)00541-1

Hagemann, D., Thayer, J. F., Naumann, E., and Bartussek, D. (2002). Does resting electroencephalograph asymmetry reflect a trait? An application of latent state-trait theory. *J. Person. Soc. Psychol.* 82, 619–641. doi: 10.1037/0022-3514.82.4.619

Hajal, N. J., Cole, P. M., and Teti, D. M. (2017). Maternal responses to infant distress: linkages between specific emotions and neurophysiological processes. *Parenting* 17, 200–224. doi: 10.1080/15295192.2017.1336001

Hakim, A., Klorfeld, S., Sela, T., Friedman, D., Shabat-Simon, M., and Levy, D. J. (2021). Machines learn neuromarketing: improving preference prediction from self-reports using multiple EEG measures and machine learning. *Int. J. Res. Market.* 38, 770–791. doi: 10.1016/j.ijresmar.2020.10.005

Harmon-Jones, E., Abramson, L. Y., Nusslock, R., Sigelman, J. D., Urosevic, S., Turonie, L. D., et al. (2008). Effect of bipolar disorder on left frontal cortical responses to goals differing in valence and task difficulty. *Biol. Psychiatry* 63, 693–698. doi: 10.1016/j.biopsych.2007.08.004

Harmon-Jones, E., Gable, P. A., and Peterson, C. K. (2010). The role of asymmetric frontal cortical activity in emotion-related phenomena: a review and update. *Biol. Psychol.* 84, 451–462. doi: 10.1016/j.biopsycho.2009.08.010

Harmon-Jones, E., Lueck, L., Fearn, M., and Harmon-Jones, C. (2006). The effect of personal relevance and approach-related action expectation on relative left frontal cortical activity. *Psychol. Sci.* 17, 434–440. doi: 10.1111/j.1467-9280.2006.01724.x

Hosseini, S. M., Fallah, P. A., Tabatabaei, S. K. R., Ladani, S. H. G., and Heise, C. (2007). Brain activity, personality traits and affect: electrocortical activity in reaction to affective film stimuli. *J. Appl. Sci.* 7, 3743–3749. doi: 10.3923/jas.2007.3743.3749

Huster, R. J., Stevens, S., Gerlach, A. L., and Rist, F. (2009). A spectralanalytic approach to emotional responses evoked through picture presentation. *Int. J. Psychophysiol.* 72, 212–216. doi: 10.1016/j.ijpsycho.2008.12.009

Inguscio, B. M. S., Cartocci, G., Modica, E., Rossi, D., Martinez-Levy, A. C., Cherubino, P., et al. (2021). Smoke signals: a study of the neurophysiological reaction of smokers and non-smokers to smoking cues inserted into antismoking public service announcements. *Int. J. Psychophysiol.* 167, 22–29. doi: 10.1016/j.ijpsycho.2021.06.010

Ischebeck, M., Endrass, T., Simon, D., and Kathmann, N. (2014). Altered frontal EEG asymmetry in obsessive-compulsive disorder. *Psychophysiology* 51, 596–601. doi: 10.1111/psyp.12214

Joaquim, M. S., Maçorano, R., Canais, F., Ramos, R., Fred, A. L., Torrado, M., et al. (2020). "Learning data representation and emotion assessment from physiological data," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. (Institute of Electrical and Electronics Engineers Inc.), 3452–3456.

Joober, R., Schmitz, N., Annable, L., and Boksa, P. (2012). Publication bias: What are the challenges and can they be overcome? *J. Psychiatry Neurosci.* 37, 149–152. doi: 10.1503/jpn.120065

Kaneko, D., Hogervorst, M., Toet, A., van Erp, J. B. F., Kallen, V., and Brouwer, A. M. (2019). Explicit and implicit responses to tasting drinks associated with different tasting experiences. *Sensors (Basel).* 19, 4397. doi: 10.3390/s19204397

Kaneko, D., Toet, A., Brouwer, A. M., Kallen, V., and van Erp, J. B. F. (2018). Methods for evaluating emotions evoked by food experiences: a literature review. *Front. Psychol.* 9, 911. doi: 10.3389/fpsyg.2018.00911

Kline, J. P., Blackhart, G. C., Woodward, K. M., Williams, S. R., and Schwartz, G. E. R. (2000). Anterior electroencephalographic asymmetry changes in elderly women in response to a pleasant and an unpleasant odor. *Biol. Psychol.* 52, 241–250. doi: 10.1016/S0301-0511(99)00046-0

Knott, V. J., Naccache, L., Cyr, E., Fisher, D. J., McIntosh, J. F., Millar, A. M., et al. (2008). Craving-induced EEG reactivity in smokers: effects of mood induction, nicotine dependence and gender. *Neuropsychobiology* 58, 187–199. doi: 10.1159/000201716

Lagast, S., de Steur, H., Gadeyne, S., Hödl, S., Staljanssens, W., Vonck, K., et al. (2020). Heart rate, electrodermal responses and frontal alpha asymmetry to accepted and non-accepted solutions and drinks. *Food Qual. Prefer.* 82, 103893. doi: 10.1016/j.foodqual.2020.103893

Lagast, S., Gellynck, X., Schouteten, J. J., de Herdt, V., and de Steur, H. (2017). Consumers' emotions elicited by food: a systematic review of explicit and implicit methods. *Trends Food Sci. Technol.* 69, 172–189. doi: 10.1016/j.tifs.2017.09.006

Lee, I. E., Latchoumane, C. -F. V., and Jeong, J. (2017). Arousal rules: An empirical investigation into the aesthetic experience of cross-modal perception with emotional visual music. *Front. Psychol.* 8, 440. doi: 10.3389/fpsyg.2017.00440

Li, D., Wang, C., Yin, Q., Mao, M., Zhu, C., and Huang, Y. (2016). Frontal cortical asymmetry may partially mediate the influence of social power on anger expression. *Front. Psychol.* 7, 73. doi: 10.3389/fpsyg.2016.00073

Lottridge, D., Chignell, M., and Yasumura, M. (2012). Identifying emotion through implicit and explicit measures: Cultural differences, cognitive load, and immersion. *IEEE Transact. Affect. Comp.* 3, 199–210. doi: 10.1109/T-AFFC.2011.36

Matosin, N., Frank, E., Engel, M., Lum, J. S., and Newell, K. A. (2014). Negativity towards negative results: a discussion of the disconnect between scientific worth and scientific culture. *Dis. Models Mech.* 7, 171–173. doi: 10.1242/dmm.015123

Mauss, I. B., and Robinson, M. D. (2009). Measures of emotion: a review. *Cogn. Emot.* 23, 209–237. doi: 10.1080/02699930802204677

McGeown, L., and Davis, R. (2018). Frontal EEG asymmetry moderates the association between attentional bias towards food and body mass index. *Biol. Psychol.* 136, 151–160. doi: 10.1016/j.biopsycho.2018.06.001

Mennella, R., Messerotti Benvenuti, S., Buodo, G., and Palomba, D. (2015). Emotional modulation of alpha asymmetry in dysphoria: results from an emotional imagery task. *Int. J. Psychophysiol.* 97, 113–119. doi: 10.1016/j.ijpsycho.2015.05.013

Miller, A., and Tomarken, A. J. (2001). Task-dependent changes in frontal brain asymmetry: effects of incentive cues, outcome expectancies, and motor responses. *Psychophysiology* 38, 500–511. doi: 10.1111/1469-8986.3830500

Missana, M., and Grossmann, T. (2015). Infants' emerging sensitivity to emotional body expressions: Insights from asymmetrical frontal brain activity. *Dev. Psychol.* 51, 151–160. doi: 10.1037/a0038469

Modica, E., Cartocci, G., Rossi, D., Levy, A. C. M., Cherubino, P., Magilone, A. G., et al. (2018). Neurophysiological responses to different product experiences. *Comput. Intell. Neurosci.* 2018, 1–10. doi: 10.1155/2018/9616301

Olszewska-Guizzo, A., Fogel, A., Escoffier, N., and Ho, R. (2021). Effects of COVID-19-related stay-at-home order on neuropsychophysiological response to urban spaces: beneficial role of exposure to nature? *J. Environ. Psychol.* 75, 101590. doi: 10.1016/j.jenvp.2021.101590

Olszewska-Guizzo, A., Sia, A., Fogel, A., and Ho, R. (2020). Can exposure to certain urban green spaces trigger frontal alpha asymmetry in the brain?—Preliminary findings from a passive task EEG study. *Int. J. Environ. Res. Public Health* 17, 394. doi: 10.3390/ijerph17020394

Papousek, I., Aydin, N., Rominger, C., Feyaerts, K., Schmid-Zalaudek, K., Lackner, H. K., et al. (2018). DSM-5 personality trait domains and withdrawal versus approach motivational tendencies in response to the perception of other people's desperation and angry aggression. *Biol. Psychol.* 132, 106–115. doi: 10.1016/j.biopsycho.2017.11.010

Papousek, I., Weiss, E. M., Perchtold, C. M., Weber, H., de Assunção, V. L., Schulter, G., et al. (2017). The capacity for generating cognitive reappraisals is reflected in asymmetric activation of frontal brain regions. *Brain Imaging Behav.* 11, 577–590. doi: 10.1007/s11682-016-9537-2

Papousek, I., Weiss, E. M., Schulter, G., Fink, A., Reiser, E. M., and Lackner, H. K. (2014). Prefrontal EEG alpha asymmetry changes while observing disaster happening to other people: cardiac correlates and prediction of emotional impact. *Biol. Psychol.* 103, 184–194. doi: 10.1016/j.biopsycho.2014.09.001

Pérez-Edgar, K., Kujawa, A., Nelson, S. K., Cole, C., and Zapp, D. J. (2013). The relation between electroencephalogram asymmetry and attention biases to threat at baseline and under stress. *Brain Cogn.* 82, 337–343. doi: 10.1016/j.bandc.2013.05.009

Pion-Tonachini, L., Kreutz-Delgado, K., and Makeig, S. (2019). ICLabel: an automated electroencephalographic independent component classifier, dataset, and website. *Neuroimage* 198, 181–197. doi: 10.1016/j.neuroimage.2019.05.026

Pönkänen, L. M., and Hietanen, J. K. (2012). Eye contact with neutral and smiling faces: effects on autonomic responses and frontal EEG asymmetry. *Front. Hum. Neurosci.* 6, 122. doi: 10.3389/fnhum.2012.00122

Poole, B. D., and Gable, P. A. (2014). Affective motivational direction drives asymmetric frontal hemisphere activation. *Exp. Brain Res.* 232, 2121–2130. doi: 10.1007/s00221-014-3902-4

Prause, N., Staley, C., and Roberts, V. (2014). Frontal alpha asymmetry and sexually motivated states. *Psychophysiology* 51, 226–235. doi: 10.1111/psyp.12173

Rabe, S., Zoellner, T., Beauducel, A., Maercker, A., and Karl, A. (2008). Changes in brain electrical activity after cognitive behavioral therapy for posttraumatic stress disorder in patients injured in motor vehicle accidents. *Psychosom. Med.* 70, 13–19. doi: 10.1097/PSY.0b013e31815aa325

Rejer, I., and Jankowski, J. (2017). Brain activity patterns induced by interrupting the cognitive processes with online advertising. *Cogn. Process.* 18, 419–430. doi: 10.1007/s10339-017-0815-8

Reuderink, B., Mühl, C., and Poel, M. (2013). Valence, arousal and dominance in the EEG during game play. *Int. J. Auton. Adapt. Commun. Syst.* 6, 45–62. doi: 10.1504/IJAACS.2013.050691

Rodrigues, J., Müller, M., Mühlberger, A., and Hewig, J. (2018). Mind the movement: frontal asymmetry stands for behavioral motivation, bilateral frontal activation for behavior. *Psychophysiology* 55. doi: 10.1111/psyp.12908

Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.* 39, 1161–1178. doi: 10.1037/h0077714

Rutherford, H. J. V., and Lindell, A. K. (2011). Thriving and surviving: approach and avoidance motivation and lateralization. *Emot. Rev.* 3, 333–343. doi: 10.1177/1754073911402392

Sargent, A., Watson, J., Ye, H., Suri, R., and Ayaz, H. (2020). Neuroergonomic assessment of hot beverage preparation and consumption: an EEG and EDA study. *Front. Hum. Neurosci.* 14, 175. doi: 10.3389/fnhum.2020.00175

Scherer, K. R., and Ekman, P. (1984). *Approaches To Emotion*. New York, NY: Psychology Press.

Schöne, B., Schomberg, J., Gruber, T., and Quirin, M. (2016). Event-related frontal alpha asymmetries: electrophysiological correlates of approach motivation. *Exp. Brain Res.* 234, 559–567. doi: 10.1007/s00221-015-4483-6

Shankman, S. A., Klein, D. N., Tenke, C. E., and Bruder, G. E. (2007). Reward sensitivity in depression: a biobehavioral study. *J. Abnorm. Psychol.* 116, 95–104. doi: 10.1037/0021-843X.116.1.95

Smith, E. E., Reznik, S. J., Stewart, J. L., and Allen, J. J. B. (2017). Assessing and conceptualizing frontal EEG asymmetry: an updated primer on recording, processing, analyzing, and interpreting frontal alpha asymmetry. *Int. J. Psychophysiol.* 111, 98–114. doi: 10.1016/j.ijpsycho.2016.11.005

Sobotka, S. S., Davidson, R. J., and Senulis, J. A. (1992). Anterior brain electrical asymmetries in response to reward and punishment. *Electroencephalogr. Clin. Neurophysiol.* 83, 236–247. doi: 10.1016/0013-4694(92)90117-Z

Songsamoe, S., Saengwong-ngam, R., Koomhin, P., and Matan, N. (2019). Understanding consumer physiological and emotional responses to food products using electroencephalography (EEG). *Trends Food Sci. Technol.* 93, 167–173. doi: 10.1016/j.tifs.2019.09.018

Stewart, J. L., Bismark, A. W., Towers, D. N., Coan, J. A., and Allen, J. J. B. (2010). Resting frontal EEG asymmetry as an endophenotype for depression risk: sex-specific patterns of frontal brain asymmetry. *J. Abnorm. Psychol.* 119, 502–512. doi: 10.1037/a0019196

Stewart, J. L., Coan, J. A., Towers, D. N., and Allen, J. J. B. (2014). Resting and task-elicited prefrontal EEG alpha asymmetry in depression: support for the capability model. *Psychophysiology* 51, 446–455. doi: 10.1111/psyp.12191

Towers, D. N., and Allen, J. J. B. (2009). A better estimate of the internal consistency reliability of frontal EEG asymmetry scores. *Psychophysiology* 46, 132–142. doi: 10.1111/j.1469-8986.2008.00759.x

Tucker, M., Stenslie, C. E., Roth, R. S., and Shearer, S. L. (1981). Right frontal lobe activation and right hemisphere performance: decrement during a depressed mood. *Arch. Gen. Psychiatry* 38, 169–174. doi: 10.1001/archpsyc.1981.01780270055007

Uusberg, A., Uibo, H., Tiimus, R., Sarapuu, H., Kreegipuu, K., and Allik, J. (2014). Approach-avoidance activation without anterior asymmetry. *Front. Psychol.* 5, 192. doi: 10.3389/fpsyg.2014.00192

Uusberg, H., Allik, J., and Hietanen, J. K. (2015). Eye contact reveals a relationship between neuroticism and anterior EEG asymmetry. *Neuropsychologia* 73, 161–168. doi: 10.1016/j.neuropsychologia.2015.05.008

Vecchiato, G., Cherubino, P., Maglione, A. G., Ezquierro, M. T. H., Marinozzi, F., Bini, F., et al. (2014). How to measure cerebral correlates of emotions in marketing relevant tasks. *Cognit. Comput.* 6, 856–871. doi: 10.1007/s12559-014-9304-x

Wacker, J., Chavanon, M. L., Leue, A., and Stemmler, G. (2008). Is running away right? The behavioral activation-behavioral inhibition model of anterior asymmetry. *Emotion* 8, 232–249. doi: 10.1037/1528-3542.8.2.232

Wacker, J., Heldmann, M., and Stemmler, G. (2003). Separating emotion and motivational direction in fear and anger: effects on frontal asymmetry. *Emotion* 3, 167–193. doi: 10.1037/1528-3542.3.2.167

Walden, K., Pornpattananangkul, N., Curlee, A., McAdams, D. P., and Nusslock, R. (2015). Posterior versus frontal theta activity indexes approach motivation during affective autobiographical memories. *Cogn. Affect. Behav. Neurosci.* 15, 132–144. doi: 10.3758/s13415-014-0322-7

Walsh, A. M., Duncan, S. E., Bell, M. A., O'Keefe, S. F., and Gallagher, D. L. (2017). Integrating implicit and explicit emotional assessment of food quality and safety concerns. *Food Qual. Prefer.* 56, 212–224. doi: 10.1016/j.foodqual.2016.11.002

Wang, F., Wang, C., Yin, Q., Wang, K., Li, D., Mao, M., et al. (2015). Reappraisal writing relieves social anxiety and may be accompanied by changes in frontal alpha asymmetry. *Front. Psychol.* 6, 1604. doi: 10.3389/fpsyg.2015.01604

Wiedemann, G., Pauli, P., Dengler, W., Lutzenberger, W., Birbaumer, N., and Buchkremer, G. (1999). Frontal brain asymmetry as a biological substrate of emotions in patients with panic disorders. *Arch. Gen. Psychiatry* 56, 78–84. doi: 10.1001/archpsyc.56.1.78

Winter, S. R., Feig, E. H., Kounios, J., Erickson, B., Berkowitz, S., and Lowe, M. R. (2016). The relation of hedonic hunger and restrained eating to lateralized frontal activation. *Physiol. Behav.* 163, 64–69. doi: 10.1016/j.physbeh.2016.04.050

Zhao, G., Zhang, Y., Ge, Y., Zheng, Y., Sun, X., and Zhang, K. (2018). Asymmetric hemisphere activation in tenderness: evidence from EEG signals. *Sci. Rep.* 8, 1–9. doi: 10.1038/s41598-018-26133-w

Conflict of Interest: DK was employed by Kikkoman Europe R&D Laboratory B.V.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership