PhD-FSTM-2026-009

Faculty of Science, Technology and Medicine

DISSERTATION

Defence held on 05/02/2026 in Esch-sur-Alzette
to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG EN INFORMATIQUE

by

Kayhan LATIFZADEH
Born on 12th January 1994 in Rasht (Iran)

Decoding Cognitive States from Neurophysiological and Behavioral Signals

Dissertation defence committee
Dr. Luis A. LEIVA, Supervisor
Professor, University of Luxembourg, Luxembourg

Dr. Martin THEOBALD, Chairman
Professor, University of Luxembourg, Luxembourg

Dr. Tomas WARD, Member
Professor, Dublin City University, Ireland

Dr. Andreas BULLING, Member
Professor, University of Stuttgart, Germany

Dr. Jean BOTEV, Member
Research scientist, University of Luxembourg, Luxembourg

A PhD Dissertation by

Kayhan LATIFZADEH

Submitted to the University of Luxembourg

Decoding Cognitive States from Neurophysiological and Behavioral Signals

Members of the thesis supervision committee (CET):

Supervisor:	Prof. Dr. Luis A. LEIVA, University of Luxembourg
Committee members:	Dr. Jean BOTEV, University of Luxembourg Prof. Dr. Tuukka Ruotsalo, LUT University and University of Copenhagen

To the people of my motherland, Iran, the bravest among them.

Acknowledgments

I would like to thank my supervisor, **Prof. Luis Leiva**, for his support, guidance, and patience throughout my PhD. His advice has been essential for my growth as a researcher.

I am also grateful to **Dr. Jean Botev** and **Prof. Tuukka Ruotsalo** for their helpful feedback during the annual Thesis Supervision Committee (CET: *comité d'encadrement de thèse*) meetings, which helped improve the direction and quality of my work.

My thanks go to **Prof. Klen Čopič Pucihar** and **Prof. Matjaž Kljun** for hosting me at the HICUP Lab ("Humans Interacting with Computers" Lab) at the University of Primorska. Their support made possible the work that resulted in the publication presented in Chapter 4.

I would also like to thank **Prof. Jacek Gwizdka** for hosting me at the Information eXperience (IX) Lab at the University of Texas at Austin, where I worked on the exploratory analysis of the AdSERP dataset included in Chapter 3.

I am thankful to **Prof. Alexander Plopski** for hosting me at the Advanced Mixed Reality Interfaces Group. Our discussions initiated an ongoing joint project, and the opportunity to give a PhD talk in his group helped me prepare for my PhD defense presentation.

I would also like to thank all my co-authors of the papers on which this thesis is based for their collaboration and valuable contributions.

Finally, I want to thank my family: my mother **Lida**, whose support and sacrifices made it possible for me to become a researcher; my brother **Pedram**; my father **Moosa**; and my partner **Saeedeh**, for her constant support, patience, and kindness. Their love has sustained me throughout this journey.

Contents

List of Figures

List of Tables

Acronyms

kNN k-Nearest Neighbor. 21, 30–33, 46, 62, 63, 65

BCI Brain-Computer Interface. 21, 24

CNN Convolutional Neural Network. 44, 46, 48, 51

EBR Eye Blink Rate. 86, 88, 91

EEG Electroencephalography. 11–16, 20–26, 30, 31, 37–43, 45, 46, 48–53, 102–104, 111, 115, 116, 118, 119

FCNN Fully Connected Neural Network. 21, 31, 32

GRU Gated Recurrent Unit. 46, 51, 62, 63, 65

GTN Gated Transformer Network. 21, 31, 32

GUI Graphical User Interface. 13, 16, 20, 40–53, 116

HCI Human–Computer Interaction. 11, 12, 14, 15, 17, 19, 21, 23, 39, 53–55, 66, 83, 99, 100, 107–109, 113, 115, 116, 118–120

IR Information Retrieval. 66

LSTM Long Short-Term Memory. 46, 58, 71, 72

MLP Multilayer Perceptron. 72, 73, 81

NDCG Normalized Discounted Cumulative Gain. 72, 73

RNN Recurrent Neural Network. 46, 94

Seq2Seq sequence-to-sequence. 67, 69, 72–75, 81

SERP Search Engine Results Page. 13, 14, 16, 55–61, 63–70, 73–76

SVM Support Vector Machine. 21, 30–33, 46, 62, 63, 65, 88, 94

TEPR Task-Evoked Pupillary Response. 86, 88, 91

TFC Total Fixation Count. 70–72, 81

TFT Total Fixation Time. 70–72, 81

ViT Vision Transformer. 71, 72, 80

Abstract

This thesis investigates how neurophysiological and behavioral signals can be used to decode human cognitive states in Human–Computer Interaction (HCI), with a focus on affect, attention, and expertise. Using Electroencephalography (EEG), eye movements, mouse movements, facial expressions, and head movements, the work explores how implicit, continuous measurements can complement or replace traditional self-reports to enable more adaptive and cognitively aware interactive systems.

First, the thesis examines efficient decoding of affective states from video-elicited EEG. By systematically varying signal length, temporal window size, and frequency bands on two benchmark datasets, it shows that reliable classification of arousal and valence is possible with recordings as short as 30–42 seconds and with short analysis windows. Models operating solely on Beta-band activity achieve accuracy comparable to using all bands, demonstrating that affect decoding can be both data- and computation-efficient. A second study on affect evaluates "good" versus "bad" Graphical User Interfaces (GUIs) using multimodal signals. EEG, eye tracking, and facial activity reveal systematic differences in emotional response, visual exploration patterns, and neural activation, enabling implicit discrimination between interface quality levels without explicit user feedback.

The second part of the thesis focuses on attention in web search. It introduces AdSERP, a large-scale public dataset that combines synchronized eye and mouse movements, HTML structure, and visual layouts for transactional search engine results pages. Analyses show that mouse trajectories closely track gaze. Building

on this dataset, the AdSight framework uses a Transformer-based sequence-to-sequence model to predict gaze-based attention metrics and whether individual result slots are visually examined, using only cursor data and slot metadata. The model achieves high accuracy and ranking quality, enabling scalable, low-cost attention estimation in multi-slot sponsored search.

The third part addresses expertise assessment in medical training. Eye and head movements recorded during simulated breech deliveries are used to differentiate trained from untrained practitioners. Head motion dynamics and pupil-based measures are particularly discriminative, supporting the objective and unobtrusive evaluation of clinical skills.

Finally, the thesis presents Gustav, a cross-device synchronization framework, and Thalamus, a multimodal user simulation toolkit, which together support reproducible, temporally precise multimodal experimentation. Overall, the findings demonstrate that multimodal implicit sensing provides a robust foundation for decoding cognitive states and designing adaptive, user-centered interactive systems.

Chapter 1

Introduction

Understanding and interpreting human cognitive states is at the core of numerous fields, including psychology, neuroscience, and Human–Computer Interaction (HCI). Advances in neurophysiological and behavioral signal processing have opened new possibilities for measuring these internal states in ways that were previously not possible. This thesis explores how such signals—specifically Electroencephalography (EEG) and eye movement data—can be used to decode cognitive states, with a focus on three key areas: affect, attention, and expertise. These areas are fundamental to the design of adaptive and user-centered systems, and their accurate measurement is essential for improving human–computer interactions across various domains, including web search, user interface design, and medical training.

1.1 Motivation and Focus of the Thesis

The motivation for this thesis stems from the growing need to create interactive systems that not only respond to user actions but also adapt based on an understanding of users' cognitive and emotional states. Traditional methods for assessing user experience, such as questionnaires, interviews, and subjective ratings, have limitations: they are time-consuming, subjective, and prone to bias.

Moreover, these methods provide only intermittent snapshots of users' experiences, often failing to capture the dynamic, real-time nature of human interaction with technology.

In contrast, *neurophysiological and behavioral signals*, such as brain activity measured by EEG and eye movements, offer continuous, objective, and real-time indicators of users' cognitive states. These signals allow researchers to passively monitor users' attention, affective states, and even expertise, providing richer and more nuanced insights into user behavior. By decoding these signals, systems can be designed to respond to users in a more natural, adaptive, and personalized manner. For instance, understanding users' *emotional states* can help tailor content or interface design to improve user experience; tracking *attention* can guide information presentation or highlight areas of interest; and assessing *expertise* can inform training systems or feedback mechanisms.

The research presented in this thesis focuses on these three cognitive dimensions: affect, attention, and expertise. Each of these areas plays a critical role in shaping user experience, and each presents unique challenges for both measurement and interpretation.

1.2 Neurophysiological and Behavioral Signals Studied

EEG and eye movement signals were chosen as the primary focus of this research due to their accessibility, noninvasive nature, and relevance to a wide range of applications in HCI. EEG provides high temporal resolution, capturing real-time brain activity in response to stimuli, making it particularly well suited for studying affective and cognitive processes. By analyzing specific frequency bands of the EEG signal, researchers can gain insights into users' emotional arousal, engagement, and mental workload. Furthermore, integrating EEG data with other modalities, such as eye movements, enables a more holistic understanding of

cognitive states.

Eye movement data, on the other hand, offer valuable insights into *attention* and *visual processing*. Eye tracking provides high spatial and temporal precision, allowing researchers to observe how users allocate their gaze across different regions of an interface or scene. This information is essential for understanding how attention is distributed and how users interact with dynamic content. For example, in the context of Search Engine Results Pages (SERPs), eye movement patterns can reveal how users focus on advertisements, organic results, and other visual elements. In addition, *pupil dilation* has been linked to cognitive load and emotional arousal, making it a useful measure for assessing attention and affect in real time.

In addition to EEG and eye movement data, this thesis also examines facial expressions, head movements, and mouse movements as valuable signals for understanding cognitive and emotional states. Facial expressions are used to assess emotional responses and complement affective state decoding, while head movements provide insights into attention and engagement, particularly in training contexts. Mouse movement patterns help track attention allocation and cognitive effort, especially in dynamic environments with multiple visual elements. Together, these signals offer a more comprehensive, multimodal approach to studying user behavior and experience.

1.3 Research Focus Areas

The main focus areas of this research are as follows:

1. **Decoding Affect**: Affective states, such as emotional arousal and valence, are fundamental to user experience. Chapter 2 explores how affective states can be decoded from EEG signals during emotionally evocative video stimuli, as well as how multimodal signals (e.g., EEG and eye movements) can be used to evaluate Graphical User Interfaces (GUIs).

2. **Decoding Attention**: Attention is a key cognitive process that shapes how users interact with information-rich environments. Chapter 3 investigates how users allocate attention on SERPs, examining how eye and mouse movements can be used to track user attention.

3. **Decoding Expertise**: Expertise influences how individuals perceive, interpret, and respond to complex tasks. Chapter 4 examines how eye and head movements can be used to assess medical training skills, focusing on simulated obstetric scenarios. By modeling behavioral and oculomotor patterns, this chapter demonstrates how multimodal signals can reveal underlying skill levels.

4. **Applications and Toolkits**: To support the empirical studies presented in this thesis, Chapter 5 introduces two tools developed to advance multimodal HCI research: *Gustav*, a framework for precise synchronization of multimodal data streams, and *Thalamus*, a simulation toolkit for prototyping and testing multimodal experimental setups. These tools address the technical challenges of collecting and analyzing complex multimodal data, enabling more efficient and reliable experimentation in HCI.

1.4 Objectives of the Thesis

The overarching goal of this thesis is to contribute to the growing body of work on *cognitive state decoding* in HCI by developing and applying novel methods for understanding users' internal states through neurophysiological and behavioral signals. Specifically, this research aims to:

- Investigate the efficiency and accuracy of decoding affective states from EEG data in real-world contexts.
- Develop methods for tracking and predicting user attention in complex, multimodal environments such as SERP.

- Explore the potential of eye and head movement data for assessing expertise in professional training scenarios.
- Design tools that facilitate the collection, synchronization, and simulation of multimodal data, supporting future HCI research and applications.

Through these contributions, this thesis aims to push the boundaries of HCI research by enabling more intuitive, adaptive, and user-centered systems that can understand and respond to users' cognitive and emotional experiences.

1.5 Research Questions

Beyond addressing these objectives, this thesis emphasizes the **novelty** of its contributions in terms of efficiency, scalability, and ecological validity in cognitive state decoding, particularly in realistic HCI settings. This thesis is guided by a set of research questions that aim to investigate how neurophysiological and behavioral signals can be used to decode human cognitive states in HCI. These questions reflect the three core focus areas of the thesis—affect, attention, and expertise—as well as the methodological challenges of multimodal data collection and analysis.

- **RQ1 (Affect):** To what extent can affective states be decoded efficiently from EEG signals under realistic constraints, such as limited signal duration, reduced temporal windows, and selective frequency bands?
- **RQ2 (Attention):** Can user attention in complex interactive environments, such as search engine results pages, be accurately inferred from behavioral signals, particularly mouse movements, as a proxy for gaze?
- **RQ3 (Expertise):** Can multimodal behavioral signals, including eye and head movements, be used to assess user expertise in complex tasks, such as simulated medical procedures?

- **RQ4 (Methodology and Tooling):** How can multimodal data collection, synchronization, and simulation be supported to enable reproducible, scalable, and efficient research in cognitive state decoding?

Each of these research questions is addressed in a dedicated part of the thesis. Chapter 2 investigates affect decoding from EEG signals, Chapter 3 focuses on attention modeling in web search, Chapter 4 explores expertise assessment using multimodal behavioral signals, and Chapter 5 presents tools that support multimodal experimentation. Together, these contributions aim to advance the understanding of cognitive state decoding and support the design of adaptive, user-centered interactive systems.

1.6 Structure of the Thesis

This thesis is organized as follows:

- **Chapter 2**: *Decoding Affect.* This chapter presents two studies on affective state decoding using EEG and multimodal physiological signals, focusing on video-elicited emotional responses and GUI evaluation.
- **Chapter 3**: *Decoding Attention.* This chapter examines attention allocation in SERPs, investigating how eye and mouse movements can be used to decode users' attention.
- **Chapter 4**: *Decoding Expertise.* This chapter explores the use of eye and head movements to assess medical expertise during simulated obstetric training scenarios.
- **Chapter 5**: *Applications and Toolkits.* This chapter introduces two key tools developed in this research: *Gustav*, a synchronization framework for multimodal experiments, and *Thalamus*, a simulation toolkit for prototyping multimodal setups.

- **Chapter 6**: *Conclusion, Discussion, and Outlook.* This chapter summarizes the findings from the previous chapters and discusses the broader implications of this research for the future of adaptive, cognitively aware systems in HCI.

Through the integration of neurophysiological and behavioral signals, as well as the development of supportive tools and frameworks, this thesis provides new insights into the dynamic and complex relationship between human cognition and technology.

1.7 Publications Resulting from This Thesis

This thesis compiles a series of individual studies, each contributing to a broader understanding of cognitive state decoding in HCI. These studies have been published in peer-reviewed journals and conferences and collectively form the empirical backbone of this thesis. Each chapter builds upon the findings of these studies, elaborating on the methods, results, and implications for real-world applications. Bibliography derived from this thesis include the following seven publications:

1. **K. Latifzadeh***, N. Gozalpour*, V. J. Traver, T. Ruotsalo, A. Kawala-Sterniuk, and L. A. Leiva. "Efficient decoding of affective states from video-elicited EEG signals: an empirical investigation". In: *ACM Transactions on Multimedia Computing, Communications and Applications* 20.10 (2024), pp. 1–24. DOI: https://doi.org/10.1145/3663669.

2. S. Haddad*, **K. Latifzadeh***, S. Duraisamy, J. Vanderdonckt, O. Daassi, S. Belghith, and L. A. Leiva. "Good GUIs, Bad GUIs: Affective Evaluation of Graphical User Interfaces". In: *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*. 2024, pp. 232–243. DOI: https://doi.org/10.1145/3627043.3659549.

3. **K. Latifzadeh**, J. Gwizdka, and L. A. Leiva. "A Versatile Dataset of Mouse and Eye Movements on Search Engine Results Pages". In: *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2025, pp. 3412–3421. DOI: https://doi.org/10.1145/3726302.3730325.

4. M. Villaizán-Vallelado, M. Salvatori, **K. Latifzadeh**, A. Penta, L. A. Leiva, and I. Arapakis. "AdSight: Scalable and Accurate Quantification of User Attention in Multi-Slot Sponsored Search". In: *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2025, pp. 255–265. DOI: https://doi.org/10.1145/3726302.3729891.

5. **K. Latifzadeh***, L. A. Leiva*, K. Čopič Pucihar*, M. Kljun, I. Devetak, and L. Steblovnik*. "Assessing Medical Training Skills via Eye and Head Movements". In: *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*. 2025, pp. 1–10. DOI: https://doi.org/10.1145/3699682.3728330.

6. **K. Latifzadeh** and L. A. Leiva. "Gustav: Cross-device cross-computer synchronization of sensory signals". In: *Adjunct Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 2022, pp. 1–3. DOI: https://doi.org/10.1145/3526114.3558723.

7. **K. Latifzadeh** and L. A. Leiva. "Thalamus: A User Simulation Toolkit for Prototyping Multimodal Sensing Studies". In: *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*. 2025, pp. 109–113. DOI: https://doi.org/10.1145/3708319.3733687.

The chapters of this thesis are organized around these studies, with each chapter discussing a specific aspect of cognitive state decoding—affective states, attention, and expertise.

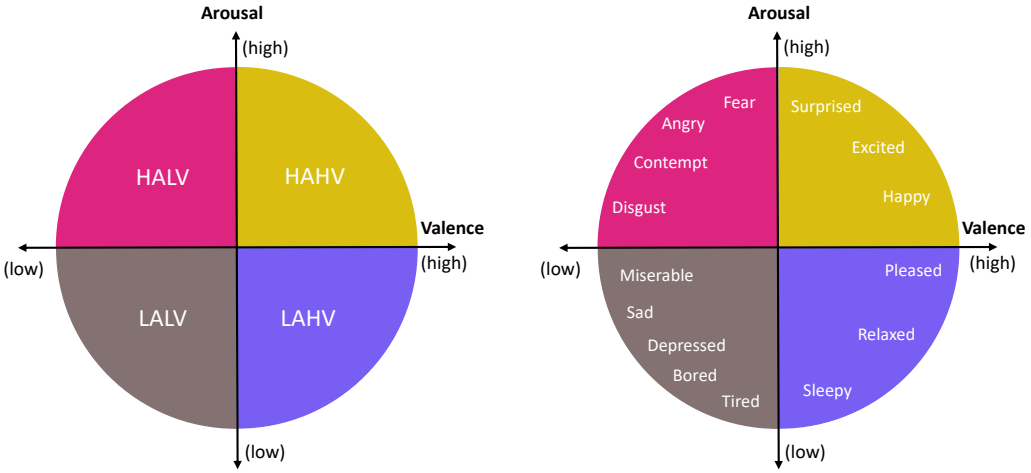
Chapter 2

Decoding Affect

2.1 Introduction

Understanding and modeling human affect is fundamental to developing interfaces that can sense and respond to users' internal experiences. As illustrated in Figure 2.1, *affect* represents the underlying, continuous dimensions of human feeling, defined by two axes: *valence*, which indicates the degree of pleasantness from negative to positive, and *arousal*, which reflects physiological activation from calm to excited states. These dimensions form a two-dimensional space divided into four quadrants: high arousal and high valence (HAHV), high arousal and low valence (HALV), low arousal and high valence (LAHV), and low arousal and low valence (LALV). In contrast, *emotions* correspond to discrete and consciously perceived states, such as happiness, fear, sadness, or relaxation, that emerge from combinations of valence and arousal within this affective space.

In the context of HCI, decoding affective states is essential for developing adaptive and user-centered systems that go beyond usability to account for how people feel during interaction. Traditional evaluation methods, such as questionnaires and interviews, rely on explicit self-reports that are subjective, time-consuming, and sensitive to bias. In contrast, *implicit evaluation* uses physiological and behav-



(a) Affect. **(b)** Emotion.

Figure 2.1 Conceptual models of affect and emotion in the valence–arousal framework. (a) Affect is represented across four quadrants defined by valence and arousal: high arousal–high valence (HAHV), high arousal–low valence (HALV), low arousal–high valence (LAHV), and low arousal–low valence (LALV), where valence spans from negative to positive. (b) Each quadrant in the emotion model contains a set of corresponding discrete emotional states.

ioral signals, such as EEG activity, eye movements, or facial expressions, to infer emotional responses automatically and unobtrusively. These implicit measures offer continuous and objective access to the user's affective state without disrupting the interaction process, enabling a more natural and scalable assessment of user experience.

The studies presented in this chapter address this motivation from two complementary perspectives. The first [1] investigates how efficiently affective states can be decoded from EEG signals recorded while participants watch emotionally evocative videos, with the aim of identifying the minimal data requirements for reliable classification of arousal and valence. The second [2] explores whether multimodal physiological signals can reveal users' emotional responses to GUIs, allowing the distinction between good and bad interface designs without explicit feedback. Together, these works advance the broader goal of implicit and data-efficient affect decoding, contributing to the development of emotion-aware

systems that can sense and adapt to users' cognitive and emotional experiences in real-world HCI contexts. This chapter addresses **RQ1** by investigating how efficiently affective states can be decoded from EEG signals under realistic constraints.

2.2 Efficient Decoding of Affective States from Video-elicited EEG Signals: An Empirical Investigation

Brain-Computer Interfaces (BCIs) offer a promising avenue for recognizing users' emotional states by leveraging non-invasive EEG signals. However, affect decoding in scenarios involving dynamic audiovisual stimuli remains underexplored. To address this gap, we investigate EEG-based emotion recognition from video content through arousal and valence classification, examining the effects of signal duration, feature extraction window size, and EEG frequency bands. We evaluate both traditional machine learning approaches (Support Vector Machines (SVMs) and k-Nearest Neighbors (kNNs)) and deep learning architectures (Fully Connected Neural Network (FCNN) and Gated Transformer Networks (GTNs)). Our findings demonstrate that: (1) reliable affect classification can be achieved with under one minute of EEG data; (2) window lengths of 6–10 seconds yield optimal performance for traditional machine learning models, whereas deep learning models perform best with much shorter windows of approximately 2 seconds; and (3) models trained exclusively on the Beta frequency band achieve performance comparable to, and in some cases exceeding, that of models trained on the full spectrum. Overall, these results suggest that EEG-based affect decoding is feasible under more realistic conditions than previously assumed, supporting its practical application in adaptive interfaces and personalized user modeling.

Motivation

This study is motivated by the increasing relevance of understanding human affect in contexts where emotional reactions influence behavior in fundamental ways, as affect is guided by biologically based action dispositions that shape decisions and actions [3]. Emotions and mood significantly impact daily life and interpersonal behavior, including how individuals interact with others [4, 5], how they make decisions [6], and how they perceive and interpret the world around them [7]. Affect also plays an important role in maintaining physical well-being [8]. As computing systems increasingly mediate human activity, researchers have focused on the need to systematically decode affective reactions to digital information in order to create systems that more naturally respond to users' internal states [7, 6]. Affective responses have traditionally been measured through subjective self-reports and different forms of neurophysiological data [4, 9, 10, 6], and recent research has begun coupling Machine Learning models with insights from neural mechanisms to decode affective states in computer-mediated environments [11].

EEG has gained particular attention for affect decoding due to its simplicity, portability, affordability, and user-friendly nature [4, 7], and the availability of large publicly distributed EEG datasets has accelerated progress by enabling replication and comparison across studies [12, 13, 14]. Despite this progress, most prior research emphasizes maximizing classification accuracy rather than understanding how affect decoding can be made efficient. Studies have focused on identifying subject-specific or cross-subject features that maximize performance [15, 16], developing feature engineering approaches [17], or improving classification using advanced Machine Learning techniques [18]. Yet these studies primarily rely on full-length recordings, assuming that using the maximum amount of data from EEG signals is necessary for optimal decoding performance, as seen in state-of-the-art systems that compute features from entire trial durations [19, 20].

This assumption does not reflect real-world scenarios where users interact with dynamic content, such as videos, in variable ways. People often skip through

videos, focus only on certain segments, or discontinue viewing before the end. In practical HCI systems, acquiring long EEG signals is time-consuming and often impractical. Moreover, affective responses to dynamic content may only emerge at specific temporal moments rather than being uniformly distributed across the entire stimulus duration. As prior studies on visual attention demonstrate, importance evolves dynamically over time [21], and similarly affective experiences fluctuate in response to environmental events [22]. For dynamic audiovisual content, the assumption that affect remains stable during exposure to short-term stimuli may not hold for longer, heterogeneous videos. Understanding which temporal windows carry affective information is therefore critical.

To address this gap, we investigate how EEG signal length, sampling window size, and frequency band selection affect the efficiency of affect decoding during dynamic video viewing. Our goal is to determine whether accurate affect classification is possible with significantly reduced signal duration, shorter feature extraction windows, or selective use of EEG frequency bands. Achieving reliable decoding under these constraints could significantly reduce computational demands, shorten required EEG recordings, and enable real-time affect-adaptive systems that function more effectively within realistic constraints. This study thus aims to challenge conventional assumptions and establish empirically grounded design guidelines for efficient EEG-based affect decoding systems.

Related Work

Previous work in HCI has increasingly examined automated methods for measuring affective responses to support systems that better align with human emotional experiences [23, 24]. This direction has been enabled by the development of affect decoding techniques that play a foundational role in user modeling, adaptive interface systems, and personalized interaction frameworks [25]. EEG-based affect classification has followed several theoretical models, including discrete emotional states such as joy or fear [26], appraisal theories linking emotion to

situational context [27], and dimensional models positioning emotions along continuous dimensions [28]. Dimensional models, especially the valence-arousal model, have proven particularly suitable for computational affect decoding due to their analytic tractability and ability to represent a wide range of human emotions [29]. In this model, arousal reflects intensity, spanning from calm to excited, and valence reflects pleasantness, from negative to positive [30]. Research indicates that all discrete emotions can be approximately localized within this two-dimensional space [31, 32], which has led many works to focus on valence and arousal classification.

EEG signals have been shown to reliably reflect affective states in various studies utilizing BCI and physiological sensing [6, 15, 33, 34]. Behavioral signals may sometimes be intentionally concealed or distorted through facial expressions or vocal modulation, whereas EEG signals are less manipulable, making them attractive for emotion recognition [12]. Consequently, a rapidly growing body of literature employs EEG-based affect decoding, as documented in recent surveys [35, 4, 36, 37]. Despite its advantages, EEG suffers from nonstationarity and high variability across individuals and sessions, complicating subject-independent decoding [12, 13].

Research has differed widely in terms of stimulus design, with earlier works frequently relying on static images [35] and more recent work considering dynamic audiovisual content such as videos [38, 39, 40, 41]. Affect elicited during video viewing, gaming [42, 43, 43, 44, 45, 46], educational settings [47, 48, 46], or cognitively demanding tasks such as stock trading [49], suggests that dynamic stimuli offer richer affective responses.

Although valence and arousal are continuous, many studies binarize them into low versus high categories due to the difficulty of predicting exact numerical values [50, 51]. Some studies extend classification to three, four, or eight discrete classes [52, 53, 39, 54], but binary classification remains the most common choice.

Various Machine Learning techniques have been applied to EEG-based affect

decoding, including Random Forests with Hjorth features [55] and more recent deep architectures such as CNNs [56, 49, 57, 53]. Deep models are challenging to train due to limited EEG data, and collecting sufficient data is costly and time-consuming [58]. Data augmentation methods such as GANs and contrastive learning have been explored with varying success [59, 60, 58, 61, 62, 63, 64, 65, 53]. For instance, contrastive learning combined with graph neural networks achieved moderate improvements on DEAP [14] and MAHNOB-HCI [66] datasets, with binary accuracies ranging from 64.84% to 71.69% [67]. However, these studies rely on full-length EEG signals and do not examine how shorter signals or specific frequency bands influence performance.

Beyond classification models, the extraction of reliable markers from EEG remains complex due to its multidimensionality, nonlinearity, and sensitivity to temporal and spatial factors [68, 69]. Frequency bands carry different discriminative information depending on the task, yet many studies restrict analysis to traditional bands or combine all bands without examining their individual contributions [68, 70, 53]. Recent work using multimodal hybrid systems integrating EEG with eye tracking shows improved performance [71, 72, 73, 41], though such methods introduce synchronization challenges [74].

Temporal aspects of EEG processing have received limited attention. Some studies show benefits from longer signal durations for disease detection [75] or motor imagery [76, 77], whereas others suggest shorter windows can be effective [78, 79]. Window sizes from 1 to 30 seconds have been used in emotion recognition, and no consensus exists [80]. Feature extraction from very short windows may be computationally expensive [81], while longer windows may smooth out important temporal fluctuations. One prior study found that the latter half of EEG signals during video watching was more discriminative for affect than earlier segments [50], but this assumes full-length signal availability. Because affective interpretations can be influenced by previous exposure to stimuli [82], analyzing segments from the onset of a stimulus may be essential for building

models suitable for real-time systems.

The overarching trend across prior work is a focus on maximizing performance with full EEG recordings rather than optimizing efficiency for realistic settings. The present study seeks to shift this focus by systematically evaluating how signal length, window size, and frequency bands affect decoding performance in dynamic content viewing contexts.

Methodology

We conducted extensive experiments using two public datasets to examine EEG-based affect decoding during video viewing. Our methodology systematically manipulates EEG signal length, temporal windows for feature extraction, and EEG frequency bands, enabling an exploration of decoding efficiency under varying conditions.

The datasets used were DEAP [14] and MAHNOB-HCI [66]. DEAP consists of 1,280 EEG recordings from 32 participants, each corresponding to a 60 second music video accompanied by self-reported arousal and valence scores on a scale of 1 to 9. The dataset includes 32 EEG channels downsampled to 128 Hz, with EOG artifacts removed and a 4–45 Hz bandpass filter applied to eliminate low frequency movement noise and high frequency interference from electromagnetic sources and muscle contractions. Figure 2.2 visually presents the sample distribution of DEAP dataset across the valence-arousal space, showing all four affect quadrants represented.

MAHNOB-HCI includes EEG data from 27 participants who watched 20 videos lasting between 35 and 117 seconds, with discrete arousal and valence ratings on a 1 to 9 scale. The EEG was originally sampled at 256 Hz across 32 channels. For comparability, we downsampled all EEG data to 128 Hz, applied Artifact Subspace Reconstruction (ASR) to remove transient large-amplitude artifacts [83], applied a 4–45 Hz bandpass filter, and performed Common Average Referencing (CAR) [84]. Since DEAP trials are uniformly 60 seconds long, MAHNOB-HCI trials shorter

than 60 seconds were removed, and trials longer than 60 seconds were truncated to the first minute, resulting in a final set of 449 standardized trials. Figure 2.3 visually presents the sample distribution of MAHNOB-HCI dataset across the valence-arousal space, showing all four affect quadrants represented.

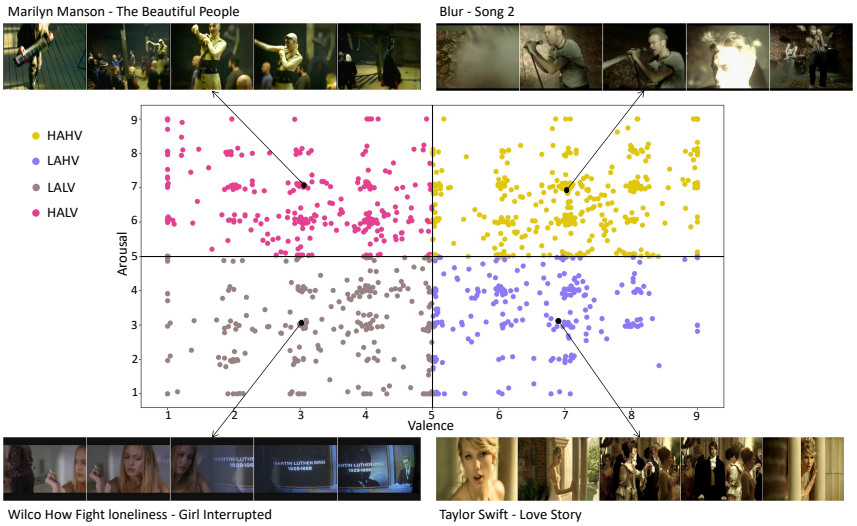


Figure 2.2 Distribution of EEG samples from the DEAP dataset within the valence–arousal space. Samples are categorized into high arousal–high valence (HAHV), high arousal–low valence (HALV), low arousal–high valence (LAHV), and low arousal–low valence (LALV). Each point represents the arousal and valence ratings reported by a participant following exposure to a music video.

The experimental design centers on two temporal parameters: signal length l and window size w. We consider window sizes of 2, 4, 6, and 10 seconds based on previous work [85, 86, 78, 79]. Signal length is defined as $l = w \cdot n$, where n is the number of windows into which the signal is partitioned. Thus, for $w = 2$ seconds, signal lengths of 2, 4, 6, …, 60 seconds are possible, while for $w = 6$ seconds, signal lengths of 6, 12, 18, …, 60 seconds are evaluated. Figure 2.4 provides a visual example of how different signal lengths are derived from a consistent window size.

Within each window, we extract features from four frequency bands defined through FFT: Theta (4–7 Hz), Alpha (8–12 Hz), Beta (13–30 Hz), and Gamma (31+

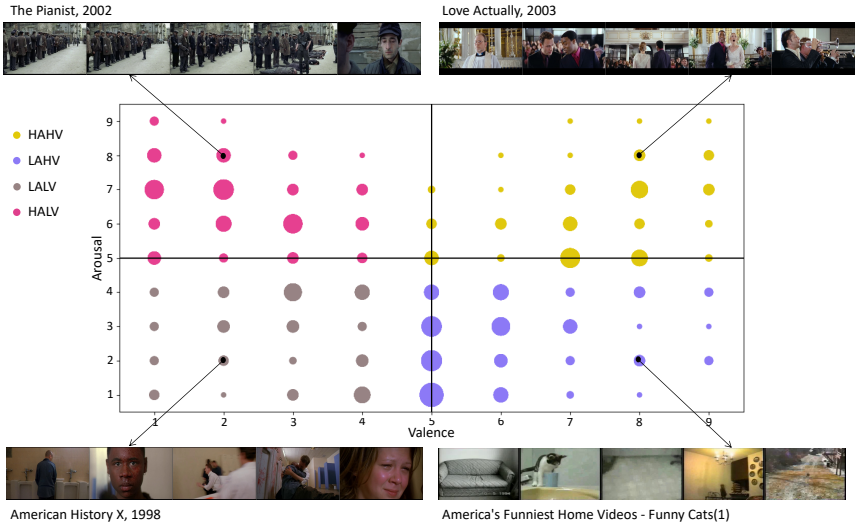


Figure 2.3 Distribution of EEG samples from the MAHNOB-HCI dataset within the valence–arousal space. Samples are grouped into high arousal–high valence (HAHV), high arousal–low valence (HALV), low arousal–high valence (LAHV), and low arousal–low valence (LALV), corresponding to combinations of high and low arousal and valence. Because ratings are reported on a 1–9 scale, the radius of each marker reflects the number of samples at a given coordinate, with larger markers indicating higher frequencies. For instance, a single sample appears at (valence, arousal) = (2, 9), whereas 10 and 24 samples are observed at (6, 4) and (5, 1), respectively.

Hz). FFT is applied to each window, and the inverse-transformed band-specific signals are then used for feature extraction.

We compute **five features**, consisting of the three classical Hjorth parameters [87] and two additional spectral measures. Hjorth parameters describe a time series in terms of its power spectrum: activity, mobility, and complexity. For a signal $x(t)$, the power spectrum is $S(m) = |X(m)|^2$, where $X(m)$ is the discrete Fourier transform of $x(t)$.

Hjorth activity indicates the variance of the time series:

$$\text{Activity} = \text{Var}(x(t)) = \frac{1}{N-1} \sum_{i=1}^{N} |x_i - \mu_x|^2 , \tag{2.1}$$

where $x(t)$ is the EEG signal expressed as a discrete time series with N values x_i,

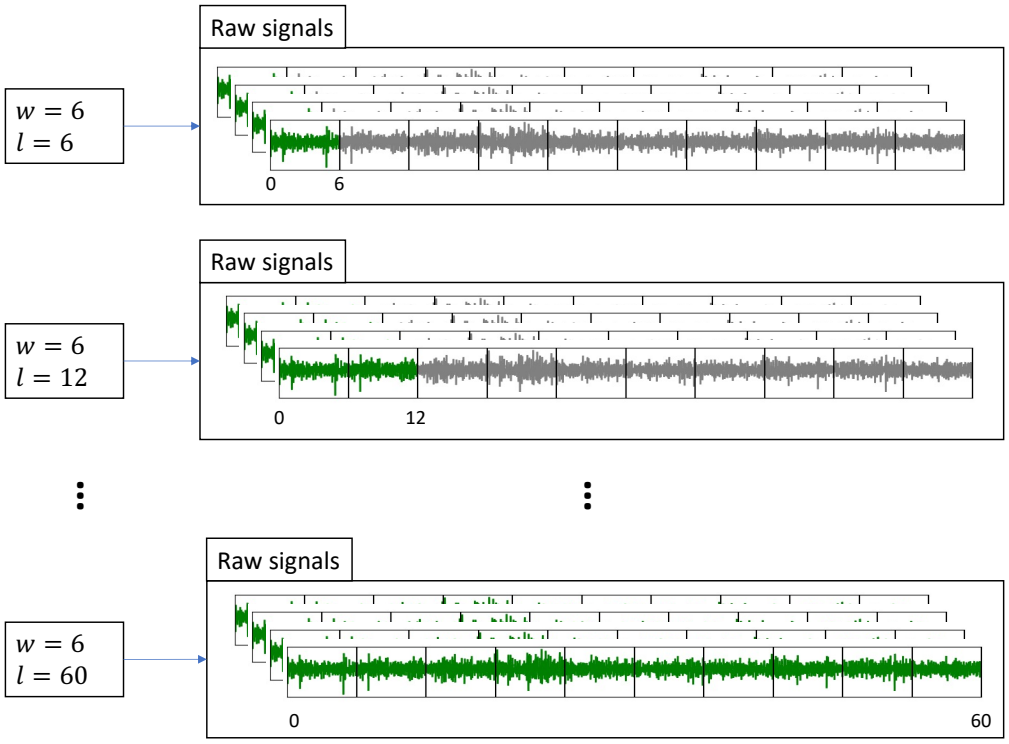


Figure 2.4 Extraction of EEG data instances using varying signal lengths within a fixed sampling window. The figure illustrates how EEG segments of incrementally different durations are analyzed while maintaining a constant window size (6 seconds in this example). Green segments indicate the extracted EEG samples used for affect decoding.

and μ_x is the sample mean.

Hjorth mobility is proportional to the standard deviation of the power spectrum:

$$\text{Mobility} = \sqrt{\frac{\text{Var}\left(\frac{d}{dt}x(t)\right)}{\text{Var}(x(t))}}, \tag{2.2}$$

Hjorth complexity quantifies the rate of change in the signal's frequency content:

$$\text{Complexity} = \frac{\text{Mobility}\left(\frac{d}{dt}x(t)\right)}{\text{Mobility}(x(t))}, \tag{2.3}$$

In addition to the Hjorth parameters, we compute two spectral-domain measures. The **spectral entropy** [88] characterizes the distribution of spectral power

and is computed following Shannon's definition:

$$H = -\sum_{m=1}^{N} P(m) \log_2 P(m), \tag{2.4}$$

where N is the number of frequency points, and for an FFT-discretized signal $X(m)$, the normalized power distribution ($P(m)$) is

$$P(m) = \frac{S(m)}{\sum_i S(i)}. \tag{2.5}$$

Finally, we extract the **signal energy**, which measures the overall strength of the time series [89]:

$$E = \int_0^N |x(t)|^2 \, dt. \tag{2.6}$$

These five features have demonstrated strong performance in prior EEG emotion-recognition literature [90, 91, 92, 55, 51, 93, 94]. Figure 2.5 illustrates the extraction of frequency bands and features from EEG data.

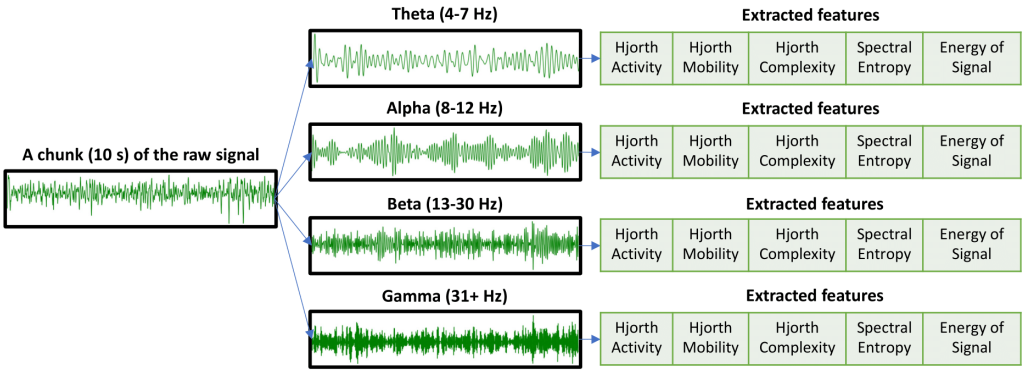


Figure 2.5 Extraction of EEG frequency bands and features. The figure presents an example of a 10 second segment of raw EEG data from the DEAP dataset, from which the corresponding frequency components are derived.

Four types of classifiers are used. SVMs incorporate a variety of kernels and hyperparameters, including C, γ, and polynomial degree, optimized via Bayesian search [95]. kNN classifiers vary in number of neighbors, distance metrics, and

weighting schemes. FCNNs consist of two hidden layers of 800 tanh-activated neurons, with batch normalization and a sigmoid output, trained using Adam with a learning rate of 10^{-3} and early stopping. GTNs, designed for multivariate time series classification, include spatial and temporal transformer branches whose outputs are combined through a gating mechanism. GTNs are trained using Adagrad with scheduled learning rates, dropout, and early stopping.

For classification, arousal and valence scores are normalized to $[0, 1]$ and binarized at 0.5, following common practice. Data are split into 90% training and 10% testing portions, maintaining class balance. SVM and kNN models use 10-fold cross-validation for hyperparameter tuning, while FCNN and GTN models use validation splits from the training set. Balanced Accuracy and F1 scores serve as evaluation metrics.

Results

The results are presented across Figure 2.6 to Figure 2.9 and Table 2.1 to Table 2.8. The analysis includes performance comparisons across signal lengths, window sizes, and frequency bands, as well as statistical tests examining differences between classifiers and design parameter choices.

In the DEAP dataset for arousal classification, the best performance using a single frequency band and a partial EEG signal was 70% Balanced Accuracy, achieved by an SVM trained on the Beta band with a 10-second window and a 30-second signal length, as described in Table 2.1. Statistical tests revealed significant differences between classifiers, with GTN performing worse than SVM but no significant differences among SVM, kNN, and FCNN models. No significant differences were found between individual frequency bands and the combined set of bands. Window size exhibited no significant main effects, and signal length from 10 to 60 seconds also showed no significant differences. Using full-length signals, Beta and all bands combined achieved competitive results, with Table 2.2 showing Balanced Accuracy values of 69%.

For DEAP valence classification, the best partial-signal performance among single bands was 65% Balanced Accuracy, achieved by an SVM trained on the Beta band with a 6 second window and a 48 second signal length, reported in Table 2.3. GTN again performed worse than SVM, while SVM, kNN, and FCNN showed no significant differences. The Beta band matched the performance of all bands combined, while the remaining bands performed significantly worse, indicating that valence information is concentrated in the Beta range. As with arousal, no significant differences existed among window sizes or signal lengths tested. Table 2.4 shows that full-length signals offered little improvement, with Balanced Accuracy for the Beta and for all bands combined at 66%.

In the MAHNOB-HCI dataset for arousal classification, the kNN classifier achieved the highest partial-signal performance, reaching 88% Balanced Accuracy using the Beta band with a 6 second window and a 42 second signal length, as detailed in Table 2.5. GTN performed significantly worse than kNN in statistical tests, while SVM, kNN, and FCNN did not differ significantly. The Beta band again matched the performance of using all bands combined, while other bands were significantly worse. Signal lengths of 40 seconds or longer yielded significantly better performance than shorter segments. Table 2.6 shows that full-length analysis yielded the highest arousal performance when all bands were used, reaching 92% Balanced Accuracy.

For MAHNOB-HCI valence classification, the best partial-signal performance was 85% Balanced Accuracy, achieved by kNN with the Beta band, a 6 second window, and a 54 second signal length, as presented in Table 2.7. Statistical tests reveal that GTN performed significantly worse than kNN, while no significant differences appeared among the other classifiers. Again, the Beta band matched all bands combined in performance, and windows of 6 and 10 seconds did not differ significantly. Longer signal lengths of around 40 seconds or more significantly improved performance. Full-length results in Table 2.8 show that the best performance for valence classification was achieved using all bands, reaching 93%

Balanced Accuracy.

Across both datasets and both target variables, the general pattern is that the Beta band performs as effectively as the combined set of all bands. Classic Machine Learning classifiers such as SVM and kNN frequently outperform or match deep models. Signal lengths of approximately 30 to 42 seconds achieve performance comparable to full 60 second signals, and windows of 6 or 10 seconds are consistently optimal for traditional models, whereas deep models benefit from shorter windows.

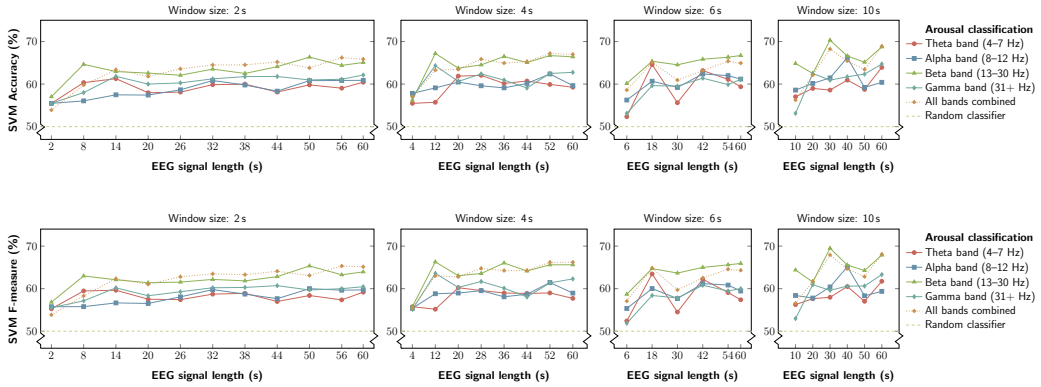


Figure 2.6 Balanced accuracy (top) and F_1 score (bottom) for **arousal** classification using support vector machines (SVMs) on the DEAP dataset. Performance is reported as a function of EEG signal length (horizontal axis, in seconds) under different sliding window sizes employed for feature extraction.

			SVM hyperparams				
Freq. band	l	w	**kernel**	C	γ	**Bal. Acc.**	**F_1 score**
Theta	18	6	RBF	1000.000	0.049	0.646	0.635
Alpha	40	10	RBF	371.830	0.136	0.662	0.651
Beta	30	10	RBF	62.657	0.213	0.703	0.695
Gamma	12	4	RBF	1000.000	0.044	0.643	0.636
All bands	30	10	RBF	128.562	0.010	0.682	0.679

Table 2.1 Best-performing support vector machine (SVM) models for **arousal** classification on the DEAP dataset using **partial** EEG signals ($l < 60$).

			SVM hyperparams				
Freq. band	l	w	**kernel**	C	γ	**Bal. Acc.**	**F_1 score**
Theta	60	10	RBF	109.299	0.235	0.639	0.617
Alpha	60	2	RBF	6.200	1.000	0.609	0.597
Beta	60	10	RBF	14.234	0.558	0.688	0.679
Gamma	60	10	RBF	1000.000	0.056	0.647	0.633
All bands	60	10	RBF	2.001	0.630	0.689	0.681

Table 2.2 Best-performing support vector machine (SVM) models for **arousal** classification on the DEAP dataset using **full-length** EEG signals ($l = 60$).

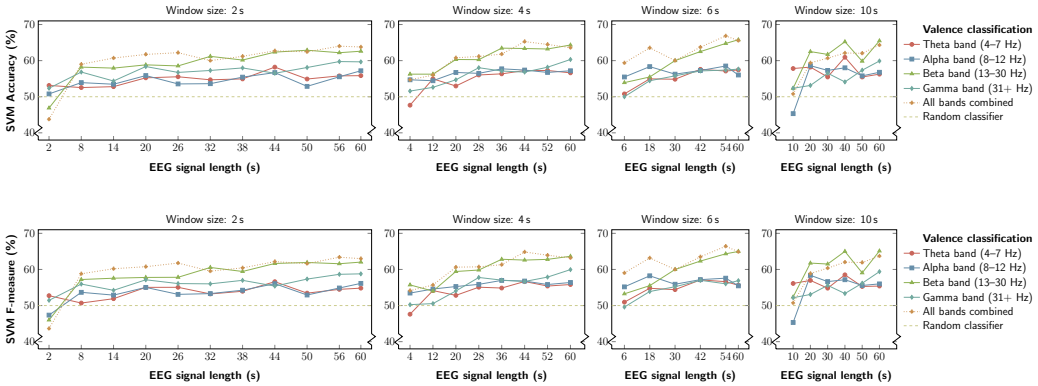


Figure 2.7 Balanced accuracy (top) and F_1 score (bottom) for **valence** classification using support vector machines (SVMs) on the DEAP dataset. Performance is shown as a function of EEG signal length (horizontal axis, in seconds) for different sliding window sizes used in feature extraction.

			SVM hyperparams				
Freq. band	l	w	**kernel**	C	γ	**Bal. Acc.**	**F_1 score**
Theta	16	4	RBF	5.264	1.000	0.600	0.591
Alpha	20	10	RBF	1000.000	0.072	0.586	0.584
Beta	48	6	RBF	5.749	1.000	0.656	0.651
Gamma	16	4	RBF	974.833	0.106	0.596	0.596
All bands	54	6	RBF	2.764	0.389	0.668	0.664

Table 2.3 Best-performing support vector machine (SVM) models for **valence** classification on the DEAP dataset using **partial** EEG signals ($l < 60$).

			SVM hyperparams				
Freq. band	l	w	**kernel**	C	γ	**Bal. Acc.**	**F_1 score**
Theta	60	4	RBF	464.061	0.203	0.566	0.558
Alpha	60	4	RBF	271.481	0.215	0.572	0.564
Beta	60	10	RBF	14.206	0.477	0.655	0.651
Gamma	60	4	RBF	1000.000	0.475	0.603	0.600
All bands	60	6	RBF	2.887	0.448	0.655	0.649

Table 2.4 Best-performing support vector machine (SVM) models for **valence** classification on the DEAP dataset using **full-length** EEG signals ($l = 60$).

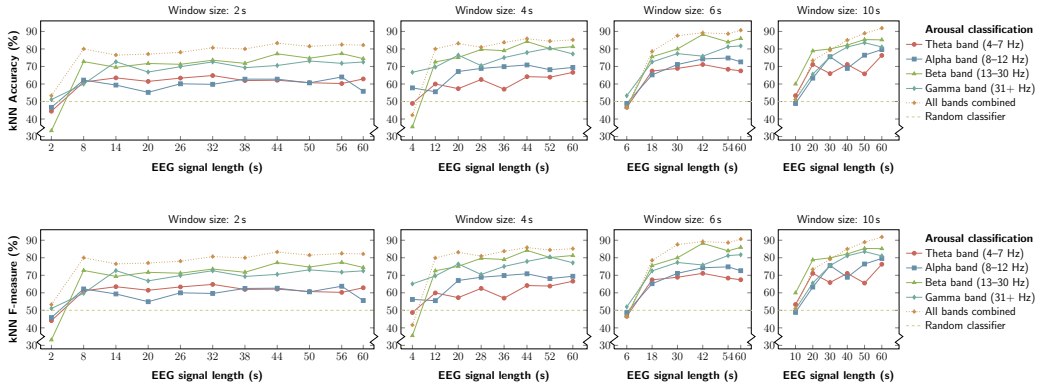


Figure 2.8 Balanced accuracy (top) and F_1 score (bottom) for **arousal** classification using k-nearest neighbor (k-NN) classifiers on the MAHNOB-HCI dataset. Performance is reported as a function of EEG signal length (horizontal axis, in seconds) for different sliding window sizes employed in feature extraction.

			k-NN hyperparams				
Freq. band	l	w	k	p	v	**Bal. Acc.**	**F_1 score**
Theta	20	10	1.0	1.0	distance	0.711	0.711
Alpha	50	10	4.0	1.0	distance	0.764	0.764
Beta	42	6	1.0	1.0	distance	0.883	0.883
Gamma	48	6	2.0	1.0	distance	0.836	0.836
All bands	48	6	1.0	1.0	distance	0.908	0.908

Table 2.5 Best-performing k-nearest neighbor (k-NN) models for **arousal** classification on the MAHNOB-HCI dataset using **partial** EEG signals ($l < 60$).

			k-NN hyperparams				
Freq. band	l	w	k	p	v	**Bal. Acc.**	**F_1 score**
Theta	60	10	1.0	1.0	distance	0.763	0.763
Alpha	60	10	6.0	1.0	distance	0.796	0.796
Beta	60	6	1.0	1.0	distance	0.860	0.860
Gamma	60	6	1.0	1.0	uniform	0.817	0.817
All bands	60	10	1.0	1.0	uniform	0.919	0.919

Table 2.6 Best-performing k-nearest neighbor (k-NN) models for **arousal** classification on the MAHNOB-HCI dataset using **full-length** EEG signals ($l = 60$).

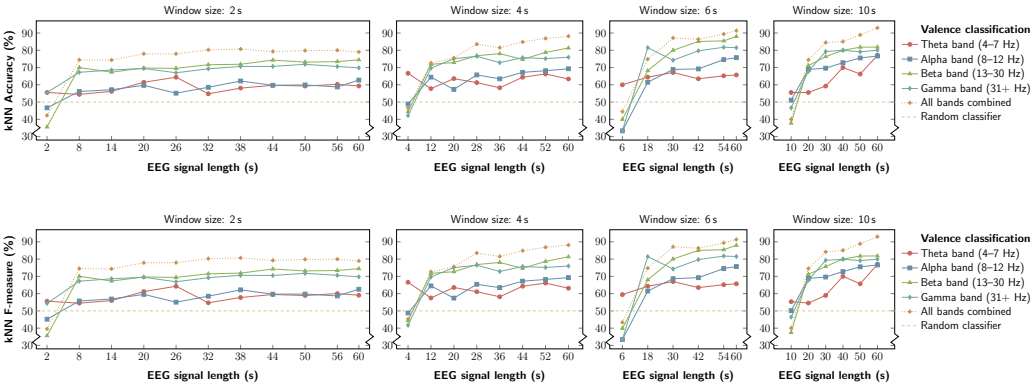


Figure 2.9 Balanced accuracy (top) and F_1 score (bottom) for **valence** classification using k-nearest neighbor (k-NN) classifiers on the MAHNOB-HCI dataset. Performance is presented as a function of EEG signal length (horizontal axis, in seconds) under different sliding window sizes used for feature extraction.

			k-NN hyperparams				
Freq. band	l	w	k	p	v	**Bal. Acc.**	**F_1 score**
Theta	24	6	1.0	1.0	uniform	0.711	0.711
Alpha	50	10	1.0	1.0	distance	0.756	0.755
Beta	54	6	4.0	1.0	distance	0.854	0.854
Gamma	54	6	2.0	1.0	distance	0.817	0.818
All bands	54	6	2.0	1.0	distance	0.894	0.894

Table 2.7 Best-performing k-nearest neighbor (k-NN) models for **valence** classification on the MAHNOB-HCI dataset using **partial** EEG signals ($l < 60$).

			k-NN hyperparams				
Freq. band	l	w	k	p	v	**Bal. Acc.**	**F_1 score**
Theta	60	10	1.0	1.0	distance	0.770	0.769
Alpha	60	10	1.0	1.0	distance	0.767	0.766
Beta	60	6	4.0	1.0	distance	0.880	0.880
Gamma	60	6	1.0	1.0	distance	0.815	0.815
All bands	60	10	1.0	1.0	distance	0.930	0.930

Table 2.8 Best-performing k-nearest neighbor (k-NN) models for **valence** classification on the MAHNOB-HCI dataset using **full-length** EEG signals ($l = 60$).

Discussion and Limitations

The findings demonstrate that efficient affect decoding from EEG is possible without requiring full-length signal processing. Performance reaches its peak within approximately 30 seconds of data for DEAP and within 40–42 seconds for MAHNOB-HCI, meaning that reliable affect recognition can be achieved even when users do not watch entire videos or when real-time adaptive systems cannot afford full-length analysis. This directly challenges the common assumption that complete signals must be used for optimal performance. Sampling windows of 6 and 10 seconds produce the most reliable performance for traditional Machine Learning models, while shorter windows benefit deep models by increasing the number of training samples, which is particularly important in data-scarce EEG contexts.

Another significant finding is that the Beta band alone carries enough discriminative information to match models trained on all frequency bands, contrasting with older assumptions that combining all bands is necessary [53] and aligning with recent work suggesting that high-frequency components contain strong affective signals [53]. Focusing on the Beta band alone reduces computational cost and lowers noise, making it advantageous for real-time systems.

Balanced Accuracy and F1 scores were consistently similar across experiments, implying that models did not show bias toward either class and that errors in false

positives and false negatives were balanced, which is valuable for applications where both misclassification types pose risks.

Despite these promising findings, several limitations temper the generalizability of the results. Both DEAP and MAHNOB-HCI are curated datasets that include videos specifically chosen to elicit affective responses. Because of this, affective responses may arise early in the signal, making short segments unusually discriminative. In real-world settings with more variable content, affective responses may occur less consistently. Another limitation is that signal segments were treated as independent classification instances across both datasets, a common but potentially problematic practice [80, 78, 79], since dependencies across trials or participants may influence performance. Although deep models were included, none outperformed simpler classifiers, likely due to the limited size of the available datasets, a well-known challenge in EEG research [35, 6]. We note that additional modalities such as electrodermal activity or heart-rate variability might improve performance in future hybrid systems. Finally, crowdsourcing approaches for gathering additional data may be promising but remain methodologically underdeveloped, especially regarding how to combine multi-participant signals to improve reliability.

Future Work

While our results suggest that efficient affect decoding is feasible under realistic constraints, several promising directions remain. First, future work should validate these efficiency findings in less curated and more heterogeneous viewing contexts (e.g., everyday streaming content, short-form clips, or user-driven skipping), where affective responses may be weaker, delayed, or less consistent than in benchmark datasets. In addition, our current setup treats signal segments as independent instances; a natural next step is to explicitly model and evaluate dependencies across participants and videos (e.g., cross-video generalization, subject-wise splits, or hierarchical models), since participant- and

stimulus-specific factors can materially shape decoding performance.

Second, expanding the data and sensing assumptions could improve robustness without sacrificing efficiency. One avenue is to explore scalable data collection (e.g., crowdsourcing) together with principled methods for fusing signals across users, including studying the trade-off between number of participants and required signal length for reliable decoding. Another is multimodal fusion with peripheral physiology (e.g., heart-rate variability or electrodermal activity) to compensate for EEG variability and strengthen inference in noisy real-world settings. Finally, future systems could move beyond fixed windowing by learning adaptive, content- or event-aware temporal sampling strategies that identify informative moments online, enabling truly real-time affect-adaptive interfaces.

Conclusion

Affect decoding is becoming increasingly central in HCI, where systems that respond to users' emotional states can create richer and more adaptive interactions. This study addresses a critical gap in the field by examining how affect decoding can be performed efficiently when users are exposed to dynamic audiovisual content. Through systematic evaluation of EEG signal length, sampling window size, and frequency band selection, we demonstrate that affective responses can be decoded accurately using relatively short EEG segments, larger window sizes for traditional models, and selective use of the Beta frequency band. These findings have significant implications for the design of real-time affective systems, enabling faster data collection, reduced computational overhead, and improved adaptability to real-world settings. By showing that shorter durations and selective features suffice for robust decoding, the study opens new pathways for building efficient, scalable, and practical EEG-based affect recognition models that function effectively in dynamic HCI environments.

2.3 Good GUIs, Bad GUIs: Affective Evaluation of Graphical User Interfaces

Affective computing can play a significant role in enhancing the design and evaluation of graphical and intelligent user interfaces by enabling emotion-aware adaptation. Despite this potential, emotional responses are rarely leveraged to assess whether a GUI design is perceived as effective or ineffective. In this work, we explore the use of physiological signals as a fast, unobtrusive, and reliable means of affective evaluation for GUI designs, eliminating the need for explicit user feedback. We carried out a controlled study involving 29 participants who viewed 20 well-designed and 20 poorly designed GUIs, while simultaneously capturing eye movements via eye tracking, facial expressions through video recordings, and neural activity using EEG. Distinct patterns emerged across the collected modalities, prompting the training and comparison of multiple computational models to discriminate between good and bad designs. Our results indicate that each sensing modality exhibits an optimal trade-off between model choice and signal duration. Overall, the findings demonstrate that physiological data can effectively differentiate between high- and low-quality GUI designs, opening the door to implicit, user-centered evaluation approaches based on physiological user modeling.

Motivation

The question of whether a GUI is good or bad remains central to interface evaluation, since design errors become more costly to repair the later they are discovered, and early, efficient, and minimally intrusive evaluation methods are needed to support iterative design processes [96, 87, 97, 98]. Traditional evaluation methods can be classified according to whether they involve users or not, and whether they rely on real or represented GUIs [99]. Methods without users have attempted to automate GUI evaluation using heuristics, guidelines, or computational aes-

thetics models [100, 101, 102, 103, 104]. Although useful, their results remain independent of actual users and contexts, and prior studies have shown that aesthetic preferences vary substantially with personal taste, psychological traits, and demographic background [105, 106, 107]. Self-reported evaluations of aesthetics are also limited due to subjectivity and carry-over effects [98].

Evaluation methods with users provide context-sensitive outcomes but are often costly, explicit, and conducted too late, requiring users to verbalize or justify their judgments [108, 109]. At the same time, the emotional impact of GUI design has gained increasing attention, as aesthetics influence not only usability judgments but also emotional reactions that contribute to engagement, brand loyalty, and user retention [110, 111, 112, 113, 114]. Positive aesthetic experiences are linked to physiological well-being [115] and are known to affect user perceptions at early stages of interaction. Conversely, poorly designed GUIs can induce negative emotions that lead to frustration or abandonment [114].

To bridge the gap between subjective aesthetic impressions and objective evaluation needs, the paper investigates whether implicit affective responses can serve as early indicators of GUI quality. Neurological and peripheral physiological signals, such as facial expressions [116], eye activity [114], and EEG [117], may reveal rapid, automatic emotional reactions that do not require explicit self-reporting. Because labeled affect datasets for GUIs are essentially nonexistent [114], building models of affective responses to interface design is challenging. However, based on intersubjectivity principles suggesting that shared reactions can become objective indicators [109, 98], the authors hypothesize that physiological signals can reliably distinguish good and bad GUI designs. This research aims to uncover how users affectively respond to interface aesthetics and content properties, and whether these responses can support implicit evaluation of GUI design quality. This study investigates whether multimodal physiological data can provide early, implicit, and efficient signals for distinguishing good and poor GUIs, thereby enabling new approaches to evaluating interface design throughout

the development lifecycle.

Related Work

The research builds on two main areas: GUI aesthetics and emotion recognition. Within GUI aesthetics research, numerous methods have been proposed to evaluate interface quality [118], with aesthetics shown to influence perceived usefulness, credibility, and likelihood of revisiting [119, 120, 121, 122]. Findings have been mixed regarding whether users consistently distinguish good from bad designs. Some studies report that both experts and non-experts can reliably detect poor designs, although experts may sometimes judge good designs more harshly [119]. Automatic evaluation systems have reinforced that people rely on similar visual cues regardless of viewing duration [123, 115, 124]. Standard aesthetic evaluation often uses questionnaires [125, 126] which require explicit user input. By contrast, physiological measurements provide a way to capture user responses during natural interaction, addressing limitations of explicit and model-based methods.

Emotion recognition research is extensive [127, 128, 129, 130, 131] and commonly relies on facial expressions, eye tracking, and EEG signals. Facial expression methods often achieve high accuracy in controlled datasets with actors [132, 133, 134, 135] but are unrealistic for GUI evaluation where users rarely display exaggerated emotional expressions [132, 133]. Eye tracking has been used extensively for GUI evaluation [100, 136, 137] and generation [136], but its use in affect recognition remains limited, with prior work noting that eye movements alone are unreliable for emotion detection [138] and are often combined with other signals such as EEG [139]. However, pupil dilation and fixation patterns have been shown to correlate with emotional arousal, task efficiency, and aesthetic appraisal [140, 141, 105]. Still, little is known about how eye movements respond specifically to GUI design quality.

EEG-based affect recognition has analyzed frequency bands associated with

cognitive and emotional states [142, 143] and achieved classification results on popular datasets such as DEAP [14] using methods ranging from unsupervised learning [102] to wavelet-based decompositions [144] and deep neural architectures [145]. Multimodal approaches combining EEG with eye tracking or voice have been shown to improve accuracy [146, 147, 148, 41] but also require complex synchronization setups [97]. The present study departs from prior work by examining GUI aesthetics rather than audiovisual or affective stimuli and by jointly considering eye activity, facial expressions, and EEG signals during exposure to real interface designs.

Methodology

We conducted a controlled within-subjects experiment to determine whether physiological signals can distinguish good and bad GUI designs. Forty GUI designs were selected from the LabintheWild dataset [126], which contains aesthetic ratings for 418 websites from approximately 32,000 participants worldwide. The top 20 most positively rated designs were categorized as good and the lowest 20 as bad, ensuring a realistic distribution of aesthetic judgments and reflecting natural variability in user preferences. Figure 2.10 illustrates examples of good and bad GUI designs. These stimuli were presented to participants across two sessions to assess affective responses and test–retest reliability.

Twenty-nine participants between ages 18 and 46 completed the experiment, after excluding three participants due to data quality issues, equipment failures, or outlier behavior. All participants had normal or corrected-to-normal vision and most had no formal training in GUI design. We collected three physiological modalities: facial expressions recorded through a webcam at 30 fps, EEG signals recorded using an EEG device with 8 electrodes sampled at 250 Hz and notch-filtered at 50 Hz, and eye-tracking data captured using an eye tracker sampling at 150 Hz and mounted to a 21.5-inch monitor. All electrodes used conductive gel to ensure high signal quality. Figure 2.11 illustrates the experimental setup.

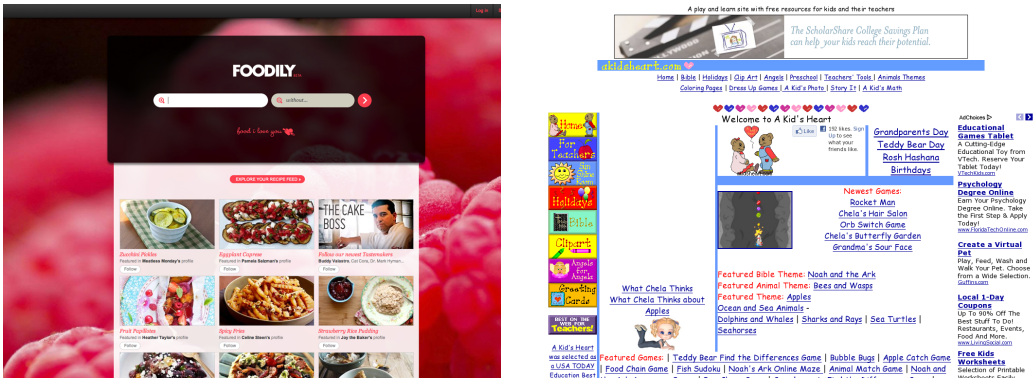


Figure 2.10 Examples of good (left) and bad (right) graphical user interface (GUI) designs. The illustrated cases correspond to the ratings closest to the centroid of each respective group.

Each trial began with a 5-second resting phase serving as a physiological baseline, followed by a 10-second exposure to a randomly selected GUI design. Participants then answered three rating questions: an overall evaluation on a 9-point scale, a valence rating from very unpleasant to very pleasant, and an arousal rating from very calm to very exciting. Participants could adjust their answers freely before confirming their responses. Each participant viewed 20 good and 20 bad designs in randomized order, followed by a second session with the same designs in a different randomized sequence, separated by a 5-minute break.

Eye activity was analyzed by examining pupil size during different time windows: the first second, the first five seconds, and the full ten-second exposure. Missing data due to blinks were interpolated and the signals were smoothed with a Savitzky-Golay filter [111]. Differences between consecutive smoothed values were normalized between 0 and 100. Fixation data were extracted for each time interval and fixation heatmaps were generated. Facial expression data consisted of frames extracted every 0.5 seconds for the first five seconds and every two seconds for the full ten seconds of exposure. Faces were detected using a Haar Cascade [149] classifier, converted to grayscale, and fed to a Convolutional

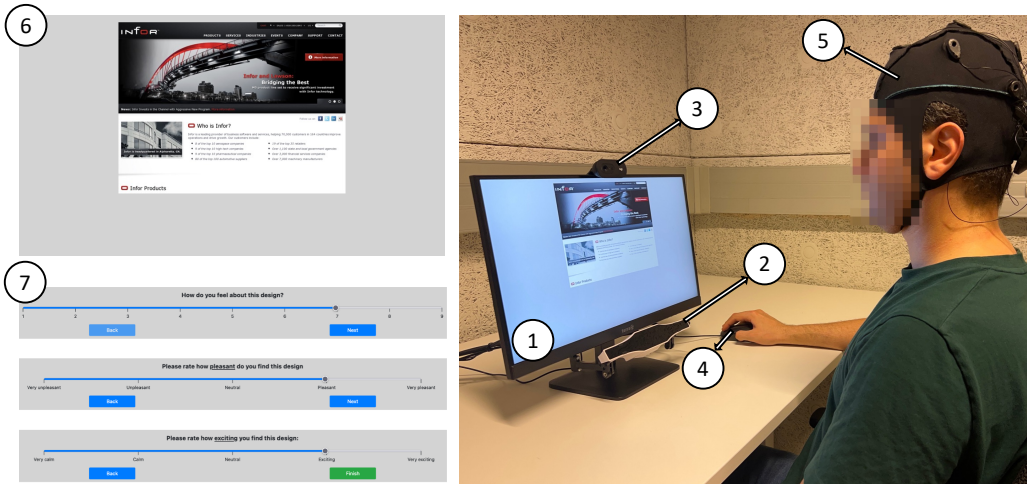


Figure 2.11 Experimental setup and apparatus. Participants wore a Unicorn Hybrid Black EEG cap (5) and were seated in front of a monitor (1) equipped with a Gazepoint GP3 eye tracker (2) and a webcam (3). A computer mouse (4) was used to provide ratings (7) after viewing each GUI design (6). User ratings were collected exclusively as a measure of perceived quality.

Neural Network (CNN) model equipped with activation-map visualization through feature-extraction hooks.

EEG data were preprocessed through EEGLAB [150] with Impulse Response (IIR) bandpass filtering between 0.05 and 80 Hz, Independent Component Analysis (ICA) artifact removal, and baseline correction. EEG was examined in 1-, 5-, and 10-second windows to assess temporal differences in neural responses to GUI quality. The methodology provided an integrated pipeline for evaluating whether eye activity, facial expressions, and EEG signals capture affective distinctions between good and bad GUI designs.

To evaluate whether physiological signals can reliably distinguish good from bad GUI designs, we trained supervised classification models on each modality. Trials with inconsistent user ratings across sessions (differences greater than two points) were removed to ensure label reliability, resulting in less than 10% discarded data. For all remaining trials, we defined two classes ("good" for ratings above 6 and "bad" for ratings below 4) and trained modality-specific models opti-

mized for the characteristics of each signal. Facial expression data were processed using a fine-tuned CNN architecture [147] originally designed for audiovisual affect recognition, relying on grayscale face crops, spatial normalization, and data augmentation. Eye-tracking signals were modeled in two ways: pupil-dilation time series were classified using kNN models across multiple temporal windows (1, 5, and 10 s), while fixation sequences were treated as spatiotemporal inputs to recurrent architectures (Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU)). EEG signals were transformed into spectral representations using FFT, extracting θ, α, β, and γ bands and computing five standard features: Hjorth activity [87], mobility [87], complexity [87], spectral entropy [88], and signal energy [89]. These features were then classified using SVM and kNN models, which are suitable for high-dimensional and low-sample EEG data. This multimodal pipeline enabled a systematic analysis of which physiological channels and temporal windows provide the strongest discriminative power for implicit GUI quality assessment.

Results

The subjective ratings demonstrated clear distinctions between good and bad GUI designs (Figure 2.12). Distributions of overall rating, valence, and arousal differed significantly between design categories. For bad designs, all rating distributions fell significantly below their medians, while for good designs, all ratings fell significantly above their medians. For example, good GUI ratings in session one yielded a $z = 11.90,\ p < .001$ and a moderate effect size $r = .47$ $(n = 640)$. Ratings for bad designs remained consistent across sessions, while ratings for good designs showed some variability. Cronbach's alpha values indicated acceptable to good inter-rater consistency across all measures, and Spearman correlation coefficients between sessions ranged from .57 to .76, indicating strong temporal stability. A comparison between participants' ratings and the original LabintheWild dataset showed a strong correlation (Pearson $r(38) = 0.97,\ p < .0001$), validating

the relevance of the original aesthetic judgments to contemporary participants (Figure 2.14). The valence–arousal distributions for good and bad GUIs further revealed clear separation, with bad designs clustering in low-valence and low-arousal regions, and good designs clustering in high-valence and high-arousal regions (Figure 2.13).

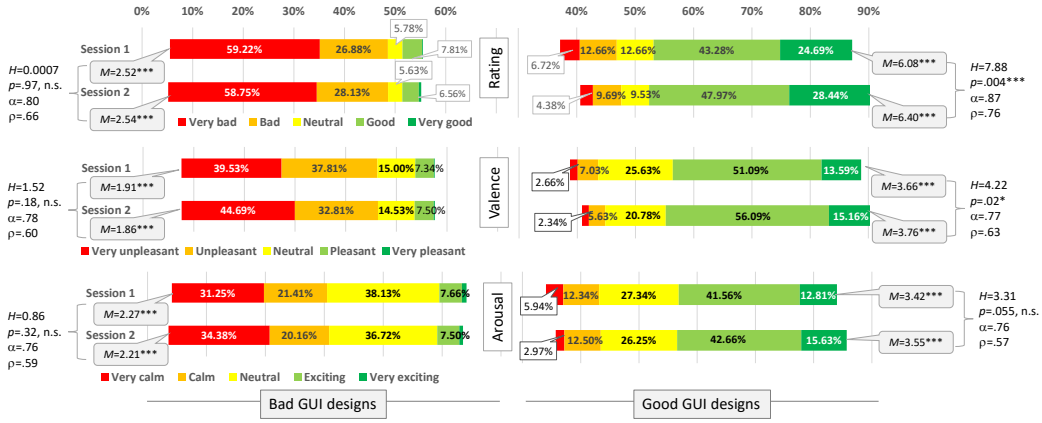


Figure 2.12 Distribution of participants' responses for rating, valence, and arousal associated with **bad** and **good** GUI designs.

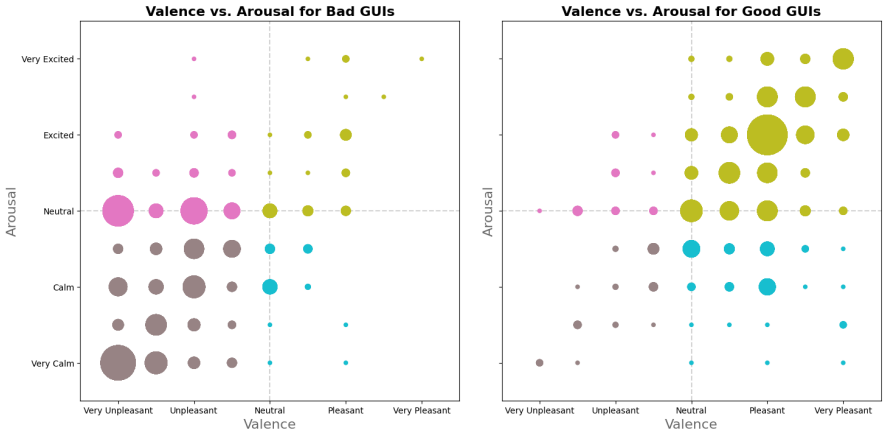
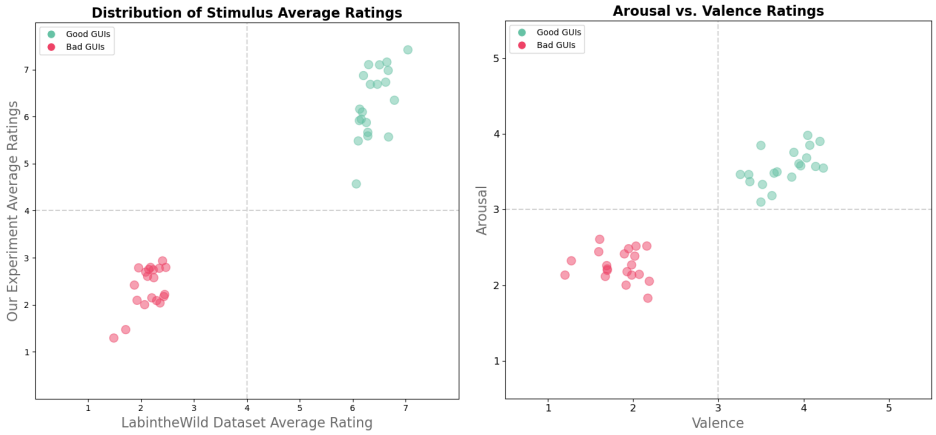


Figure 2.13 Distribution of samples in the valence–arousal plane for **bad** (left) and **good** (right) GUI designs. The radius of each circle represents the relative sample density, while colors denote values associated with each quadrant.

Eye activity analysis revealed distinct early differences in pupil dilation pat-



(a) Comparison of user ratings.

(b) GUI designs in the valence-arousal plane.

Figure 2.14 Distribution of samples across user ratings (left) and within the valence–arousal plane (right).

terns (Figure 2.15). Within the first second of exposure, pupil size changes differed significantly between good and bad designs ($p < .05$). An initial increase in pupil dilation occurred around 200 ms, followed by a decrease within the first 500 ms when viewing bad designs, suggesting rapid cognitive engagement or strain. This indicates that affective discrimination may be detectable within the first second of exposure. Fixation counts (Figure 2.16), however, did not reliably differentiate between good and bad designs. Fixation heatmaps (Figure 2.17) exhibited stronger symmetry and centralized focus for good designs, consistent with design principles of visual consistency and symmetry [151]. Paired samples t-tests showed significant differences at all durations.

Facial expression analysis showed that CNN activation maps differed significantly between responses to good and bad GUIs (Figure 2.18). Feature map distributions for Conv2 and Conv3 layers showed statistically significant differences, indicating that facial activity contains discriminative information regarding users' affective responses to GUI quality.

EEG data analysis (Figure 2.19 and Figure 2.20) revealed temporal and spec-

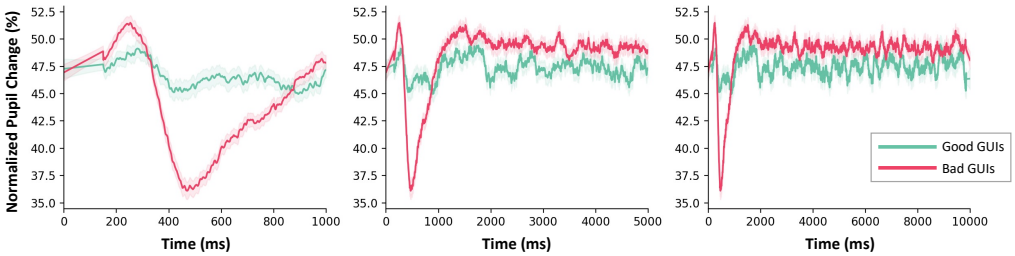


Figure 2.15 Normalized pupil change for **good** and **bad** GUIs across different trial durations. Shaded regions denote 95% confidence intervals.

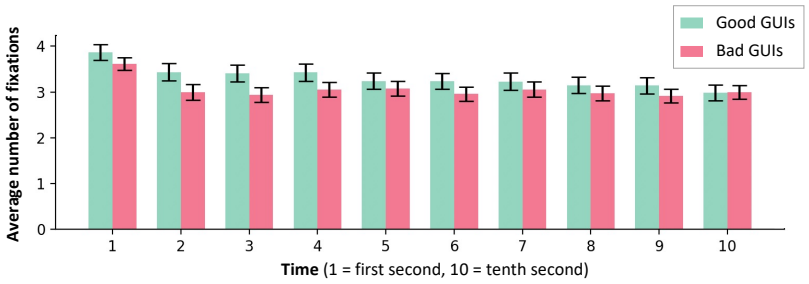


Figure 2.16 Average number of fixations for **good** and **bad** GUI designs across various trial durations. Error bars indicate 95% confidence intervals.

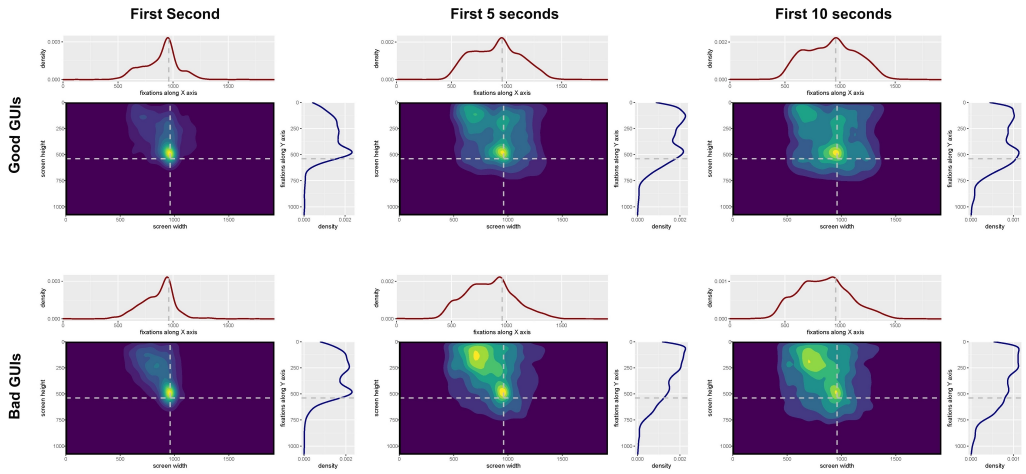
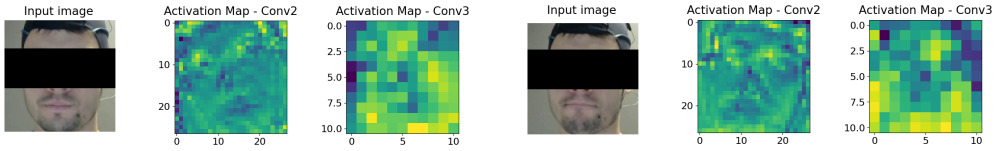


Figure 2.17 Heatmaps of eye fixations for **good** and **bad** GUI designs across different trial durations.

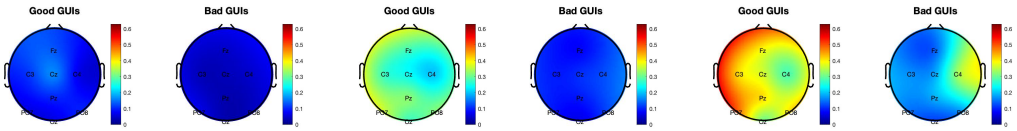
tral differences associated with GUI quality. Although detailed results provide modality-specific distinctions, EEG signals contributed meaningful differences



(a) Good GUI example **(b)** Bad GUI example

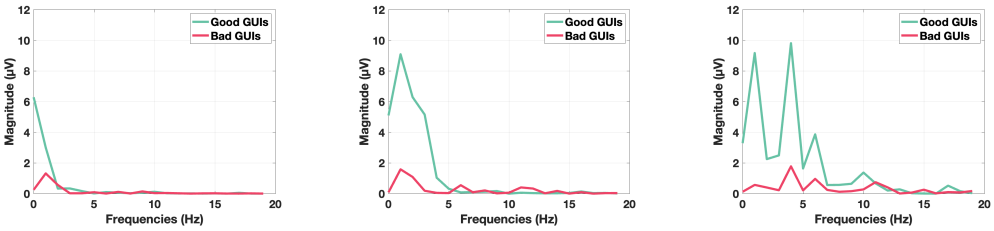
Figure 2.18 Visualization of facial expression activation maps during the viewing of GUI designs.

between conditions and supported the multimodal analysis presented in the paper.



(a) First second **(b)** First 5 seconds **(c)** First 10 seconds

Figure 2.19 Brain topographic maps illustrating the distribution of power across different trial durations.



(a) First second **(b)** First 5 seconds **(c)** First 10 seconds

Figure 2.20 EEG spectral activity across different trial durations.

Collectively, these results indicate that subjective ratings, early visual responses, facial expressions, and EEG activity all contain evidence that users respond affectively to GUI quality. Each modality exhibits its own performance characteristics depending on the temporal window of analysis.

Classification performance varied across modalities, confirming that each physiological signal exhibits its own performance sweet spot. Facial expressions

Modality	Accuracy	AUC	Optimal window	Model
Facial expressions	**74%**	**74%**	10 s	CNN
Pupil dilation	**73%**	**71%**	5 s	7-NN
EEG (β band)	67%	67%	5 s	1-NN
Fixations	54%	53%	5 s	GRU

Table 2.9 Summary of best-performing classification models across different modalities.

produced the strongest results, with the CNN reaching 72% accuracy (71% AUC) in the first 5 seconds and up to 74% accuracy (74% AUC) over the full 10 second exposure. Pupil dilation also proved highly discriminative: a 7-NN classifier achieved 73% accuracy (71% AUC) using only the first 5 seconds of data. EEG-based models performed moderately well, with the best configuration, a 1-NN classifier using the β band, achieving 67% accuracy and AUC. In contrast, fixation-sequence models based on recurrent networks performed poorly (approximately 54% accuracy), indicating that gaze trajectories alone do not reliably differentiate GUI quality. Table 2.9 summarizes the best-performing models for each modality and highlights the temporal windows that yielded optimal results.

Discussion and Limitations

The findings demonstrate that users exhibit clear affective responses to GUI designs, measurable across subjective ratings, eye activity, facial expressions, and EEG signals. Physiological signals revealed reliable early distinctions between good and bad designs, indicating that implicit affective evaluation is feasible. The early pupil dilation differences, observable within the first 500 ms, suggest that cognitive and emotional responses to GUI quality occur rapidly and automatically, supporting the possibility of efficient real-time affective evaluation. Fixation heatmaps further revealed that good designs induce more structured and symmetrical visual scanning patterns, reinforcing established design principles regarding layout clarity and symmetry [151]. Facial expression models captured

differences not visible to human observers, demonstrating that microexpression analysis provides a sensitive and implicit indicator of user experience. EEG results contributed additional evidence that neural dynamics differ between good and bad GUI exposure.

Despite these promising findings, the study includes several limitations. The dataset used consisted of static GUI screenshots rather than interactive interfaces, which may limit ecological validity since affective responses during active use may differ from passive viewing. The exposure duration of 10 seconds was fixed, which may not reflect natural browsing patterns. Although the LabintheWild dataset provided ground truth judgments, it is a decade old, raising questions about temporal relevance, although correlations with current ratings mitigate this concern. Physiological recordings were performed in a controlled laboratory environment, which may not generalize to real-world settings where noise and device variability can affect signal quality. The sample size of 29 participants, while adequate for a within-subjects design, may limit generalizability across demographic groups or cultures, especially since aesthetic preferences can vary with psychological and demographic factors [105, 106, 126]. The study did not include multimodal fusion models, leaving open questions about how best to integrate physiological data for affective GUI evaluation.

Future Work

Several promising directions emerge from this work. A natural extension is to move beyond static screenshots and investigate affective responses during interaction with fully functional and dynamic GUIs. Studying affect under realistic tasks such as navigation, form completion, and error recovery would improve ecological validity and allow the analysis of how aesthetics, usability, and affect co-evolve over time. Future studies could also vary exposure duration and interaction complexity to better reflect real-world usage patterns. Expanding the participant pool to include more diverse cultural, demographic, and expertise backgrounds

would further clarify how individual differences modulate affective responses to interface design.

From a modeling perspective, future work should explore multimodal fusion strategies that integrate eye activity, facial expressions, and EEG signals within a unified framework, potentially leveraging late or hybrid fusion to accommodate modality-specific temporal characteristics. Personalized or adaptive models that account for individual baselines and preferences may improve robustness and predictive power. Finally, translating these findings into practical design tools such as real-time affect-aware prototyping systems or implicit evaluation dashboards could enable designers to receive rapid and non-intrusive feedback early in the design lifecycle, bridging the gap between affective computing research and everyday HCI practice.

Conclusion

The study provides strong evidence that good and bad GUI designs elicit distinguishable affective responses measurable through eye activity, facial expressions, and EEG signals. Subjective ratings validated the original dataset categorization and revealed consistent distinctions between design groups. Eye activity revealed immediate differentiation, with pupil dilation changes emerging within the first 500 ms, suggesting that affective appraisal of GUI design occurs rapidly and automatically. Fixation patterns and heatmaps aligned with established design principles, while facial expression analysis showed that subtle emotional cues can be detected through CNN feature activation maps. EEG analysis further demonstrated that brain activity differs across design quality categories.

Together, these results indicate that implicit physiological evaluation of GUI designs is feasible, and that each modality exhibits an optimal temporal window for effective discrimination. This suggests that affective evaluation tools can be integrated into early stages of design, providing rapid, non intrusive feedback without requiring explicit user input. The research contributes a foundation for

building multimodal affect aware evaluation systems and opens pathways for future work in interface design, affective computing, and HCI.

Chapter 3

Decoding Attention

3.1 Introduction

Understanding how humans allocate attention is central to advancing HCI research. Attention manifests through observable behavioral and physiological signals that reveal cognitive processes otherwise hidden from direct inspection. As technology increasingly mediates how people search, learn, and make decisions, decoding attentional processes becomes essential not only for evaluating user experience but also for building adaptive and intelligent systems that respond to users' cognitive demands.

In digital information environments such as SERPs, attention is a scarce and valuable resource. Organic search results, advertisements, and interactive modules compete simultaneously for the user's focus, shaping decision-making and influencing both user satisfaction and commercial outcomes. Accurate modeling of visual attention therefore serves a dual purpose: it informs the design of interfaces that efficiently guide the user's perceptual flow, and it supports the development of scalable computational models capable of approximating human gaze behavior using accessible interaction data such as mouse movements.

A unifying motivation in this part of the thesis is the need for objective, con-

tinuous, and scalable decoding of attentional processes. Traditional self-reports, click logs, or aggregated behavioral metrics lack temporal precision and provide only indirect insights into moment-to-moment focus. In contrast, behavioral and physiological sensing—particularly eye and mouse tracking—offers rich, time-resolved data for inferring how users distribute their attention in complex visual environments. The studies presented in this chapter address these challenges by combining controlled experimental designs with machine learning methods that transform implicit user behavior into interpretable measures of attention allocation. This line of work contributes toward the development of attention-aware systems capable of supporting more adaptive interfaces and human-centered computational models. This chapter addresses **RQ2** by examining whether user attention can be inferred from behavioral signals, particularly mouse movements, in web search environments.

3.2 A Versatile Dataset of Mouse and Eye Movements on Search Engine Results Pages

We present AdSERP as a comprehensive dataset designed to support research on user attention and purchasing behavior within SERPs. Prior studies have often employed mouse interaction data as a scalable behavioral surrogate while relying on post-task self-reported labels as ground truth, an approach that can suffer from inaccuracy and response bias. To overcome these limitations, we employ eye-tracking technology to derive an objective and continuous measure of visual attention. The dataset consists of 2,776 transactional queries issued on Google SERPs by 47 participants and includes: (1) HTML source files with associated CSS and images; (2) rendered SERP screenshots; (3) eye-tracking recordings; (4) mouse interaction logs; (5) bounding boxes for both direct display and organic advertisements; and (6) scripts to facilitate data preprocessing. In addition, we describe the dataset in detail and report baseline classification experiments intended

to illustrate its utility and to encourage future research directions.

Motivation

Understanding how users allocate attention on SERPs is essential for optimizing design, relevance modeling, advertisement placement, and the general structure of SERP interfaces. Designers and researchers have long relied on eye tracking to measure human visual attention on SERPs, since gaze patterns provide a direct indicator of what users notice, inspect, and consider relevant when performing web search tasks [152, 153, 154, 155]. Despite its value, eye tracking requires specialized hardware and controlled laboratory settings, which limits scalability and makes large-scale deployment impractical. Consequently, mouse movements have been widely adopted as a scalable proxy for gaze, especially for large online experiments, because they can be collected unobtrusively and at very low cost while users browse in natural environments [156, 157, 158, 159, 107].

Earlier work attempting to decode user attention from mouse data on SERPs has depended heavily on self-reported ground-truth labels or post-task questionnaires, which are prone to inaccuracy and bias [160, 161, 162]. Post-hoc responses do not reliably represent continuous attention over time nor capture fast, subtle moments of visual engagement. This gap reinforces the need for datasets that provide objective, fine-grained, continuous ground-truth measurements for attention, ideally through eye fixations.

Many existing datasets contain mouse movement logs but lack corresponding eye-tracking information, or contain only small samples of eye and mouse data that were not publicly shared. Some contain screenshots of SERPs but no programmatic AOI segmentations or HTML source code, making large-scale semantic analysis difficult. Because of these constraints, the field has lacked a comprehensive dataset pairing dense eye and mouse movement recordings with detailed SERP-level structural information, especially for transactional queries that play a central role in commercial search and online purchasing behavior.

This study addresses these gaps by introducing AdSERP, a large, publicly available in-lab dataset containing synchronized mouse trajectories, eye movements, SERP screenshots, HTML source code, advertisement bounding boxes, and preprocessing scripts for AOI computation. Its primary motivation is to enable the research community to model and understand user attention with objective labels, to revisit prior assumptions about mouse–eye coordination, and to build predictive models that can infer attention from inexpensive behavioral signals on real SERPs. The dataset is especially valuable because it focuses on transactional search queries, where ad positions and user goal-directed behavior directly affect commercial outcomes. Grounding machine learning models in eye-based attention labels helps researchers overcome the issues of bias, noise, and unreliability associated with previous self-reported approaches.

Related Work

Mouse movements have served as a rich behavioral signal revealing many aspects of user behavior, including demographics [163], identity [164], satisfaction [165, 166], preferences [167, 168], experience [169, 170], decision making [171, 172], future activity [173, 174], and visual attention [160, 159]. Some studies explored whether mouse trajectories can predict gaze patterns. Smucker et al. [175] examined mouse movements during relevance judgments but found that for simple tasks they did not reliably indicate attention. Other studies proposed models incorporating clicks, attention, and satisfaction to improve behavioral prediction [165].

Certain prior work demonstrated that mouse signals can outperform traditional click-based models in predicting user engagement with SERP modules [156] and can enhance click prediction and relevance estimation [176]. Horizontal mouse movement patterns have been linked to reading direction and engagement [177, 178]. More recent work has explored LSTM-style models for predicting user behavior from cursor data [179].

Alignment between mouse and eye movements remains inconsistent across studies. Johnson et al. [180] showed high correspondence, but only because participants were explicitly instructed to move the mouse in synchrony with their gaze. Other studies using naturalistic search tasks showed substantial divergence between gaze and cursor position, with eye and mouse behaviors often decoupled [157, 181, 182, 183].

Few datasets systematically combine eye tracking and mouse tracking on SERPs. Datasets by Guo and Agichtein [184] and Huang et al. [157] contained both modalities but were small and not publicly released. Mao et al. [185] included both signals but did not release the dataset. Liu et al. [186] provided a dataset with eye and mouse data but lacked AOI segmentation and HTML files. Chen et al. [187] collected mouse information for satisfaction prediction but did not release data. The Attentive Cursor Dataset [160, 159] was larger and publicly shared but lacked eye movement data and relied on self-reported attention labels.

No existing publicly available dataset offers transactional queries, large scale logs, both eye and mouse data, HTML for AOI derivation, and programmatic AOI identification simultaneously (Table 3.1). AdSERP is the first dataset to meet all these requirements.

Ref.	Transactional	Large	Eye	Mouse	AOIs	HTML	Availability
[184]	✗	✗	✓	✓	✗	✗	✗
[157]	✗	✓	✓	✓	✗	✗	✗
[185]	✓	✗	✓	✓	✗	✗	✗
[186]	✗	✓	✓	✓	✗	✗	✓
[187]	✗	✓	✗	✓	✗	✗	✗
[160, 159]	✓	✓	✗	✓	✓	✓	✓
Ours	✓	✓	✓	✓	✓	✓	✓

Table 3.1 Summary of prior studies, ordered by publication year. The *Large* column denotes studies that analyzed more than 1,000 log files.

Methodology

The dataset was constructed through a controlled laboratory experiment with 47 participants. Demographics include ages 19–44, gender distribution of 27 male and 20 female participants, and English proficiency at B2 or above. Participants had varied shopping preferences, with many preferring Amazon, and all reported normal or corrected vision. Figure 3.1 illustrates the demographics of our participants.

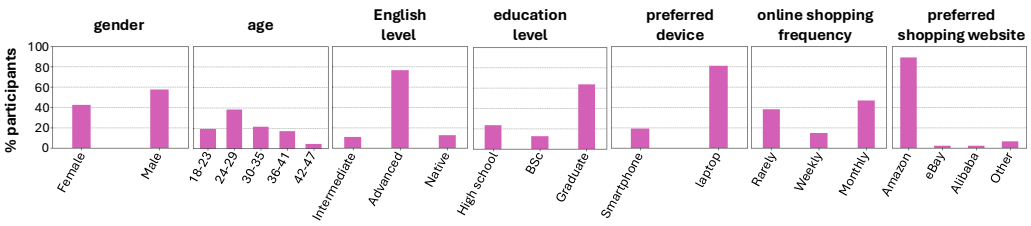


Figure 3.1 Demographic characteristics of the user sample ($N = 47$ participants).

Transactional search queries were generated from product titles in the Amazon Product Reviews dataset [188] by prepending the word "buy" to each title. This yielded 3,020 unique English-language SERPs containing mixtures of organic results and direct display advertisements. SERP layouts followed consistent patterns, with organic ads appearing at the top or bottom of SERPs and direct display ads appearing at top-left or top-right. Figure 3.2 shows examples of these layouts.

The SERPs were batched into 302 sets of 10 SERPs, with each participant viewing exactly one unique set. Two additional batches were used for familiarization and were the same for all participants. The apparatus included a 17-inch Dell LCD monitor at 1280×1024 px and 60 Hz, a Gazepoint GP3 HD eye tracker at 150 Hz, and a Dell MS116 mouse. SERPs were displayed full-screen. A custom web-based application guided participants through each trial.

As illustrated in Figure 3.3, each trial began with a product title and query. Participants were instructed to imagine they intended to purchase the item, after which they viewed the SERP for up to one minute. They could scroll freely, inspect

results, and click an option. After clicking, they could confirm their selection or continue browsing for another minute.

The dataset includes HTML files with embedded CSS and images, full-page SERP screenshots, XML logs describing trial metadata, and programmatically derived AOI bounding boxes for ads. AOI coordinates were extracted using Selenium WebDriver[1] and stored as (x, y, w, h) tuples. Eye tracking was recorded both as pupil size streams (timestamp t, coordinates x, y, and pupil diameters) and as fixations (timestamp t, coordinates x, y, duration d). Mouse events included timestamp t, cursor position x, y, event type e, and DOM XPath. Preprocessing scripts include visualization tools for eye and mouse data such as heatmaps, trajectory plots, and combined overlays following prior work [160]. Additionally, screen recordings (Figure 3.5) are provided for every trial. Moreover, we incorporated a set of mouse-movement visualizations (Figure 3.4) that were proposed in a previous study by Arapakis and Leiva [160]. Of the 2,820 trials, 2,776 remained after removing malformed logs. These represent combinations of ad placements. The dataset also includes scripts for computing AOIs beyond advertisements, enabling custom segmentation of SERP components.

To evaluate whether mouse movements can reliably predict visual attention on SERPs, we conducted baseline classification experiments that map cursor trajectories to fixation-derived attention labels. Each trial was assigned a binary label indicating whether users attended to a specific advertisement AOI, computed via the fixation-duration based attention metric. A simple attention metric was introduced:

$$\text{Attention}_{\text{trial}} = \frac{\sum \text{Fixation Duration}_{\text{AOI}}}{\sum \text{Fixation Duration}_{\text{total}}} \tag{3.1}$$

where $\text{Attention}_{\text{trial}}$ is a value ranging from 0 to 1. This measure can also be transformed into a binary indicator to assess whether an AOI received relatively

[1] https://www.selenium.dev/

sufficient attention or not:

$$\text{label}_{\text{trial}} = \begin{cases} 1 & \text{if Attention}_{\text{trial}} > \tau, \\ 0 & \text{otherwise.} \end{cases} \tag{3.2}$$

where τ is a user-defined cutoff value, typically derived from the collected data and influenced by the AOI type as well as the input duration. For example, Figure 3.6 illustrates how attention values vary across different exposure lengths for DD ads. Thus, the choice of threshold is a tunable parameter that can be adjusted according to the specific context or objectives of the study. In our study, τ was set to the median value of the attention scores for each experiment.

We tested multiple input representations, including Mouse2Vec [174] embeddings and raw mouse time series concatenated with AOI bounding box coordinates, and examined how predictive performance varies with different temporal windows (first 5, 10, 15, and 20 seconds of interaction). Three model families were implemented: SVM, kNN, and GRU-based recurrent networks, using stratified train and test splits and early stopping for optimization. This methodology enables systematic comparison of traditional machine-learning approaches and sequence-based neural models in predicting fine-grained attentional states from inexpensive cursor data.

Results

The analysis covers 2,776 valid trials. The mean trial duration was 22.16 seconds with a median of 20 seconds. Most trials (97.4%) were completed with a single click, 2.4% with a second click, and 0.2% with three clicks. Left-aligned direct display ads were selected in 8.61% of trials, right-aligned direct display ads in 3.78%, and organic ads in 5.19%. Non-ad clicks constituted 82.42% of selections.

Fixation counts (Figure 3.7) on ad and non-ad areas were computed across layout conditions, showing that direct display ads attract substantial attention.

Organic ads attracted less attention. Fixation duration (Figure 3.8), an indicator of attentional depth [189], showed a right-skewed distribution consistent with prior findings [190]. The dataset contained an average of 84.42 fixations per trial and an average fixation duration of approximately 218 ms.

Heatmaps (Figure 3.9) comparing eye and mouse distributions across SERP layouts demonstrated differing degrees of alignment between modalities. Mutual information between mouse and eye behavior remained low across all layouts: 0.02 for SERPs with right-aligned direct display and organic ads, 0.01 for SERPs with only organic ads, and 0.06 for SERPs with left-aligned direct display and organic ads. ANOVA testing revealed no statistically significant differences. KL divergence values were 19.89, 21.90, and 17.27 respectively, with significant differences across conditions. Post-hoc tests showed greater divergence between eye and mouse for SERPs with only organic ads.

Distances between gaze and cursor positions (Figure 3.10) were also analyzed. Euclidean distance averaged 372.89 px, substantially larger than findings reported by Huang et al. [158]. Divergence was greater in the vertical axis than the horizontal axis.

Classification results (Table 3.2 and Table 3.3) show that early-stage mouse trajectories contain strong signals for inferring visual attention. GRU models consistently outperformed SVM and kNN baselines across all durations and input types, achieving the best performance when using only the first five seconds of cursor data. For SERPs with organic ads, the GRU reached an F1 score of 93%, while performance for direct-display ads was lower but still strong (F1 = 73%). SVM and kNN models performed noticeably worse, especially when relying on Mouse2Vec embeddings, likely due to a domain mismatch between SERP cursor behavior and the embedding model's pretraining corpus. These results demonstrate that attention can be predicted accurately and very early in the interaction, well before users complete the task, highlighting the potential of sequence-based models for scalable, low-cost attention inference.

Model type	Duration (s)	F_1 score	AUC
SVM	5	0.67	0.71
k-NN	5	0.57	0.57
GRU	5	0.70	0.78
SVM	10	0.63	0.67
k-NN	10	0.54	0.55
GRU	10	0.73	0.82
SVM	15	0.59	0.59
k-NN	15	0.54	0.54
GRU	15	0.72	0.77
SVM	20	0.60	0.60
k-NN	20	0.55	0.55
GRU	20	0.69	0.74

Table 3.2 Performance of baseline models under different input representations and mouse trajectory durations on SERPs containing various combinations of DD ads.

Model type	Duration (s)	F_1 score	AUC
SVM	5	0.89	0.94
k-NN	5	0.86	0.84
GRU	5	0.93	0.97
SVM	10	0.92	0.95
k-NN	10	0.85	0.82
GRU	10	0.93	0.97
SVM	15	0.88	0.90
k-NN	15	0.77	0.76
GRU	15	0.90	0.96
SVM	20	0.86	0.91
k-NN	20	0.71	0.71
GRU	20	0.89	0.95

Table 3.3 Performance of baseline models under different input representations and mouse trajectory durations on SERPs containing only organic ads.

Discussion and Limitations

The AdSERP dataset provides the first large-scale combination of eye tracking, mouse trajectories, SERP HTML, and AOI segmentation for transactional queries.

The results reveal expected and unexpected properties of mouse–eye coordination. Direct display ads attract strong visual attention. Organic ads receive less attention, which may affect commercial outcomes.

Cursor–gaze agreement is limited and varies by layout. Mouse signals often diverge significantly from eye fixations, especially vertically, confirming earlier findings [157, 182, 183] at a larger scale and with objective attention labels.

Early mouse movements proved highly predictive of visual attention, with GRU models outperforming SVM and kNN approaches. Mouse2Vec performed less effectively due to domain mismatch.

The dataset has limitations. It includes only desktop browsing with a physical mouse and does not represent trackpad behavior or mobile interactions. Mobile devices lack cursor data, which limits comparability. Most participants preferred Amazon, which may influence browsing strategies. Queries were fixed rather than user-generated. The one-minute trial limit may restrict observations of extended search behavior. The dataset focuses exclusively on transactional queries, leaving out informational or navigational ones.

Future Work

The AdSERP dataset opens several promising directions for future research on attention modeling and user behavior on SERPs. A natural extension is to move beyond advertisement-level AOIs toward finer-grained (see Figure 3.11), element-level attention prediction, such as isolating prices, images, ratings, or textual descriptions within ads and organic results. Leveraging the provided HTML and DOM structure enables precise semantic segmentation, allowing researchers to study how visual attention aligns with content relevance, visual saliency, and commercial attributes. Additionally, combining gaze data with computer vision techniques applied to SERP screenshots could support multimodal models that jointly reason over layout, visual features, and interaction signals.

Future modeling work could also explore representation learning approaches

that better adapt to the SERP domain. For instance, pretraining or fine-tuning cursor embeddings such as Mouse2Vec on AdSERP data may improve generalization compared to off-the-shelf embeddings trained on non-search tasks. More expressive architectures, including transformer-based sequence models or multimodal fusion networks integrating mouse, eye, and visual inputs, may further enhance early attention prediction. Beyond modeling, extending data collection to include different input devices (e.g., trackpads), mobile interactions, user-generated queries, and non-transactional search intents would broaden the ecological validity of findings and enable comparative studies across devices, intents, and interaction paradigms. Together, these directions position AdSERP as a foundation for advancing attention-aware systems in search and advertising research.

Conclusion

The paper introduces AdSERP, a comprehensive publicly available dataset for analyzing user attention on SERPs through combined eye and mouse tracking (see Figure 3.12 for illustrative examples). Its contributions include complete HTML source files, fine-grained AOI boundaries, SERP screenshots, and continuous eye-based attention signals. Analyses provide new insights about user engagement, including strong attention capture by direct display ads and limited agreement between mouse and eye behavior. Baseline models show that visual attention can be predicted from early mouse movements with high accuracy, particularly for organic ads.

This dataset enables research into computer vision on SERP screenshots, DOM-level semantic analysis, AOI modeling, and improved representation learning for mouse trajectories. It provides a foundation for attention-aware models in search, advertising, and user modeling. The dataset has been publicly released [2] in full and is intended as a versatile resource for the Information Retrieval (IR) and HCI

[2] https://zenodo.org/records/15236546

communities.

3.3 AdSight: Scalable and Accurate Quantification of User Attention in Multi-Slot Sponsored Search

Modern SERPs feature increasingly complex layouts in which multiple elements compete for user attention. Accurate attention modeling is essential for improving web design and computational advertising, as attention-related metrics can guide advertisement placement and monetization strategies. We present AdSight, a scalable and precise approach for estimating user attention in multi-slot environments such as SERPs by leveraging mouse cursor trajectories. AdSight is built upon a novel Transformer-based sequence-to-sequence (Seq2Seq) architecture in which the encoder processes embeddings of cursor movements and the decoder integrates slot-specific features, allowing reliable attention prediction across diverse SERP configurations. We evaluate the proposed method on two machine learning tasks: (1) regression, aimed at predicting fixation durations and counts; and (2) classification, focused on identifying whether specific slot types attracted attention. The results demonstrate that AdSight achieves highly accurate attention predictions, providing valuable insights for both research and practical applications.

Motivation

The increasing complexity of modern SERPs has transformed how user attention must be understood, modeled, and quantified. Early web pages were largely static and text based, where saliency was primarily driven by textual relevance. As web interfaces evolved to include images, logos, banners, interactive elements, and dynamic content, attention became influenced not only by content but also by

layout and visual hierarchy [191]. Contemporary SERPs now contain heterogeneous elements such as organic results, direct display advertisements, featured snippets, and side panels, all competing simultaneously for user attention. This competition makes accurate attention modeling essential for effective web design and computational advertising.

The importance of attention modeling is particularly pronounced in sponsored search, where advertising revenue underpins a large portion of search engine business models. Ads contribute roughly one third of advertising revenue for major search engines, amounting to tens of billions of dollars annually [192]. However, complex SERP layouts introduce challenges such as good abandonments, where users obtain sufficient information without clicking, complicating traditional metrics like clicks and impressions [193, 194]. As a result, clicks alone no longer provide a reliable proxy for user attention or ad effectiveness.

Accurate attention prediction enables more strategic ad placement, reduces intrusive advertising experiences, and improves monetization efficiency. It also supports emerging paradigms such as pay-per-attention auction schemes, which charge advertisers based on whether users actually noticed ads rather than whether they clicked them [195]. Despite the promise of these approaches, existing attention measurement methods often rely on eye tracking, which is expensive, intrusive, and impractical at scale. Consequently, scalable alternatives are required to support large-scale deployment.

Mouse cursor tracking has emerged as a viable proxy for visual attention, as cursor movements correlate with gaze behavior on SERPs and can be collected unobtrusively at scale [196, 197, 157, 158, 159, 198]. Nevertheless, existing mouse-based approaches typically analyze single ads in isolation, fail to account for multiple slots simultaneously, or rely on simplified models that do not generalize well across different SERP layouts. Moreover, many approaches lack objective ground truth labels derived from eye tracking.

This study introduces AdSight to address these challenges. AdSight is de-

signed to quantify user attention in multi-slot sponsored search environments using mouse cursor trajectories while maintaining scalability and accuracy. By leveraging a Transformer-based Seq2Seq architecture that integrates cursor dynamics with slot-specific metadata, the method aims to predict attention across heterogeneous SERP layouts. The motivation is to provide a low-cost, scalable alternative to eye tracking that delivers precise attention estimates suitable for real-world advertising systems.

Related Work

Research on mouse cursor analysis has a long history across experimental psychology, cognitive science, and information retrieval. Mouse trajectories have been used to study motor control, decision making, consumer choice, and cognitive processes during everyday tasks such as e-shopping [199, 200, 201, 202, 203, 204]. In information retrieval, eye tracking has traditionally been used to analyze user behavior and relevance judgments [205, 206, 207, 208, 209], but its reliance on specialized hardware limits scalability.

Mouse tracking offers a cost-effective alternative that can be deployed in natural browsing environments without additional equipment [210]. Prior work has demonstrated that mouse movements can serve as a reasonable proxy for gaze, particularly on SERPs [156, 197, 157, 158, 159, 198]. Mouse data have been used to inform usability testing [211], predict user engagement and intent [212, 213, 214], detect frustration and abandonment [193, 215, 216], and model open-ended behaviors [179].

Within user modeling, early approaches relied on coarse-grained cursor features to infer user interest [217, 218], while later work employed fine-grained features that improved prediction accuracy [213, 219]. Mouse data have been used to predict search satisfaction and success [184, 220], and to extend click models for improved relevance estimation [221, 158]. Mouse cursor motifs have also been introduced to capture recurring behavioral patterns associated with

attention and relevance [222, 220].

Several studies specifically investigated visual attention inference from mouse movements. Mouse trajectories have been used to predict reading behavior [223], hesitation patterns [224], and whether users are looking at the content pointed to by the cursor [225]. Research on direct displays has shown that mouse data can support attention measurement and enable pay-per-attention advertising models [196, 195]. However, these studies typically focus on individual elements rather than jointly modeling attention across multiple slots.

In sponsored search, earlier scanning models assumed linear top-to-bottom exploration, but this assumption no longer holds due to heterogeneous SERP components [226, 227, 228]. More recent models incorporate ancillary modules and non-linear scanning behavior [229, 221, 198]. Liu et al. [230] showed that users often skim SERPs before reading, proposing a two-stage examination model that can be predicted from mouse movements.

The closest prior work to AdSight is by Arapakis and Leiva [231], who developed diagnostic technologies to measure attention to direct displays using mouse data [196, 231, 195]. However, these approaches analyze one slot at a time and do not support simultaneous attention prediction across multiple slot types. The present work differs by introducing a Transformer-based framework that models attention to multiple slots jointly and by using eye tracking to provide objective ground truth labels for training and evaluation.

Methodology

The study is based on AdSERP dataset [232], as described in section 3.2. Attention prediction is formulated as two learning problems. The regression task predicts Total Fixation Time (TFT) and Total Fixation Count (TFC) for each individual slot. The classification task predicts whether a user *noticed* each of the four slot categories.

Mouse cursor data are represented using both temporal and visual encodings.

For time-series representations, each mouse event is encoded using normalized cursor coordinates $(x, y) \in [0, 1]$, the time spent at that location, a categorical feature indicating the slot type at the cursor position (outside$= -1$, direct-top$= 0$, direct-right$= 1$, organic-top$= 2$, organic-bottom$= 3$), and a normalized sequence index in $[0, 1]$. Cursor events are asynchronous, and consecutive duplicate coordinates are removed. Transformer-based models operate on variable-length sequences, while LSTM-based models use fixed-length sequences of 250 timesteps, obtained via truncation or zero-padding.

Visual representations (see Figure 3.13) of cursor behavior include heatmaps, color-coded trajectories, and slot-aware trajectory images, which are processed using Vision Transformer (ViT) embeddings. These embeddings serve as alternatives to time-series cursor representations within the same modeling framework.

Each slot is described using metadata consisting of its normalized center coordinates (x_c, y_c) and its categorical slot type. To improve spatial granularity when the cursor is outside predefined slots, *auxiliary slots* are introduced in the main vertical results area and the right column using heuristic placement. The number of auxiliary slots is tuned experimentally, with the best performance achieved using three auxiliary slots.

Ground truth attention labels are derived from eye-tracking data. For regression, TFT is computed as the sum of fixation durations within a slot, and TFC as the total number of fixations within the slot; fixations shorter than 100 ms are excluded. For classification, fixations within each slot are clustered using a density-based procedure, and clusters are characterized by TFT and TFC. A slot is labeled as noticed if its cluster exceeds the median TFT and TFC thresholds; a slot category is labeled positive if at least one slot within that category is noticed. The resulting fixation rates are 42% for direct-top, 46% for direct-right, 44% for organic-top, and 29% for organic-bottom slots.

AdSight employs an encoder–decoder Transformer architecture (see Figure 3.14) to map cursor behavior to slot-level attention predictions. Cursor

features are first projected into a latent space of dimension $l \in \{16, 32, 64\}$ using a shared multilayer perceptron, and then processed using either a two-layer bidirectional LSTM or a two-layer Transformer encoder with two attention heads and a feed-forward dimension of 512. For visual inputs, a frozen ViT backbone is used, with only the final layers trained and the hidden width scaled by a factor $k \in \{2, 4, 8\}$. Slot metadata are embedded via a shared network, concatenated, and projected back to dimension l. The decoder attends over the encoded cursor representations and produces slot-level predictions via a shared Multilayer Perceptron (MLP) readout. ReLU activations ensure non-negative outputs for fixation time and count.

Models are trained using either mean squared error loss or listwise rank loss for regression, and binary cross-entropy loss for classification. Losses from auxiliary slots are weighted by a factor $\alpha \in [0, 1]$, with best results obtained using a moderate value of $\alpha = 0.33$. Training uses the Adam optimizer with learning rates in $\{10^{-3}, 10^{-4}, 10^{-5}\}$, batch sizes in $\{16, 32, 64\}$, and a maximum of 100 epochs with early stopping after 25 epochs. Hyperparameters are optimized via Bayesian optimization with 3-fold cross-validation. Baselines replace the Seq2Seq decoder with an MLP readout sized to the maximum of 14 slots per SERP and filtered to the actual slots present. Regression performance is evaluated using mean squared error and Normalized Discounted Cumulative Gain (NDCG), while classification is evaluated using area under the ROC curve (AUC) and F1 score.

Results

Across all cursor representations and embedding strategies, Seq2Seq models consistently outperform MLP readout baselines. The best-performing configuration uses time-series cursor data with a Transformer encoder and a Seq2Seq decoder. For TFT prediction, this model achieves an average MSE of 2.86 ± 0.02, corresponding to an average absolute error of approximately 1.69 seconds per slot, and an NDCG of 96.07 ± 0.04. For TFC prediction, the same configuration achieves an

MSE of 50.07 ± 0.04 and an NDCG of 96.36 ± 0.05. Models optimized using list-wise rank loss consistently achieve higher NDCG scores than those trained with mean squared error. Wilcoxon signed-rank tests confirm that Seq2Seq models significantly outperform MLP baselines across all evaluated settings ($p < 0.05$).

In the classification task, the Seq2Seq Transformer with time-series cursor embeddings achieves the best overall performance, with an average AUC of 81.24 ± 0.06 and an average F1 score of 76.25 ± 0.05. Performance varies across slot categories: direct-top ads achieve an AUC of 80.79 and F1 of 73.90, direct-right ads an AUC of 81.72 and F1 of 75.07, organic-top ads an AUC of 71.87 and F1 of 67.27, and organic-bottom ads an AUC of 85.85 and F1 of 82.57. Time-series representations perform best for direct-display ads, while visual cursor representations are particularly competitive for organic-bottom slots. Previously proposed baselines based on BiLSTM and ResNet architectures are consistently outperformed.

Ablation studies demonstrate the importance of slot metadata, with the removal of slot type causing the largest performance degradation. Among cursor features, the normalized sequence index is particularly influential. The inclusion of auxiliary slots yields consistent improvements, with performance increasing as the number of auxiliary slots grows from zero and peaking at three auxiliary slots. Moderate weighting of auxiliary-slot loss further improves accuracy, confirming the benefit of jointly modeling attention transitions across the entire SERP.

Discussion and Limitations

The results demonstrate that user attention on complex SERPs can be accurately quantified using mouse cursor trajectories when combined with appropriate modeling techniques. AdSight shows that attention prediction is feasible at scale and can achieve precision comparable to eye tracking for both regression and classification tasks. The use of a Transformer-based Seq2Seq architecture allows the model to capture complex relationships between cursor behavior and SERP

structure.

The findings highlight the importance of modeling multiple slots jointly rather than in isolation. Slot metadata and auxiliary slots enable the model to generalize across diverse layouts and improve robustness. The superior performance of Seq2Seq models over multilayer perceptron baselines underscores the value of sequence-aware architectures for attention modeling.

The study has limitations. The dataset was collected in a laboratory setting, which may differ from real-world browsing behavior. The focus on desktop SERPs limits applicability to mobile environments, where cursor data may not be available. While mouse movements correlate with gaze, they do not perfectly capture visual attention. The approach also relies on eye tracking data for training, which may limit applicability where such data are unavailable.

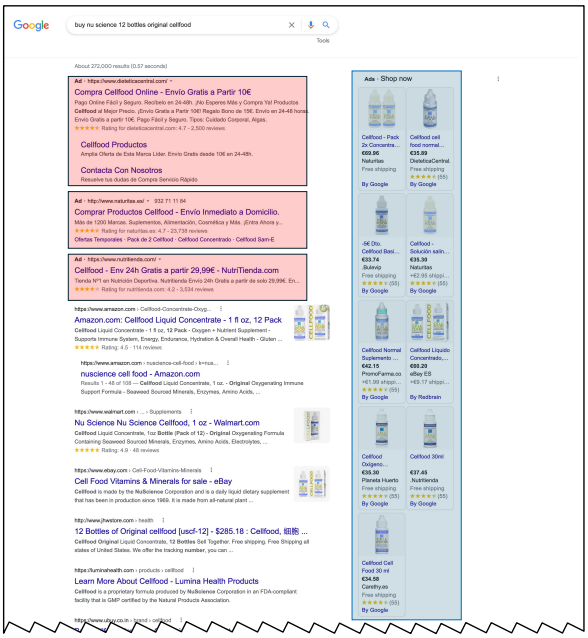
Future Work

Several promising directions emerge from this work. First, future research should validate AdSight in real-world, large-scale deployment settings to assess robustness under naturalistic browsing behavior, diverse user populations, and evolving SERP designs. Extending the approach to mobile and touch-based interfaces represents another important avenue, requiring alternative interaction signals such as touch gestures or scrolling dynamics in the absence of mouse cursor data. Additionally, incorporating richer contextual signals such as query intent, task complexity, or temporal session information could further improve attention prediction and personalization. From a modeling perspective, exploring multimodal architectures that jointly leverage cursor data, page content embeddings, and lightweight visual features may enhance generalization across heterogeneous layouts. Finally, reducing or eliminating the reliance on eye-tracking supervision through weakly supervised or self-supervised learning would increase practicality and facilitate broader adoption of attention-aware advertising and search systems building on AdSight.

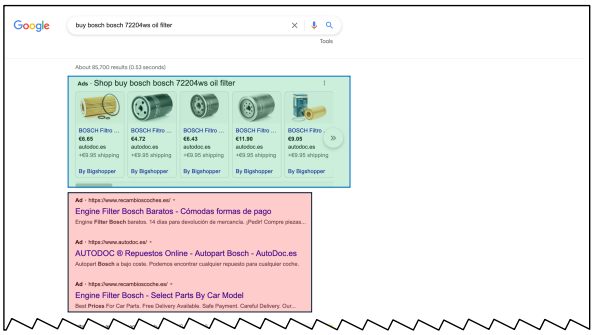
Conclusion

We introduced AdSight, a scalable and accurate method for quantifying user attention in multi-slot sponsored search environments. By leveraging mouse cursor trajectories, slot metadata, and a Transformer-based Seq2Seq architecture, the approach achieves strong performance in predicting fixation time, fixation count, and slot noticeability. AdSight consistently outperforms traditional multilayer perceptron baselines and demonstrates robustness across diverse SERP layouts.

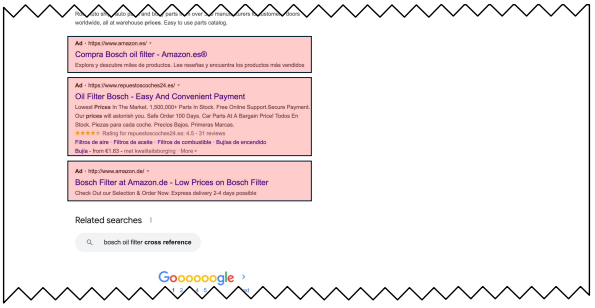
The work advances attention modeling for sponsored search by enabling low-cost, scalable inference of user attention suitable for real-world deployment. It provides a foundation for improved ad placement, pay-per-attention models, and future research on attention-aware search interfaces. The findings indicate that mouse-based attention modeling combined with modern sequence modeling techniques can serve as a practical alternative to eye tracking in complex web environments.



(a) Organic and right-aligned DD ads



(b) Organic and left-aligned DD ads



(c) Organic ads

Figure 3.2 Examples of SERPs containing advertisements. Organic ads are highlighted in light red, left-aligned DD (direct-display) ads in light green, and right-aligned DD (direct-display) ads in light blue.

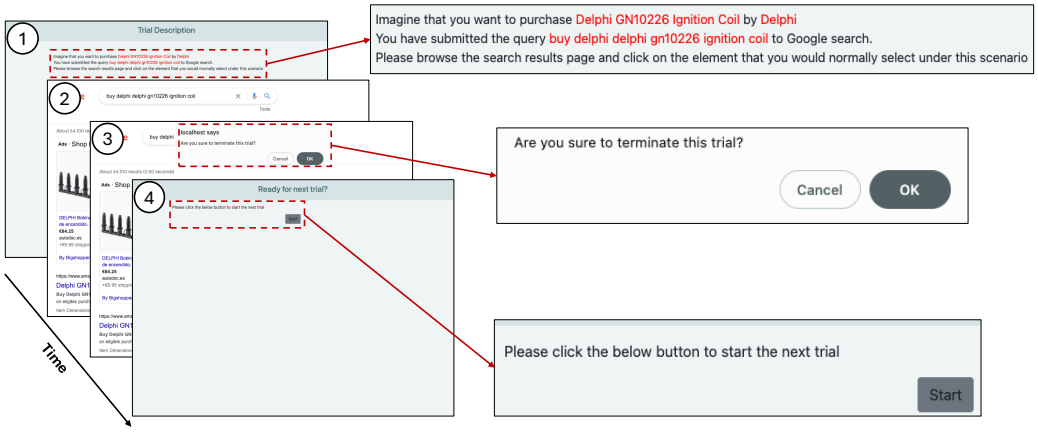


Figure 3.3 Example of an experimental trial. Participants are first presented with information about the product and its associated query (1). They then examine the SERP (2). Upon clicking an item, participants may choose to terminate the current trial or continue (3). Finally, they proceed to the next trial when ready (4).

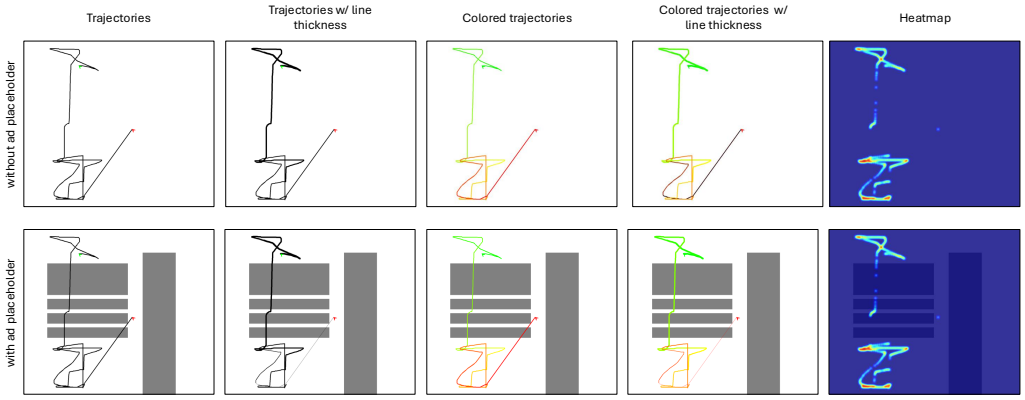


Figure 3.4 Different types of mouse movement visualizations, as proposed by Arapakis and Leiva [160], which are also provided in this work.

Figure 3.5 Example frames from a screen recording. Gaze locations and cursor positions are overlaid on the displayed page.

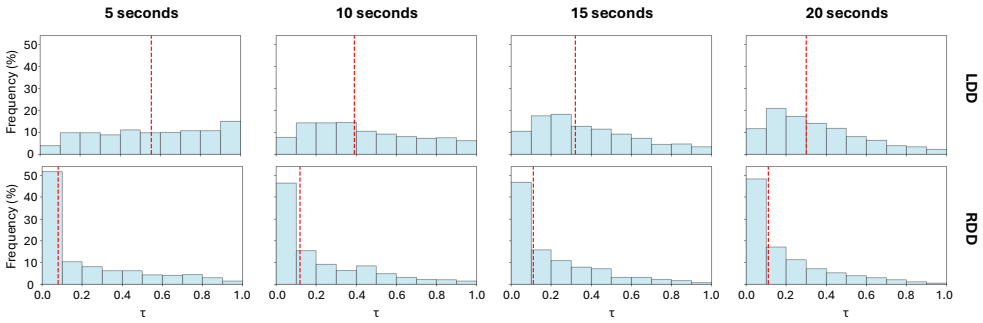


Figure 3.6 Distribution of attention on RDD and LDD advertisements across all trials. Dashed red lines indicate the median values.

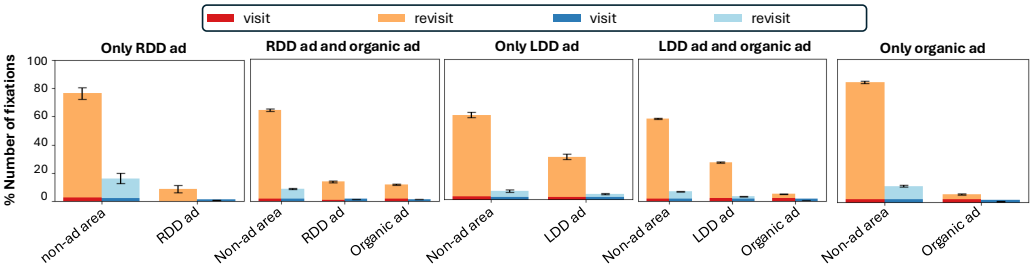


Figure 3.7 Percentage distribution of fixations on advertisement and non-advertisement areas. Error bars indicate the standard error of the mean. RDD denotes right-aligned direct display ads, and LDD denotes left-aligned direct display ads. *Visit* refers to the first fixation within an advertisement, whereas *revisit* denotes a subsequent fixation on the same advertisement after fixating elsewhere.

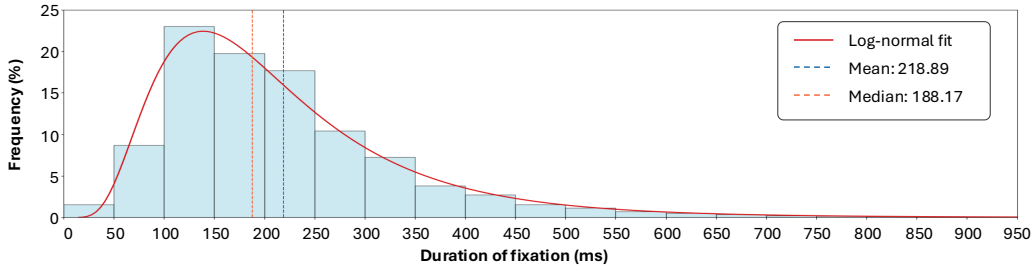
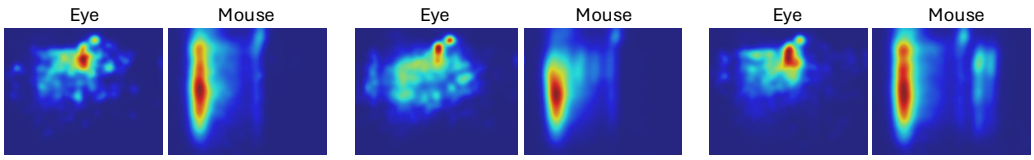


Figure 3.8 Distribution of fixation durations across all participants.



(a) Only organic ads **(b)** LDD and organic ads **(c)** RDD and organic ads

Figure 3.9 Heatmaps of eye gaze and mouse cursor locations within a 1280×1024 px viewport for different advertisement combinations.

Figure 3.10 Coordination between mouse and eye movements across all SERPs.

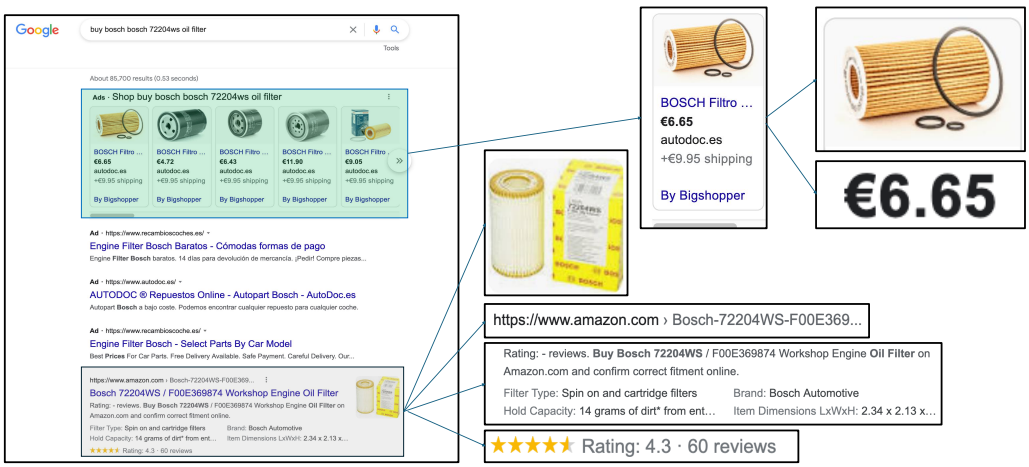


Figure 3.11 Examples of fine-grained areas of interest (AOIs) with associated metadata. For DD (direct-display) ads, image content and price information can be extracted. For organic items, the corresponding links, descriptions, ratings, and review information can also be extracted.

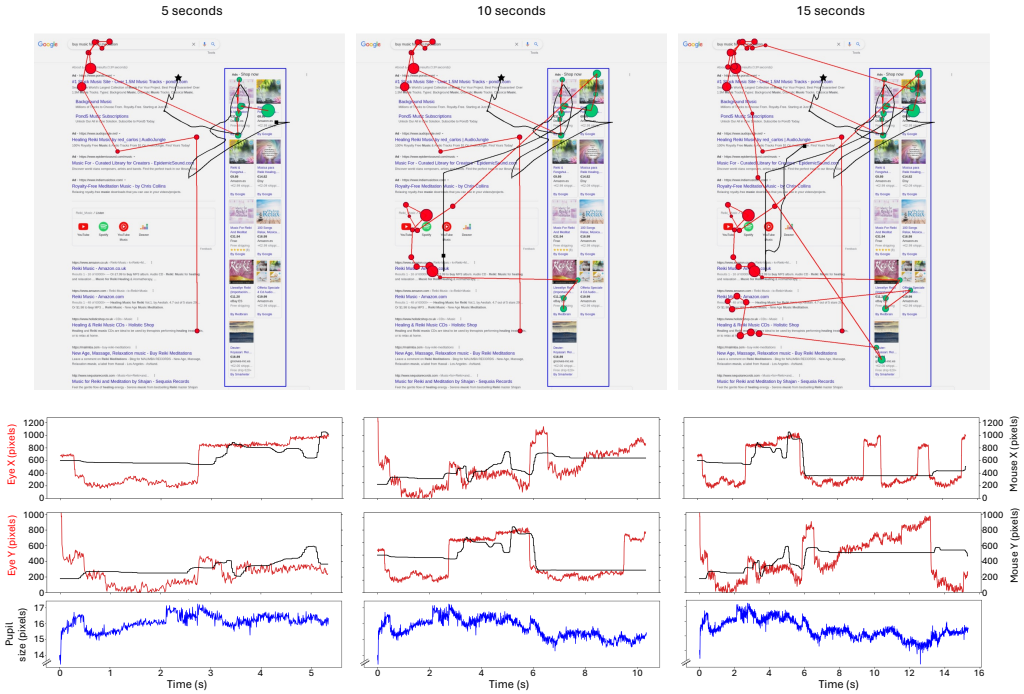
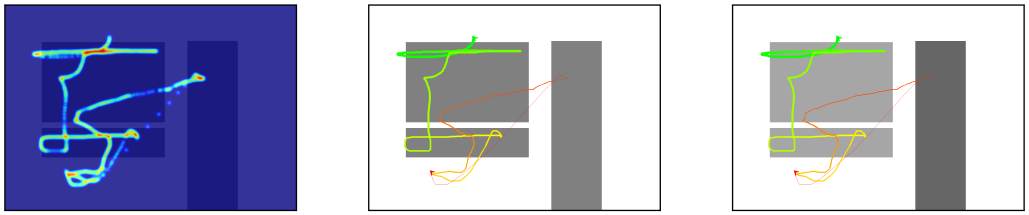
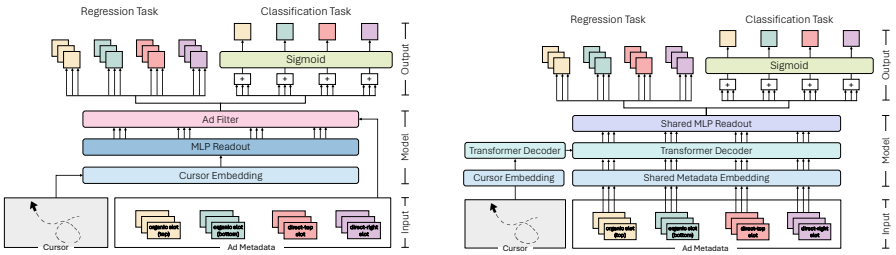


Figure 3.12 Visualization of mouse cursor trajectories and eye fixations from a single participant's trial at three time points: 5, 10, and 15 seconds after trial onset. The boundary of the right-aligned advertisement is outlined in blue. Fixations within the advertisement region are shown as green circles, while fixations outside this region are shown as red circles; circle size reflects fixation duration. Mouse movement is depicted by a continuous black trajectory, with a star indicating the starting position and a square marking the endpoint. The lower panels display the corresponding mouse and eye screen coordinates, along with pupil size, for each time interval.



(a) Heatmap **(b)** Color trajectories **(c)** Color trajectories and slot-specific color

Figure 3.13 Visual encodings of mouse movement data used to train the ViT models, all incorporating slot placeholders.



(a) Baseline workflow **(b)** Seq2Seq workflow

Figure 3.14 Comparison between the MLP baseline (Figure 3.14a) and the Seq2Seq approach (Figure 3.14b). The baseline maps cursor data into a dense latent representation via an MLP configured for the maximum number of slots per document, combined with metadata-driven filtering. In contrast, the Seq2Seq model embeds cursor data and metadata into a shared latent space, employing a Transformer Encoder–Decoder to capture sequential dependencies. Regression heads predict quantities (TFT/TFC), while classification aggregates type-wise scores and applies a sigmoid to obtain probabilities.

Chapter 4

Decoding Expertise

4.1 Introduction

Understanding how humans develop expertise is central to advancing HCI research, especially in domains where perceptual and motor skills shape performance. Expertise reveals itself through behavioral and physiological signals that uncover internal cognitive processes not directly observable. As digital technologies increasingly mediate professional training and complex decision-making, decoding expertise becomes crucial for evaluating learning progression and for building intelligent systems that provide meaningful, individualized feedback.

In professional training contexts such as medicine, expertise emerges through the refinement of perceptual strategies, sensorimotor coordination, and decision-making under pressure. Subtle variations in gaze patterns, head movements, and physiological effort can reliably distinguish novices from experts, offering objective markers of skill acquisition that go beyond traditional instructor-based evaluations. Such multimodal signals capture how experienced practitioners perceive and prioritize information, how they anticipate events, and how efficiently they interact within simulated clinical environments.

A unifying motivation in this chapter is the need for objective, continuous, and

scalable measures of expertise. Conventional assessments are subjective, episodic, and often insensitive to the nuanced behavioral signatures that indicate growing mastery. In contrast, multimodal sensing—through eye tracking, head tracking, and physiological monitoring—provides fine-grained data enabling the inference of underlying skill levels. The study in this chapter combine controlled simulation environments with machine learning approaches that transform implicit behavior into interpretable indicators of expertise. Taken together, this research advances the development of expertise-aware systems capable of supporting personalized feedback, adaptive training, and more precise evaluation of professional competencies. This chapter addresses **RQ3** by exploring whether multimodal behavioral signals can be used to assess expertise in complex tasks.

4.2 Assessing Medical Training Skills via Eye and Head Movements

We investigated eye and head movement patterns to better understand skill acquisition in clinical environments. Twenty-four practitioners took part in simulated childbirth training sessions. We extracted several key measures, including pupillary response rate, fixation duration, and angular velocity. The analysis shows that eye- and head-tracking data can reliably distinguish between trained and untrained practitioners, particularly during labor-related tasks. For instance, head-based features achieved an F1 score of 85% and an AUC of 86%, while pupil-based features reached an F1 score of 77% and an AUC of 85%. These findings establish a foundation for computational approaches to implicit skill assessment and training support in clinical contexts, using affordable eye-tracking glasses as a complement to conventional evaluation methods such as subjective scoring.

Motivation

Simulation-based training plays an essential role in preparing medical professionals for high stakes clinical procedures, offering a safe environment to practice technical and non-technical skills while receiving immediate feedback from instructors. However, traditional assessment methods remain subjective, time-consuming, and potentially biased. To complement subjective evaluation, biosignals such as eye and head movements provide an unobtrusive and objective means of assessing performance that can be recorded using lightweight wearable devices like eye-tracking glasses, which are increasingly used in clinical research environments [233, 234, 235, 236, 237]. Although eye and head tracking have been explored extensively in many areas of simulation training, their ability to assess skill acquisition in specific medical contexts remains insufficiently understood. In particular, the relative impact of various metrics such as fixations, saccades, pupil dilation, angular velocity, and cumulative rotation remains unclear, especially across different phases of complex procedures [238]. This gap is especially relevant for breech delivery, a challenging and high-risk childbirth scenario that has not previously been analyzed using these signals.

The study aims to investigate whether eye and head movement data can reliably indicate the development of expertise during simulation-based breech delivery training. By analyzing two training sessions for each practitioner and relating eye- and head-movement-derived metrics to expert-assigned skill scores, we seek to determine whether these signals can differentiate trained from untrained practitioners. The motivation further includes enabling scalable, objective, and implicit skill assessment, providing medical educators with tools that support more efficient quality assurance and personalized feedback using commodity-level sensing technology. The study also introduces a new dataset [1] including synchronized signals, video, annotated time segments, and post-session skill scores, intended to serve as a foundation for future computational models of

[1] https://zenodo.org/records/15163456

medical skill assessment.

Related Work

The paper reviews prior work on eye tracking, head tracking, and their applications to clinical training, highlighting both potential and limitations. Eye-tracking glasses have become widely used in medical research to analyze practitioner behavior during procedures, capturing key metrics such as fixations, saccades, and pupil size to infer attention and decision-making [239, 240]. Prior research has demonstrated the value of fixation-related metrics for evaluating cognitive load, visual focus, and expertise development in procedures such as epidural blocks [241] and needle insertion [242], as well as complex tasks requiring situational awareness [243, 244]. However, most eye-tracking studies rely on AOI-based measures, which are difficult to automate in dynamic environments due to constant viewpoint changes caused by head motion, making AOI annotation particularly challenging when using head-mounted eye trackers [245, 246, 247].

Beyond fixations, cognition-related metrics such as the Task-Evoked Pupillary Response (TEPR) have been associated with cognitive load and clinical performance [248, 249], while Eye Blink Rate (EBR) correlates with cognitive flexibility and abilities related to learning [250, 251]. Saccadic metrics, including amplitude, velocity, and acceleration, have been linked to visual attention and expertise in medical tasks such as neonatal intubation [252]. Yet, studies show mixed results regarding whether training reliably affects saccadic behavior [253].

Head movements have also been used to assess surgical skill, with metrics such as head acceleration helping distinguish novices from experts during laparoscopic suturing [254]. Angular velocity and cumulative rotation offer further behavioral indicators that can reflect efficiency and concentration during tasks [255, 256].

Despite extensive research, the literature lacks a detailed comparison between eye- and head-tracking metrics across different phases of obstetric emergency training, and no prior work examines breech delivery in particular. This study

positions itself to fill these gaps by analyzing multiple eye- and head-movement features across five expert-defined segments of breech delivery, testing explicit hypotheses derived from prior literature, and training machine-learning models to evaluate their predictive power.

Methodology

The study was conducted as a controlled simulation experiment during a full-day obstetric training program held at the Medical Simulation Centre in Ljubljana. As part of this training, participants took part in a simulated breech delivery, a childbirth scenario. The simulation followed a standardized educational protocol using a high-fidelity childbirth manikin and was designed to combine predefined clinical procedures with realistic, moment-to-moment decision making.

Twenty-four practitioners participated, all holding a 6-year medical degree and working in obstetric settings, at various stages of a 5-year specialization program. Each completed two simulation sessions, each approximately seven minutes long, separated by roughly 60 minutes of other training activities. Participants wore Tobii Pro Glasses 2, which recorded eye movements, pupil size, blinks, saccades, and fixations at 100 Hz, as well as six-degree-of-freedom inertial head-motion data, synchronized with a high-definition scene-camera video stream. The simulated delivery room setup included a NOELLE S550 manikin, CTG monitor, infusion pump, medical trolley, and supporting staff, including a midwife and an expert doctor who operated the manikin and provided post-session feedback. Two ceiling-mounted cameras captured panoramic environmental video. Figure 4.1 shows the experimental setup.

Each practitioner's performance was graded by an expert doctor on a 1–5 scale for each of 14 clinical skills covering communication, preparation, judgment, and technical execution. To analyze temporal dynamics, we collaborated with the expert to segment each simulation into five standardized phases (see Figure 4.2): Anamnesis, Vaginal examination, Preparation, Awaiting, and Labor.

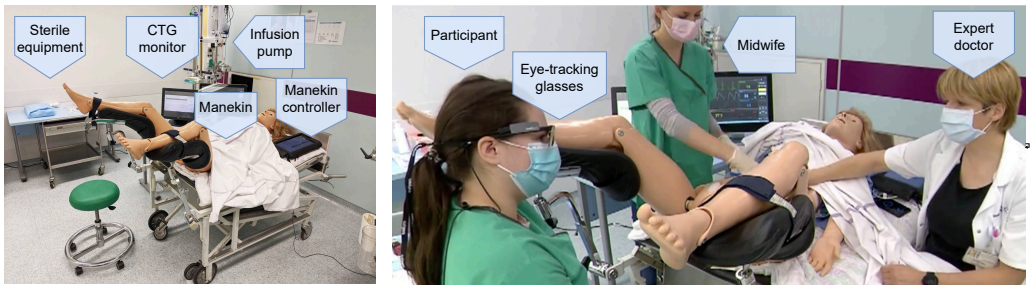


Figure 4.1 Apparatus and delivery room setup. Left: Configuration including a manikin, controller, CTG monitor, infusion pump, and sterile equipment arranged on a trolley. Right: An expert physician providing feedback to a participant following simulation-based training.

These reflect sequences of semantically coherent medical actions and align with typical breech-delivery workflows. All eye and head signals within each segment were normalized to span time 0–1 for comparability.

The analysis quantified fixation count, fixation duration, TEPR (based on min-max-normalized pupil size), EBR (excluding blinks under 100 ms [257, 258]), saccade amplitude, velocity, and acceleration, as well as head angular velocity and cumulative rotation. Paired two-tailed t-tests compared first-session (untrained) and second-session (trained) metrics within each segment.

To investigate predictive power, we built machine-learning classifiers using SVMs with Bayesian-optimized hyperparameters. For each modality (pupil, fixation, saccade, blink, head), the time series were divided into fixed-length sampling windows, from which handcrafted statistical features (Table 4.1) were extracted. Segment codes were appended to inform models of contextual task phase. Labels indicated session number (0 = untrained, 1 = trained). Separate SVMs were trained on each segment and on all segments combined, using 80 percent training, 10 percent validation, and 10 percent testing splits.

Figure 4.2 Example screenshots from a recorded breech delivery video, ordered chronologically from top to bottom across different time segments.

Results

Participants showed a clear improvement in expert-assigned skill scores (Figure 4.3) between sessions, with significantly higher overall ratings in the second session (M = 4.67 vs. M = 4.13; $t(23) = -5.08$, $p < .001$). Time-on-task analysis (Table 4.2) indicated that practitioners completed Anamnesis, Vaginal examination, and Preparation significantly faster in the second session, while taking more time during Labor, suggesting greater deliberation or deeper engagement in the most critical phase.

For fixation-related metrics, neither fixation count (Figure 4.5) nor fixation du-

Modality	Features	Sampling window
Pupil	Timestamp X and Y coordinates Normalized pupil size Segment code	100 data points (1 s) 10751 feat vectors
Fixation	Timestamp X and Y coordinates Duration Segment code	10 successive fixations 3361 feat vectors
Saccade	Timestamp Amplitude Peak velocity Peak acceleration Segment code	10 successive saccades 8729 feat vectors
Blinks	Timestamp Duration Segment code	10 successive blinks 1051 feat vectors
Head	Timestamp Rotational speed in X, Y, Z Segment code	100 data points (1 s) 15107 feat vectors

Table 4.1 List of handcrafted features used for support vector machine (SVM) classifiers.

Segment	Mean		Mdn		SD	
	S1	S2	S1	S2	S1	S2
Anamnesis	56.79	41.67	54.5	42	14.14	12.24
Vaginal exam.	44.08	31.08	40.5	31.5	15.28	13.72
Preparation	81.12	58.04	79.5	56.5	19.11	12.48
Awaiting	105.5	101.12	108.5	106.5	33.29	39.79
Labor	39.04	67.92	30	64.5	16.50	17.89

Table 4.2 Duration (in seconds) of each segment for both training sessions (S1: first session; S2: second session).

ration (Figure 4.6) differed significantly across sessions for any segment. Heatmaps (Figure 4.4) show visually denser fixations during Vaginal examination in the

second session, though statistical tests revealed no reliable differences.

For cognition-related indices, TEPR (Figure 4.7) was significantly higher only during Labor ($t(23) = -2.41$, $p = .017$), indicating increased cognitive load or engagement. EBR (Figure 4.8) showed no significant differences across any segment.

Saccade amplitude (Figure 4.9) and acceleration (Figure 4.11) were significantly higher in the first session during the Awaiting segment (amplitude: $t(23) = 2.95$, $p < .01$; acceleration: $t(23) = 2.78$, $p < .01$), suggesting more abrupt or less controlled eye movements early in training. Saccade velocity (Figure 4.10) did not yield significant differences.

Head-movement metrics revealed the strongest differences. Angular velocity (Figure 4.12) was significantly higher in the first session during Anamnesis ($t(23) = -2.897$, $p < .01$). Cumulative rotation (Figure 4.13) was significantly larger in the first session for Anamnesis, Vaginal examination, and Preparation (all $p < .001$), whereas during Labor cumulative rotation was significantly lower in the first session ($t(23) = -12.08$, $p < .001$), suggesting that trained participants made fewer extraneous head movements.

Machine-learning results (Table 4.3) show that head-based features provided the best discriminative performance. For the Labor segment, head-movement models achieved 85% F1 and 86% AUC, the strongest overall. Eye-movement features were also effective, with pupillary responses achieving 77% F1 and 85% AUC in Labor. Fixations, saccades, and blinks achieved moderate performance, with fixations reaching up to 81% AUC in Preparation. Across modalities, Labor consistently produced the highest predictive accuracy, indicating its importance for automated skill assessment.

Discussion and Limitations

The results demonstrate that while many low-level metrics do not significantly differ across sessions, specific metrics, particularly TEPR during Labor, saccadic

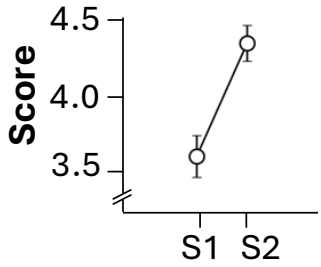


Figure 4.3 Comparison of skill scores across sessions (S1: first session; S2: second session). Error bars indicate the standard error of the mean.

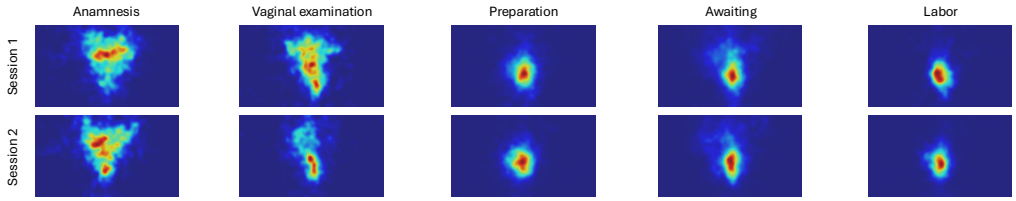


Figure 4.4 Heatmaps of eye fixations.

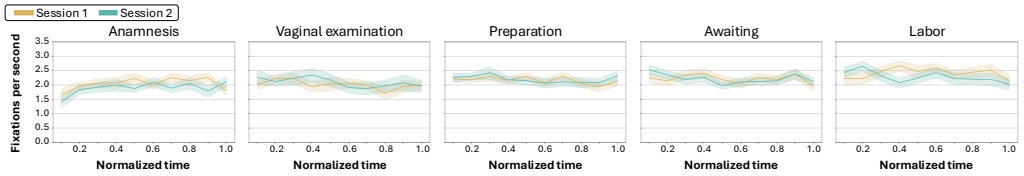


Figure 4.5 Fixation count across different segments. Shaded regions represent the standard error of the mean.

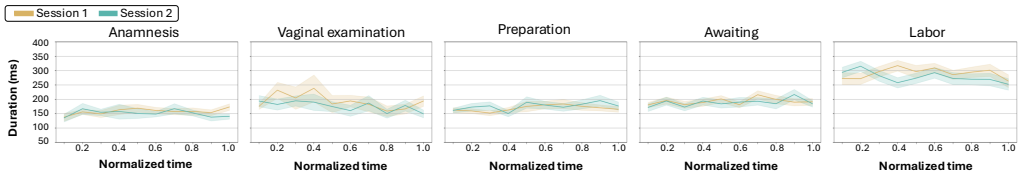


Figure 4.6 Fixation duration across different segments. Shaded regions represent the standard error of the mean.

indices during Awaiting, and multiple head-movement measures, successfully capture behavioral differences between trained and untrained practitioners. Greater pupil dilation and longer Labor durations suggest that trained practitioners engage in more focused decision-making during the most critical phase. The head-

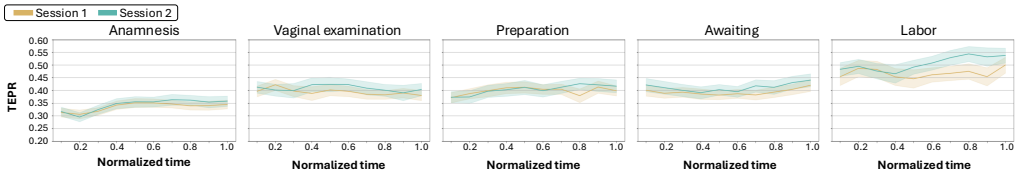


Figure 4.7 Pupil size across different segments. Shaded regions represent the standard error of the mean.

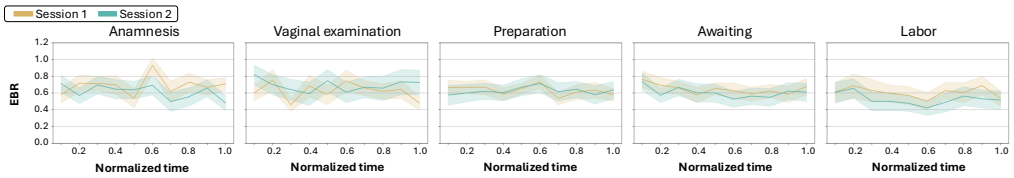


Figure 4.8 Blink rate across different segments. Shaded regions represent the standard error of the mean.

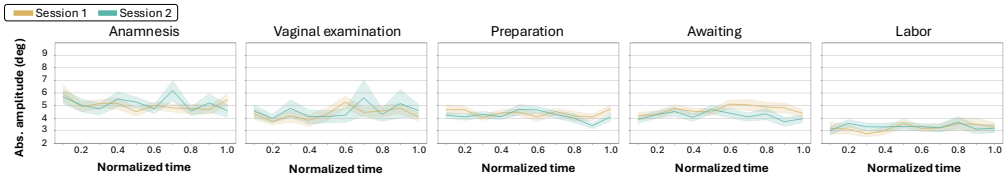


Figure 4.9 Saccade amplitude across different segments. Shaded regions represent the standard error of the mean.

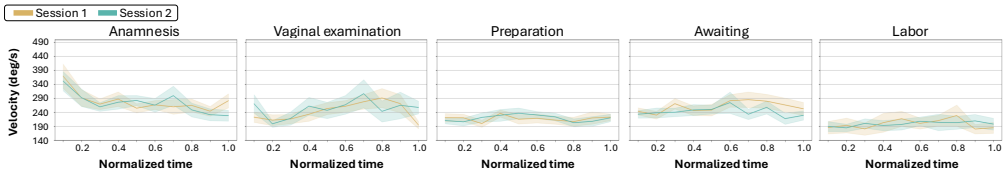


Figure 4.10 Saccade velocity across different segments. Shaded regions represent the standard error of the mean.

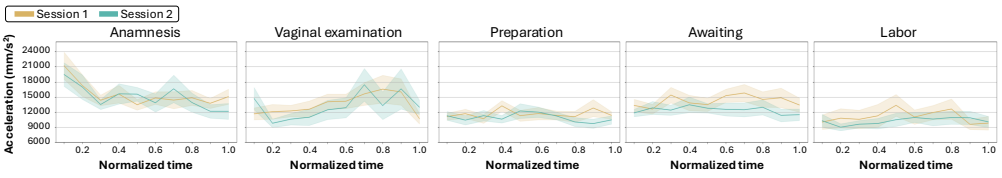


Figure 4.11 Saccade acceleration across different segments. Shaded regions represent the standard error of the mean.

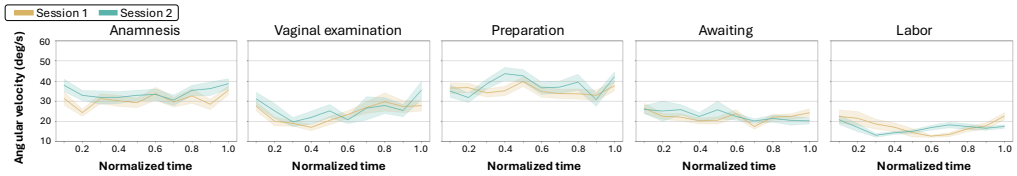


Figure 4.12 Angular velocity of head movements. Shaded regions represent the standard error of the mean.

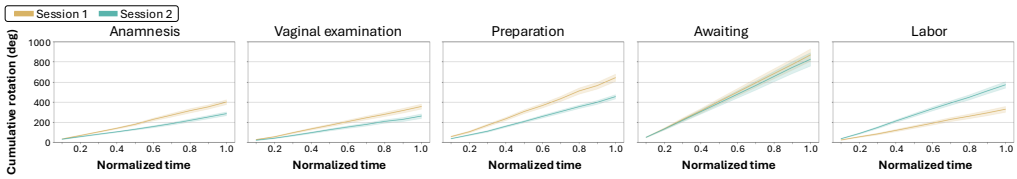


Figure 4.13 Cumulative rotation of head movements. Shaded regions represent the standard error of the mean.

movement findings indicate that trained practitioners exhibit more stable and efficient motion patterns, with reduced unnecessary rotation except during Labor, where more precise movements may be required.

The study confirms that head-movement patterns can be particularly powerful indicators of skill development and may complement eye-movement metrics in future multimodal models. The machine-learning results further show that even without AOI annotations, statistical summaries of gaze and head motion can reliably classify practitioner expertise.

Limitations include the modest sample size of 24, a common issue in medical-simulation research. Time normalization assumes comparable relative progression through segments, though Awaiting displayed variability. Eye-tracking glasses operated at 100 Hz, which is lower than the recommended 200 Hz for precise saccade analysis, though results still proved meaningful. Additionally, RNN and XGBoost models underperformed relative to SVMs, likely due to insufficient data length per time series. Future work may include data augmentation, multimodal fusion, higher-frequency sensors, analysis of specific head-motion types, and application to other clinical tasks.

Future Work

Future work should explore richer multimodal models that integrate eye movements, head movements, and additional physiological signals to improve robustness and generalizability of automated skill assessment. In particular, multimodal fusion strategies, both early and late, may capture complementary aspects of perceptual attention, motor efficiency, and cognitive load that are only partially reflected in individual modalities. Increasing the temporal resolution of eye tracking and incorporating fine-grained head-movement descriptors, such as directional patterns or task-specific motion primitives, could further enhance sensitivity to subtle changes in expertise. Expanding the dataset to include more participants, repeated sessions over longer training trajectories, and varying levels of clinical complexity would also enable the use of data-hungry models such as deep sequence learners and support longitudinal analyses of skill acquisition.

Beyond technical improvements, future research should investigate how inferred expertise signals can be translated into actionable feedback for trainees and instructors. Integrating expertise-aware models into simulation platforms could enable real-time feedback, adaptive scenario difficulty, or targeted debriefing based on detected performance patterns. Additionally, validating the proposed methods across other medical procedures and clinical domains would help establish the external validity of eye- and head-movement-based assessment, ultimately contributing to scalable, objective, and personalized training systems that complement expert judgment rather than replace it.

Conclusion

This study investigated eye- and head-movement signals to assess skill development during breech-delivery simulation training. The results demonstrate that these signals can distinguish between trained and untrained practitioners, especially during the Labor segment, and that head-movement features are particularly

discriminative. The findings suggest that commodity eye-tracking glasses can support implicit, objective, and scalable medical skill assessment, complementing traditional evaluation methods. The provided dataset offers a foundation for future computational models, enabling more efficient, data-driven feedback and reducing the burden on instructors while minimizing bias. The study contributes toward developing automated systems that enhance both medical training and patient safety.

Modality	Segment	Hyperparameters			Adj. F_1	AUC
		kernel	C	γ		
Pupil	Anamnesis	linear	47.35	0.01	0.68	0.74
	Vaginal exam.	linear	12.58	0.01	0.58	0.69
	Preparation	linear	100.0	10.0	0.70	0.79
	Awaiting	RBF	25.21	0.58	0.67	0.74
	Labor	linear	100.0	0.85	**0.77**	**0.85**
	All segments	RBF	19.46	0.40	0.65	0.70
Fixation	Anamnesis	linear	6.64	0.01	0.62	0.64
	Vaginal exam.	RBF	1.70	6.29	0.44	0.54
	Preparation	linear	54.39	0.01	0.62	0.81
	Awaiting	RBF	32.98	0.01	0.59	0.61
	Labor	linear	73.52	0.80	**0.67**	**0.80**
	All segments	RBF	13.50	0.01	0.56	0.58
Saccade	Anamnesis	linear	44.23	10.0	0.59	0.63
	Vaginal exam.	linear	27.37	0.62	0.61	0.64
	Preparation	linear	56.50	0.01	0.63	**0.69**
	Awaiting	RBF	100.0	0.02	0.54	0.56
	Labor	linear	27.36	0.62	**0.64**	**0.69**
	All segments	RBF	93.19	0.01	0.55	0.57
Blink	Anamnesis	RBF	49.38	0.32	0.42	0.52
	Vaginal exam.	RBF	0.64	9.89	0.42	0.44
	Preparation	RBF	100.0	7.33	0.52	0.48
	Awaiting	polynomial	21.58	2.83	0.39	0.58
	Labor	RBF	15.92	0.03	**0.66**	**0.84**
	All segments	RBF	27.36	0.62	0.46	0.52
Head	Anamnesis	linear	77.60	9.97	0.63	0.69
	Vaginal exam.	linear	81.10	0.01	0.65	0.72
	Preparation	linear	100.0	0.56	0.77	0.84
	Awaiting	RBF	0.1	0.49	0.47	0.48
	Labor	linear	31.76	10.0	**0.85**	**0.86**
	All segments	RBF	100.0	0.15	0.55	0.58

Table 4.3 Classification performance results. The best result is highlighted in bold, and the second-best result is underlined.

Chapter 5

Applications and Toolkits

5.1 Introduction

The increasing sophistication of multimodal sensing in HCI has enabled researchers to decode complex cognitive states by combining neurophysiological and behavioral signals. However, these methodological advances also introduce practical challenges. Multimodal data are difficult to collect, synchronize, and validate, particularly when several recording devices, platforms, and operating systems are involved. This chapter presents two tools developed to address these challenges: *Gustav* [74] and *Thalamus* [259], which together form the methodological foundation supporting the empirical studies of this thesis.

Both tools emerged from the need to make multimodal research more reproducible, accessible, and efficient. While the earlier chapters focused on decoding affect, attention, and expertise from real-world experiments, this chapter turns to the instrumentation layer that made such experiments technically possible. *Gustav* provides precise temporal synchronization across devices and computers, ensuring data integrity in multimodal experiments. *Thalamus* extends this functionality by offering a simulation environment for multimodal signal generation that allows researchers to prototype, test, and refine their experimental setups

before collecting real data. Together, they contribute to a broader vision of open, scalable, and reliable multimodal experimentation in HCI. This chapter addresses **RQ4** by introducing tools that support reproducible and scalable multimodal experimentation.

5.2 Gustav: Cross-device Cross-computer Synchronization of Sensory Signals

Achieving precise temporal alignment of behavioral and physiological data streams collected from multiple devices, and in some cases across different computers, has long been a challenge in HCI, neuroscience, psychology, and related fields. Existing solutions typically rely on dedicated hardware, specialized software, or post hoc signal alignment algorithms. However, these approaches are often vendor dependent, lack generality, or are difficult to deploy in real-world settings. We propose a simple yet highly effective alternative that instruments the stimulus presentation software by inserting supervisory event-related timestamps, followed by a post-processing phase applied to the recorded log files. Building on this mechanism, we introduce Gustav, a framework for coordinating the acquisition of sensory signals across devices and machines. Gustav guarantees that all recorded signals are precisely aligned with the duration of each experimental condition, achieving millisecond-level accuracy. The system is released as open source software for public use.

Motivation

The collection of behavioral and physiological signals is central to understanding human cognition, interaction, and motor control across fields such as human computer interaction, neuroscience, and psychology. Signals such as brain activity, heart rate variability, keystrokes, mouse movements, and eye movements

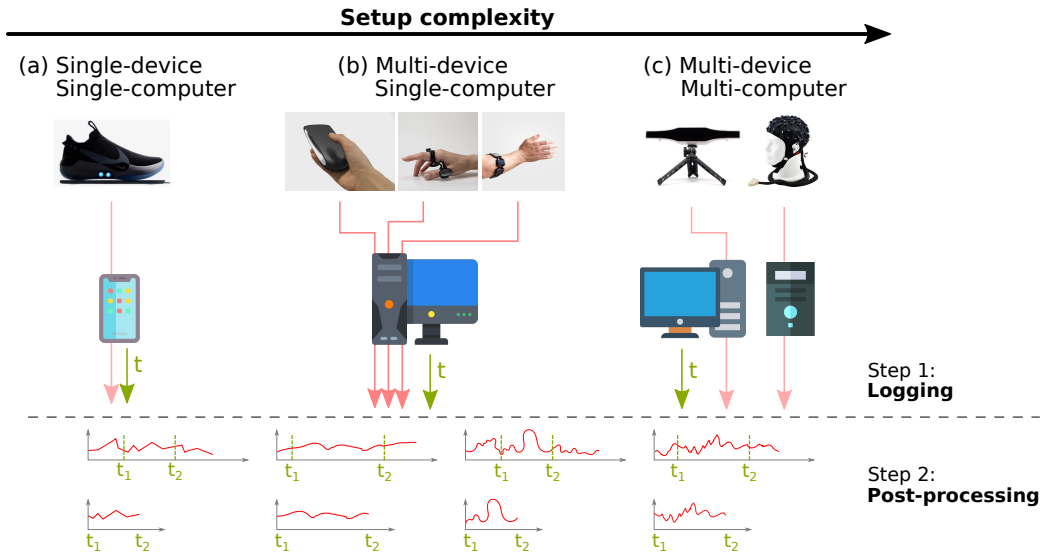


Figure 5.1 *Gustav* injects a supervisory timing signal `t` to coordinate experimental conditions across devices and computers, supporting configurations ranging from simple (a) to complex (c). During post-processing, temporal offsets among recorded signals are corrected, and uniform start and end timestamps are assigned, thereby ensuring precise synchronization across all signals.

provide complementary views of cognitive and motor processes and are often recorded simultaneously to support richer inference and analysis [260, 261, 262, 263, 264, 265]. However, when these signals are collected through multiple devices and sometimes across multiple computers, temporal synchronization becomes a critical challenge. If signals are not accurately aligned in time, analyses may produce contradictory or misleading interpretations of user behavior and mental state.

Temporal misalignment can result in situations where different signals appear to indicate conflicting cognitive states, even though they originate from the same experimental condition. For example, neural signals may indicate resting states while eye tracking signals simultaneously suggest high cognitive processing, solely due to offset timestamps. As a result, precise temporal synchronization is essential to ensure valid multimodal interpretation and to avoid erroneous conclusions [266, 267, 268].

Existing solutions to synchronization typically rely on dedicated hardware, specialized software, or alignment algorithms that attempt to match signals post hoc. These approaches often suffer from limitations such as vendor lock-in, lack of generality, operating system restrictions, or difficult adoption in real experimental workflows. The motivation of this paper is to provide a general, simple, and hardware-agnostic solution that can be integrated into virtually any experimental setup without altering existing recording pipelines.

This work introduces Gustav, a software-level approach designed to orchestrate the recording and synchronization of sensory signals across devices and computers. Gustav aims to standardize timelines, remove irrelevant data segments outside experimental conditions, and correct timestamp offsets with millisecond precision. By focusing on instrumenting stimulus presentation software rather than the sensing hardware, the approach seeks to offer a practical, scalable, and open solution for multimodal experiments regardless of their complexity.

Related Work

Prior research on multimodal synchronization has followed three main strategies: using dedicated hardware, relying on dedicated software environments, or applying signal alignment algorithms after data collection. Hardware-based approaches include systems that synchronize specific sensor combinations, such as EEG and Electromyography (EMG), but these solutions are often limited in scope and support only a narrow range of signals [7]. More comprehensive toolkits have been proposed, but they are frequently restricted to specific operating systems, such as Windows, which limits their applicability in diverse research environments [269, 270].

Other hardware-based techniques involve sending synchronization signals through cables or external devices, such as light-emitting diodes or transistor-transistor logic triggers, to align data streams [271, 272]. While effective, these methods require additional equipment and careful calibration, increasing setup

complexity and reducing accessibility.

Software-based solutions have also been proposed, but many are tightly coupled to specific platforms or lack support for heterogeneous sensors. Some researchers have developed custom synchronization programs without releasing technical details, making replication and reuse difficult [273].

Algorithmic approaches attempt to synchronize signals by identifying common patterns or landmarks across modalities. Examples include cross-correlation of temporal amplitudes, frequency-domain alignment, blink detection in eye tracking data, and extraction of characteristic events from EEG or driving simulator signals [274, 275, 276, 277, 278]. These methods assume the presence of shared signal features across devices, which is not always guaranteed, and they often require complex signal processing pipelines.

In contrast to these approaches, Gustav is positioned as an offline, software-agnostic solution that can be integrated with any existing recording software. By avoiding reliance on hardware triggers or signal-specific alignment assumptions, it offers a general and practical alternative for synchronizing sensory signals across diverse experimental setups.

Methodology

Gustav's approach to temporal synchronization consists of two independent components (Figure 5.1): an event logger and a file post-processor. The event logger is integrated into the stimulus presentation software and records event-related timestamps using Unix time format, representing the number of milliseconds elapsed since January 1, 1970. This standardized time format is widely supported across operating systems and programming languages, enabling consistent computation of time differences across devices.

The logged events are user-defined, although typical use involves recording the beginning and end of each stimulus or experimental condition. These timestamps are stored in a CSV file on the computer running the stimulus presentation

software. The event logger is implemented in multiple programming languages, including Python, JavaScript for both browser and NodeJS environments, and Java, allowing it to be embedded into a wide range of experimental platforms.

The file post-processor operates offline and takes as input the event timestamp file along with any number of log files produced by the recording devices. These log files contain the sensory data and are expected to be in CSV format. For devices that output data in other formats, Gustav provides a conversion utility to transform tabular data such as Matlab or EDF files into CSV. The post-processor performs synchronization in two main steps. First, it slices each log file according to the event-related timestamps, retaining only the data corresponding to each experimental condition. Second, it corrects timestamp offsets so that all signals share identical start and end times for each condition.

The post-processor is implemented in Python to ensure platform independence. It can automatically detect whether a column contains Unix timestamps. If timestamps are missing, the tool can infer them using file creation times or device-specific timing information, such as elapsed time since calibration or device startup. This design allows Gustav to synchronize signals even when devices do not natively record absolute timestamps.

Results

We demonstrated the effectiveness of Gustav through a multi-device, multi-computer use case experiment based on the Posner cueing task, a well-established paradigm for assessing visual attention [279]. In this experiment, participants respond to targets appearing on either side of the screen following directional cues. The setup involved recording EEG signals, eye movements, and mouse movements using separate devices connected to two different computers.

The stimulus presentation software was implemented in PsychoPy and instrumented with Gustav's event logger by inserting logging hooks before and after stimulus events. EEG signals were recorded on one computer using dedicated

recording software, while eye tracking, mouse tracking, and stimulus presentation were conducted on a second computer. After data collection, the Gustav post-processor was applied to the recorded logs.

The results show that Gustav successfully sliced and synchronized all recorded signals so that they aligned precisely with the duration of each experimental condition. Timestamp offsets were corrected with millisecond precision, and all data streams shared the same temporal boundaries. The synchronized data enabled coherent multimodal analysis without the inconsistencies that would arise from unsynchronized recordings.

Discussion and Limitations

Gustav provides a general solution to a longstanding problem in multimodal data collection. By focusing on stimulus-level event logging rather than sensor-level triggers, the approach simplifies experimental setup and reduces dependency on specific hardware or software ecosystems. The offline post-processing design allows researchers to conduct experiments as usual and apply synchronization afterward, minimizing disruption to existing workflows.

However, the approach has limitations. Gustav requires manual insertion of logging hooks into the stimulus presentation software, which may introduce some overhead for experimenters. Gustav cannot compensate for hardware-induced latencies, although such delays are typically small, often under 10 milliseconds, and below the sampling resolution targeted by the system [278]. In multi-computer setups, Gustav relies on accurate system clocks across machines. This requirement is generally satisfied through network time protocol synchronization, provided that the computers have internet access. The approach also operates offline, meaning it does not provide real-time synchronization. While real-time operation is outside the current scope, the authors suggest that it could be achieved in the future through socket-based data streaming to a centralized server.

Future Work

Future work will focus on further reducing the integration effort required to adopt Gustav in real-world experimental workflows. In particular, developing native plugins for popular experiment design platforms such as PsychoPy, PsychoJS, and OpenSesame would eliminate the need for manual insertion of logging hooks and lower the barrier for non-technical researchers. Another promising direction is extending Gustav toward optional real-time synchronization by streaming event logs and sensor data to a centralized server, enabling online monitoring and adaptive experimental paradigms. Additionally, future versions could incorporate systematic estimation and reporting of device-specific latencies, providing researchers with clearer bounds on temporal accuracy. Finally, broader empirical validation across diverse sensing modalities and large-scale deployments would help further characterize Gustav's robustness, scalability, and applicability in complex multimodal research settings.

Conclusion

We introduced Gustav, a simple yet effective approach for synchronizing sensory signals across devices and computers in multimodal experiments. By injecting event-related timestamps into stimulus presentation software and applying offline post-processing, Gustav ensures that all recorded signals are temporally aligned with millisecond precision. The approach is hardware-agnostic, software-independent, and compatible with a wide range of experimental setups, from single-device systems to complex multi-device, multi-computer configurations.

Gustav is released as open source software, making it accessible to the research community. The approach addresses key limitations of existing synchronization methods and provides a practical foundation for accurate multimodal data collection and analysis in human-centered computing, neuroscience, and related fields.

5.3 Thalamus: A User Simulation Toolkit for Prototyping Multimodal Sensing Studies

User studies involving physiological and behavioral data collection are often costly and time intensive, as they require careful experimental design, device calibration, and extensive software testing. We introduce Thalamus, a software toolkit for collecting and simulating multimodal signals that enables researchers to prepare for unforeseen scenarios prior to recruiting participants and even before installing or purchasing specific sensing devices. Thalamus supports the modification, synchronization, and broadcasting of physiological data streams originating from multiple devices, which may be distributed across different locations. The toolkit is cross-platform, device agnostic, and easy to use, making it a practical and valuable resource for HCI research.

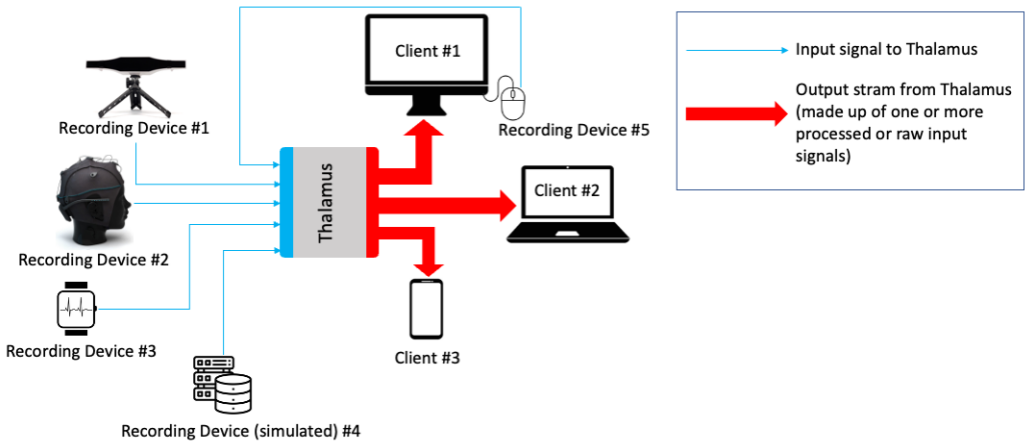


Figure 5.2 Conceptual diagram of the *Thalamus* toolkit, illustrating its ability to simulate multiple types of recording devices and to synchronize heterogeneous data streams, such as neural signals and eye movements. The toolkit can receive feedback from any connected device; in this example, client #1 also operates as a recording device (Recording Device #5), enabling real-time multimodal data acquisition and analysis.

Motivation

Human computer interaction research increasingly relies on multimodal behavioral and physiological signals to capture the complex and dynamic nature of user interaction with technology. Collecting data such as eye movements, brain activity, facial expressions, mouse movements, and other physiological signals simultaneously allows researchers to develop richer and more nuanced interpretations of user behavior that cannot be achieved through a single modality alone [280, 281, 282, 283]. These multimodal measurements support the design and evaluation of interactive systems by revealing how different sensory channels interact during user engagement.

Despite their benefits, multimodal studies are challenging to conduct, particularly when data collection occurs in real time. Experimental setups require careful planning, device calibration, and extensive software testing, and errors during data collection can lead to significant financial and time costs related to lab preparation, participant compensation, and post-processing [284]. These challenges are amplified when dealing with heterogeneous devices that generate data streams with different formats, sampling rates, and reliability characteristics. As a result, researchers face substantial risks when conducting multimodal experiments without first validating their setups.

Simulation provides an effective strategy to mitigate these risks. By simulating data streams before running a real experiment, researchers can test whether their equipment, data pipelines, and analysis methods function as intended. Simulation also enables researchers to identify corner cases such as missing data, device disconnections, noise, and synchronization issues early in the design process. Prior work has demonstrated the value of simulation for prototyping, training, and evaluating interactive systems in a controlled and cost-effective manner [285, 286].

Thalamus addresses the lack of dedicated software for simulating multimodal sensing studies in HCI. Existing solutions focus primarily on data collection

rather than simulation, forcing researchers to rely on ad hoc tools that are time-consuming to build and difficult to replicate [277, 287, 269, 273, 271, 276, 278, 275, 274, 272, 270]. To address this gap, we introduce Thalamus, a toolkit designed to capture, simulate, modify, synchronize, and broadcast multimodal data streams without requiring early access to real users or physical sensing devices. Thalamus aims to support cross-platform and cross-device experimentation and reduce the cost and complexity of preparing multimodal studies.

Related Work

Previous research has highlighted the importance of multimodal sensing for understanding user behavior and mental states in interactive systems [288, 283]. Simultaneous collection of physiological and behavioral signals has been applied in domains such as affective computing, autonomous driving, learning environments, and gaming [280, 282, 273]. However, collecting and synchronizing such data streams remains technically challenging, especially in real-time settings [284].

Many researchers have relied on ad hoc solutions to record multiple signals simultaneously, often combining eye tracking, EEG, ECG, and interaction logs [277, 287, 269, 273, 271, 276, 278, 275, 274, 272, 270]. While these approaches can solve specific experimental problems, they often lack generalizability and reproducibility, making them difficult to reuse across studies [74]. Other work has proposed software systems for multimodal data acquisition, but these systems typically focus on recording real sensor data rather than simulating signals [284, 270].

Simulation has been recognized as a valuable tool for HCI research, enabling researchers to test interfaces, train users, and explore system behavior before full deployment [285, 286]. In physiological computing, publicly available datasets such as DREAMER [289], MAHNOB [66], and SEED [53] have been used to study emotional responses and cognitive states. However, there has been no widely adopted toolkit that integrates simulation, synchronization, and real-time

streaming of multimodal signals in a unified framework.

This work positions Thalamus as a novel contribution that fills this gap by offering a general-purpose toolkit for simulating and coordinating multimodal data streams. Unlike prior solutions, Thalamus is designed to be cross-platform, device-agnostic, and capable of integrating both real and simulated data sources within the same experimental setup.

Methodology

Thalamus is implemented as a modular software toolkit consisting of three main components (Figure 5.2): a central coordination hub called Thalamus Core, a set of recording devices that may be real or simulated, and a set of clients that receive and process data streams. The toolkit is implemented in Python and uses socket-based communication over the TCP/IP protocol to support real-time data transfer across devices and operating systems.

Thalamus Core functions as the central hub that receives, organizes, synchronizes, and broadcasts data streams. Inspired by the biological thalamus, it coordinates information flow between recording devices and clients. Devices send data to the Core in JSON format, which provides cross-platform compatibility and self-describing data structures. Each data sample includes a coordinated universal time timestamp, ensuring consistent temporal alignment across devices.

Recording devices can be physical sensors or simulated sources. Simulated devices inject data from public datasets through controllers that parse and reformat the data to match the expected transport format. When real devices do not provide coordinated universal time timestamps or direct access to data streams, custom controllers can be implemented to extract samples from files or proprietary formats and forward them to Thalamus Core.

Clients connect to Thalamus Core via socket connections and can subscribe to specific signals based on experimental needs. Clients are not restricted by operating system or programming language. In addition to consuming data

streams, clients can also act as recording devices by sending signals back to the Core, enabling bidirectional data sharing such as broadcasting mouse movements between participants.

The toolkit includes built-in features for handling common issues in multimodal data collection. These include simulating missing values by sending placeholder values, applying filters such as Savitzky Golay [290] and Kalman filters [291], synchronizing signals based on Unix timestamps, injecting noise of various types, and simulating delays. Temporal synchronization operates at millisecond precision, which is sufficient for most human computer interaction studies. These features allow researchers to stress-test their experimental pipelines and analysis methods before conducting real experiments.

Results

We demonstrate the functionality of Thalamus through several illustrative use cases rather than formal quantitative evaluation. Visual examples show how the toolkit handles missing values in eye-tracking data, applies filters to mouse cursor trajectories, synchronizes EEG and pupil diameter signals, injects Gaussian noise, and simulates delayed signals. These demonstrations illustrate the toolkit's ability to manipulate and align multimodal data streams effectively.

Several practical scenarios are presented to showcase the benefits of simulation. In one example, researchers simulate delays and noise to test real-time applications subject to unstable network conditions. In another scenario, researchers use Thalamus to evaluate different EEG devices by simulating data from public datasets such as DREAMER [289], MAHNOB [66], and SEED [53] before purchasing hardware. Additional use cases include device stress-testing, remote data collection for online lectures, and broadcasting identical physiological signals to multiple clients in telemedicine studies.

Across these examples, Thalamus enables researchers to validate experimental designs, optimize data processing pipelines, and anticipate potential failures. By

performing these evaluations before involving real participants, the toolkit helps reduce costs and improve experimental reliability.

Discussion and Limitations

Thalamus enables researchers to perform dry runs of multimodal experiments, reducing the risk of technical failures and wasted resources. By integrating simulation, synchronization, and signal manipulation in a single toolkit, Thalamus provides a streamlined approach to preparing complex studies. Its cross-platform and open-source design further supports adoption and extensibility.

However, the toolkit has limitations. Thalamus currently supports simulation from structured data sources such as CSV and JSON files and does not natively support relational databases, though this is planned for future work. Writing custom controllers for proprietary devices may require programming expertise, which could be a barrier for some researchers. Additionally, while Thalamus provides millisecond-level synchronization, it relies on accurate system clocks across devices.

Future Work

Future work can extend Thalamus along several complementary directions. First, the toolkit could be expanded to support additional data backends and formats such as relational databases and streaming platforms, enabling seamless integration with large-scale or longitudinal datasets. Second, higher-level simulation capabilities could be introduced, including generative models that synthesize realistic multimodal signals based on configurable user states or interaction scenarios, allowing researchers to explore hypothetical conditions beyond existing datasets. Third, improving usability remains an important goal, and a graphical user interface for configuring devices, pipelines, and simulations would lower the barrier to adoption for researchers with limited programming experience.

Finally, future research could systematically evaluate Thalamus in real-world studies to quantify its impact on experimental reliability, setup time, and cost, and to establish best practices for simulation-driven preparation of multimodal HCI experiments.

Conclusion

In this work, we introduce Thalamus, a user simulation toolkit designed to support prototyping and preparation of multimodal sensing studies in HCI. By enabling researchers to simulate, synchronize, modify, and broadcast physiological and behavioral signals across devices, Thalamus addresses a critical gap in experimental preparation tools. The toolkit reduces costs, improves reliability, and supports flexible experimentation without requiring early access to participants or hardware.

Thalamus is released as open-source software, making it accessible to the research community. Its design supports cross-device and cross-platform operation, providing a practical and extensible solution for multimodal experimentation. The toolkit represents an important step toward more robust, efficient, and reproducible multimodal studies in human-centered computing.

Chapter 6

Conclusion, Discussion and Outlook

This thesis examined how neurophysiological and behavioral signals can be leveraged to decode three fundamental cognitive states, namely affect, attention, and expertise, during human–computer interaction. Across a series of empirical publications, the thesis demonstrated that these signals provide continuous, objective, and scalable alternatives to traditional explicit measures such as questionnaires and self-reports. In parallel, the work addressed methodological challenges inherent to multimodal experimentation by introducing dedicated toolkits that support synchronization, simulation, and reproducibility.

Taken together, the contributions of this thesis advance the understanding of how cognitive states manifest across modalities including EEG, eye movements, mouse behavior, facial expressions, and head movements. Beyond individual results, the work establishes a coherent methodological perspective showing that cognitive state decoding in HCI can be both efficient and practically deployable when guided by principled design choices and supported by appropriate tooling.

6.1 Summary of Findings

This thesis investigated how neurophysiological and behavioral signals can be used to decode human cognitive states in HCI. The novelty of this work lies not only in improving performance, but in systematically demonstrating how cognitive states can be decoded efficiently and at scale under realistic constraints using multimodal signals. In particular, the thesis emphasizes efficiency, scalability, and ecological validity, which are often overlooked in prior research.

Decoding Affect (Chapter 2). The first set of contributions focused on affective state decoding, with an emphasis on efficiency and ecological validity. The results showed that reliable classification of arousal and valence does not require long EEG recordings or exhaustive feature sets. Instead, affective information emerged early and could be captured using relatively short signal segments. Across two benchmark datasets, EEG-based affect decoding remained accurate with approximately 30 to 42 seconds of data, and short temporal windows were sufficient for robust feature extraction. Importantly, the Beta frequency band alone consistently provided discriminative power comparable to models trained on all frequency bands, indicating that affective neural signatures are more concentrated than commonly assumed.

The second affective study extended this perspective to interface evaluation. By combining EEG, eye tracking, facial expressions, and pupillometry, the work demonstrated that users' affective responses to good versus bad GUI designs can be differentiated implicitly. Early pupil dynamics, fixation patterns, and neural activity revealed systematic differences between interface quality levels, even without relying on explicit judgments. Together, these findings establish that affect can be decoded efficiently and unobtrusively in realistic HCI settings.

Decoding Attention (Chapter 3). The second core contribution addressed attention allocation in information-rich environments, focusing on web search.

The AdSERP dataset provided a large-scale multimodal resource combining synchronized eye movements, mouse trajectories, and page structure information. Analyses of this dataset confirmed a strong relationship between gaze and cursor behavior, while also revealing differences in how users attend to different types of advertisements.

Building on this dataset, the AdSight framework demonstrated that Transformer-based sequence-to-sequence models can infer fine-grained gaze-based attention metrics using mouse data alone. The results showed that direct-display advertisements attract substantially more visual attention than previously assumed, and that slot-centered representations enable robust generalization across heterogeneous search engine results page layouts. These findings indicate that scalable behavioral sensing can approximate attention with high accuracy, significantly lowering the cost and complexity associated with eye-tracking-based approaches.

Decoding Expertise (Chapter 4). The third empirical contribution examined expertise assessment in a medical training context. By analyzing eye and head movements during simulated obstetric procedures, the study demonstrated that trained and untrained practitioners exhibit systematically different behavioral signatures. In particular, head movement dynamics, including angular velocity and cumulative rotation, emerged as highly discriminative features. Oculomotor measures further reflected differences in cognitive load and task structure across training phases.

These results show that expertise manifests not only in task outcomes but also in stable, measurable patterns of perception and motor behavior. The findings support the use of multimodal sensing as an objective and continuous means of assessing skill acquisition, complementing traditional instructor-based evaluations.

Applications and Toolkits (Chapter 5). Finally, the Gustav and Thalamus toolkits addressed a recurring challenge across all empirical studies, namely the

technical difficulty of collecting, synchronizing, and validating multimodal data. Gustav enables precise cross-device synchronization, while Thalamus supports simulation and prototyping of multimodal sensing pipelines. Together, these tools improve reproducibility, reduce setup costs, and facilitate experimentation under controlled and simulated conditions, directly supporting the empirical contributions of this thesis.

6.2 Discussion

Across the empirical studies presented in this thesis, several overarching themes emerge that cut across cognitive states, sensing modalities, and application domains. These themes highlight not only what can be decoded from neurophysiological and behavioral signals, but also how such decoding should be approached in realistic HCI settings.

Multimodality as a necessary but flexible design principle

A central insight of this thesis is that cognitive states are inherently multimodal phenomena. Affect, attention, and expertise are not expressed through a single channel, but are distributed across EEG activity, eye movements, head dynamics, facial expressions, and interaction behavior such as mouse trajectories. Each modality captures a different facet of internal cognitive processing, and their combination enables richer and more robust inference than any single signal alone.

At the same time, the results show that multimodality does not necessarily imply maximal sensor complexity. In several cases, partial modality substitution proved effective. For example, mouse movements approximated gaze-based attention with high accuracy on search engine results pages, reducing reliance on eye tracking hardware. Similarly, affective EEG decoding achieved competitive performance using a single frequency band rather than the full spectrum.

These findings suggest that multimodality should be treated as a design space in which modalities can be selectively combined or replaced depending on practical constraints, rather than as a fixed requirement to collect all possible signals.

Efficiency as a first-class objective in cognitive decoding

Another recurring theme is the importance of efficiency in both data collection and modeling. Across affective, attentional, and expertise-related tasks, the studies consistently demonstrated that reliable decoding can be achieved under constrained conditions. Short EEG segments, limited temporal windows, and reduced feature sets were sufficient for robust inference, and in many cases performed comparably to more exhaustive configurations.

This challenges a common assumption in physiological computing and machine learning research, namely that performance necessarily scales with longer recordings, larger feature spaces, or more complex models. Instead, the results indicate that cognitive signals often contain early and concentrated informative patterns, particularly in dynamic interaction scenarios. From an HCI perspective, this emphasis on efficiency is critical, as systems requiring long calibration phases or heavy sensing infrastructure are unlikely to be adopted outside laboratory settings.

Implicit sensing versus explicit self-report

A further important discussion point concerns the relationship between implicit sensing and traditional explicit measures. Throughout the thesis, physiological and behavioral signals revealed distinctions in affect, attention, and expertise that either preceded or complemented subjective reports. This suggests that implicit signals provide access to continuous and often pre-conscious processes that are not fully captured by post-hoc self-assessment.

Rather than replacing self-reports, the findings support a hybrid evaluation

paradigm in which implicit sensing augments traditional methods. Such an approach is particularly valuable in adaptive systems that require continuous assessment while minimizing user burden and interaction disruption.

Domain-specific manifestations of cognitive states

While the thesis adopts a unifying framework around affect, attention, and expertise, the results also demonstrate that cognitive states manifest differently depending on task and context. Affective decoding relies heavily on neural dynamics and autonomic responses, attentional allocation emerges through coordinated visual and interaction behavior, and expertise is reflected in stable perceptual-motor strategies.

These observations highlight the importance of domain-sensitive modeling. Head movement dynamics proved especially informative in the medical training scenario, whereas cursor behavior played a central role in web search. This cautions against overly generic decoding models and motivates approaches grounded in task structure and environmental context.

Methodological robustness and reproducibility

Finally, the thesis highlights the importance of methodological infrastructure in multimodal cognitive research. Synchronization errors, inconsistent pipelines, and limited opportunities for pre-study validation can undermine experimental validity. By introducing Gustav and Thalamus, this work treats synchronization, simulation, and reproducibility as first-class research concerns rather than implementation details, supporting more rigorous and transferable HCI research.

6.3 Outlook

The findings of this thesis point toward several promising directions for future research and application.

Short-window models and efficient feature sets enable real-time cognitive-aware systems that can adapt dynamically to users' affective states, attentional focus, or expertise level. Such systems could adjust content presentation, feedback, or interface complexity without requiring intrusive sensing or long calibration phases.

As eye trackers, inertial sensors, and interaction logging become increasingly common in consumer devices, the methods developed in this thesis can be integrated into real-world applications. Mouse-based attention modeling, in particular, offers a practical pathway for large-scale deployment in web search and online advertising.

Future work should also extend these approaches beyond controlled laboratory settings to more naturalistic environments. The combination of Gustav for synchronized data collection and Thalamus for simulation-based preparation provides a foundation for conducting robust in-the-wild studies that capture greater behavioral variability.

Finally, as cognitive state decoding becomes more powerful and accessible, ethical considerations will become increasingly important. Transparent data handling, privacy-preserving modeling, and responsible interpretation of inferred cognitive states will be essential to ensure user trust and societal acceptance.

Overall, this thesis demonstrates that multimodal implicit signals provide a feasible, efficient, and reliable foundation for decoding human cognition, supporting the development of adaptive and user-centered interactive systems.

References

[1] K. Latifzadeh, N. Gozalpour, V. J. Traver, T. Ruotsalo, A. Kawala-Sterniuk, and L. A. Leiva. "Efficient decoding of affective states from video-elicited EEG signals: an empirical investigation". In: *ACM Transactions on Multimedia Computing, Communications and Applications* 20.10 (2024), pp. 1–24.

[2] S. Haddad, K. Latifzadeh, S. Duraisamy, J. Vanderdonckt, O. Daassi, S. Belghith, and L. A. Leiva. "Good GUIs, Bad GUIs: Affective Evaluation of Graphical User Interfaces". In: *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*. 2024, pp. 232–243.

[3] P. J. Lang. "The emotion probe: Studies of motivation and attention." In: *American psychologist* 50.5 (1995), p. 372.

[4] S. M. Alarcao and M. J. Fonseca. "Emotions recognition using EEG signals: A survey". In: *IEEE Trans. Affect. Comput.* 10.3 (2017), pp. 374–393.

[5] S. Brave and C. Nass. "Emotion in human-computer interaction". In: *The human-computer interaction handbook*. 2007, pp. 103–118.

[6] X. Li, Y. Zhang, P. Tiwari, D. Song, B. Hu, M. Yang, Z. Zhao, N. Kumar, and P. Marttinen. "EEG based emotion recognition: A tutorial and review". In: *ACM Computing Surveys* 55.4 (2022), pp. 1–57.

[7] M. Kostyunina and M. Kulikov. "Frequency characteristics of EEG spectra in the emotions". In: *Neurosci. Behav. Physiol.* 26.4 (1996), pp. 340–343.

[8] W. Hu, G. Huang, L. Li, L. Zhang, Z. Zhang, and Z. Liang. “Video-triggered EEG-emotion public databases and current methods: a survey”. In: *Brain Sci. Adv.* 6.3 (2020), pp. 255–287.

[9] M. Egger, M. Ley, and S. Hanke. “Emotion recognition from physiological signal analysis: A review”. In: *Electronic Notes in Theoretical Computer Science* 343 (2019), pp. 35–55.

[10] M. Imani and G. A. Montazer. “A survey of emotion recognition methods with emphasis on E-Learning environments”. In: *Journal of network and computer applications* 147 (2019), p. 102423.

[11] R. W. Picard, E. Vyzas, and J. Healey. “Toward machine emotional intelligence: Analysis of affective physiological state”. In: *IEEE transactions on pattern analysis and machine intelligence* 23.10 (2001), pp. 1175–1191.

[12] Y. Cimtay and E. Ekmekcioglu. “Investigating the use of pretrained convolutional neural network on cross-subject and cross-dataset EEG emotion recognition”. In: *Sensors* 20.7 (2020), p. 2034.

[13] A. Kawala-Sterniuk, N. Browarska, A. Al-Bakri, M. Pelc, J. Zygarlicki, M. Sidikova, R. Martinek, and E. J. Gorzelanczyk. “Summary of over fifty years with brain-computer interfaces—a review”. In: *Brain sciences* 11.1 (2021), p. 43.

[14] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. “Deap: A database for emotion analysis; using physiological signals”. In: *IEEE transactions on affective computing* 3.1 (2011), pp. 18–31.

[15] M. L. R. Menezes, A. Samara, L. Galway, A. Sant’Anna, A. Verikas, F. Alonso-Fernandez, H. Wang, and R. Bond. “Towards emotion recognition for virtual environments: an evaluation of EEG features on benchmark dataset”. In: *Pers. Ubiquit. Comput.* 21 (2017), pp. 1003–1013.

[16] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu. “Identifying Stable Patterns over Time for Emotion Recognition from EEG”. In: *IEEE Trans. Affect. Comput.* 10.3 (2019), pp. 417–429.

[17] J. Wagner, J. Kim, and E. André. “From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification”. In: *Proc. ICME*. 2005, pp. 940–943.

[18] A. Appriou, A. Cichocki, and F. Lotte. “Modern machine-learning algorithms: for classifying cognitive and affective states from electroencephalography signals”. In: *IIEEE Trans. Syst. Man Cybern. Syst.* 6.3 (2020), pp. 29–38.

[19] T. Song, W. Zheng, P. Song, and Z. Cui. “EEG Emotion Recognition Using Dynamical Graph Convolutional Neural Networks”. In: *IEEE Trans. Affect. Comput.* 11.3 (2020), pp. 532–541.

[20] P. Zhong, D. Wang, and C. Miao. “EEG-based Emotion Recognition Using Regularized Graph Neural Networks”. In: *IEEE Trans. Affect. Comput.* 13.3 (2022), pp. 1290–1301.

[21] V. J. Traver, J. Zorío, and L. A. Leiva. “Glimpse: A Gaze-Based Measure of Temporal Salience”. In: *Sensors* 21.9 (2021).

[22] E. A. Butler. “Emotions are temporal interpersonal systems”. In: *Curr. Opin. Psychol.* 17 (2017), pp. 129–134.

[23] L. S. Chen and T. S. Huang. “Emotional expressions in audiovisual human computer interaction”. In: *Proc. ICME*. 2000, pp. 423–426.

[24] R. W. Picard. *Affective computing*. MIT press, 2000.

[25] C. L. Lisetti and F. Nasoz. “MAUI: a multimodal affective user interface”. In: *Proc. ACM MM*. 2002, pp. 161–170.

[26] R. W. Levenson. “Blood, sweat, and fears: The autonomic architecture of emotion”. In: *Ann. N. Y. Acad. Sci.* 1000.1 (2003), pp. 348–366.

[27] P. C. Ellsworth and K. R. Scherer. *Appraisal processes in emotion.* Oxford University Press, 2003.

[28] E. Duffy. "Emotion: an example of the need for reorientation in psychology." In: *Psychol. Rev.* 41.2 (1934), p. 184.

[29] I. Lopatovska and I. Arapakis. "Theories, methods and current research on emotions in library and information science, information retrieval and human–computer interaction". In: *Inf. Process. Manag.* 47.4 (2011), pp. 575–592.

[30] P. E. G. Bestelmeyer, S. A. Kotz, and P. Belin. "Effects of emotional valence and arousal on the voice perception network". In: *Soc. Cogn. Affect. Neurosci.* 12.8 (2017), pp. 1351–1358.

[31] P. J. Lang. "The emotion probe. Studies of motivation and attention". In: *Am. Psychol.* 50 (1995), pp. 372–385.

[32] R. J. Larsen and E. Diener. "Promises and problems with the circumplex model of emotion". In: *Review of personality and social psychology*. Ed. by M. Clark. Vol. 13. 1992, pp. 25–59.

[33] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang. "A review of emotion recognition using physiological signals". In: *Sensors* 18.7 (2018), p. 2074.

[34] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. "A survey of affect recognition methods: audio, visual and spontaneous expressions". In: *Proc. ICMI*. 2007, pp. 126–133.

[35] A. Al-Nafjan, M. Hosny, Y. Al-Ohali, and A. Al-Wabil. "Review and classification of emotion recognition based on EEG brain-computer interface system research: a systematic review". In: *Appl. Sci.* 7.12 (2017), p. 1239.

[36] E. P. Torres, E. A. Torres, M. Hernández-Álvarez, and S. G. Yoo. "EEG-based BCI Emotion Recognition: A Survey". In: *Sensors* 20.18 (2020), p. 5083.

[37] K. P. Wagh and K. Vasanth. “Electroencephalograph (EEG) based emotion recognition system: A review”. In: *Innov. Electron. Commun. Eng.* 33 (2019), pp. 37–59.

[38] F. Feradov, I. Mporas, and T. Ganchev. “Evaluation of features in detection of dislike responses to audio–visual stimuli from EEG signals”. In: *Computers* 9.2 (2020), p. 33.

[39] I. Mazumder. “An analytical approach of EEG analysis for emotion recognition”. In: *Proc. DevIC*. 2019, pp. 256–260.

[40] M. S. Özerdem and H. Polat. “Emotion recognition based on EEG features in movie clips with channel selection”. In: *Brain Inform.* 4.4 (2017), pp. 241–252.

[41] W.-L. Zheng, B.-N. Dong, and B.-L. Lu. “Multimodal emotion recognition using EEG and eye tracking data”. In: *Proc. EMBC*. 2014, pp. 5040–5043.

[42] T. B. Alakus, M. Gonen, and I. Turkoglu. “Database for an emotion recognition system based on EEG signals and various computer games – GAMEEMO”. In: *Biomed. Signal Process. Control* 60 (2020), p. 101951.

[43] G. Du, W. Zhou, C. Li, D. Li, and P. X. Liu. “An Emotion Recognition Method for Game Evaluation Based on Electroencephalogram”. In: *IEEE Trans. Affect. Comput.* 14.1 (2023), pp. 591–602.

[44] A. Kumar and A. Kumar. “DEEPHER: Human Emotion Recognition Using an EEG-Based DEEP Learning Network Model”. In: *Eng. Proc.* 10.1 (2021).

[45] E. P. Torres, E. A. Torres, M. Hernández-Álvarez, and S. G. Yoo. “Real-Time Emotion Recognition for EEG Signals Recollected from Online Poker Game Participants”. In: *Proc. Advances in Artificial Intelligence, Software and Systems Engineering*. Ed. by T. Z. Ahram, W. Karwowski, and J. Kalra. 2021, pp. 236–241.

[46] Y. Zhou, T. Xu, S. Li, and R. Shi. “Beyond Engagement: An EEG-Based Methodology for Assessing User’s Confusion in an Educational Game”. In: *Univers. Access Inf. Soc.* 18.3 (2019), pp. 551–563.

[47] L. Bai, J. Guo, T. Xu, and M. Yang. “Emotional Monitoring of Learners Based on EEG Signal Recognition”. In: *Procedia Comput. Sci.* 174 (2020), pp. 364–368.

[48] T. Xu, Y. Zhou, Z. Wang, and Y. Peng. “Learning Emotions EEG-based Recognition and Brain Activity: A Survey Study on BCI for Intelligent Tutoring System”. In: *Procedia Comput. Sci.* 130 (2018), pp. 376–382.

[49] S. Thejaswini, K. M. R. Kumar, and A. N. J. L. “Analysis of EEG based emotion detection of DEAP and SEED-IV databases using SVM”. In: *Proc. ICETSE*. 2019.

[50] N. Kumar, K. Khaund, and S. M. Hazarika. “Bispectral analysis of EEG for emotion recognition”. In: *Procedia Comput. Sci.* 84 (2016), pp. 31–35.

[51] F. Galvão, S. M. Alarcão, and M. J. Fonseca. “Predicting exact valence and arousal values from EEG”. In: *Sensors* 21.10 (2021), p. 3414.

[52] E. S. Pane, A. D. Wibawa, and M. H. Pumomo. “Channel selection of EEG emotion recognition using stepwise discriminant analysis”. In: *Proc. CENIM*. 2018, pp. 14–19.

[53] W.-L. Zheng and B.-L. Lu. “Investigating Critical Frequency Bands and Channels for EEG-based Emotion Recognition with Deep Neural Networks”. In: *IEEE Trans. Auton. Ment. Dev.* 7.3 (2015), pp. 162–175.

[54] Y. Liu and O. Sourina. “EEG-based subject-dependent emotion recognition algorithm using fractal dimension”. In: *Proc. SMC*. 2014, pp. 3166–3171.

[55] R. M. Mehmood, M. Bilal, S. Vimal, and S.-W. Lee. “EEG-based affective state recognition from human brain signals by using Hjorth-activity”. In: *Measurement* 202 (2022), p. 111738.

[56] J. Chen, P. Zhang, Z. Mao, Y. Huang, D. Jiang, and Y. Zhang. "Accurate EEG-based emotion recognition on combined features using deep convolutional neural networks". In: *IEEE Access* 7 (2019), pp. 44317–44328.

[57] Y. Zhang, J. Chen, J. H. Tan, Y. Chen, Y. Chen, D. Li, L. Yang, J. Su, X. Huang, and W. Che. "An investigation of deep learning models for EEG-based emotion recognition". In: *Front. Neurosci.* 14 (2020), p. 622759.

[58] E. Lashgari, D. Liang, and U. Maoz. "Data augmentation for deep-learning-based electroencephalography". In: *J. Neurosci. Methods* 346 (2020), p. 108885.

[59] S. Chang and H. Jun. "Hybrid deep-learning model to recognise emotional responses of users towards architectural design alternatives". In: *J. Asian Archit. Build. Eng.* 18.5 (2019), pp. 381–391.

[60] M. P. Kalashami, M. M. Pedram, and H. Sadr. "EEG Feature Extraction and Data Augmentation in Emotion Recognition". In: *Comput. Intell. Neurosci.* 2022 (2022).

[61] Y. Luo and B.-L. Lu. "EEG data augmentation for emotion recognition using a conditional Wasserstein GAN". In: *Proc. EMBC*. 2018, pp. 2535–2538.

[62] Y. Luo, L.-Z. Zhu, and B.-L. Lu. "A GAN-based data augmentation method for multimodal emotion recognition". In: *Proc. ISNN*. 2019, pp. 141–150.

[63] E. S. Salama, R. A. El-Khoribi, M. E. Shoman, and M. A. W. Shalaby. "EEG-based emotion recognition using 3D convolutional neural networks". In: *Int. J. Adv. Comput. Sci. Appl.* 9.8 (2018).

[64] F. Wang, S.-h. Zhong, J. Peng, J. Jiang, and Y. Liu. "Data augmentation for EEG-based emotion recognition with deep convolutional neural networks". In: *Proc. MMM*. 2018, pp. 82–93.

[65] Z. Zhang, S.-h. Zhong, and Y. Liu. “GANSER: A Self-supervised Data Augmentation Framework for EEG-based Emotion Recognition”. In: *IEEE Trans. Affect. Comput.* 14.3 (2023), pp. 2048–2063.

[66] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. “A multimodal database for affect recognition and implicit tagging”. In: *IEEE Trans. Affect. Comput.* 3.1 (2011), pp. 42–55.

[67] S. S. Gilakjani and H. Al Osman. “A Graph Neural Network for EEG-Based Emotion Recognition with Contrastive Learning and Generative Adversarial Neural Network Data Augmentation”. In: *IEEE Access* 12 (2024), pp. 113–130.

[68] N. Kamel and A. S. Malik. *EEG/ERP Analysis: Methods and Applications*. CRC Press, Taylor & Francis, 2015.

[69] S. Saha, K. A. Mamun, K. Ahmed, R. Mostafa, G. R. Naik, S. Darvishi, A. H. Khandoker, and M. Baumert. “Progress in brain computer interface: challenges and opportunities”. In: *Front. Syst. Neurosci.* 15.578875 (2021).

[70] D.-H. Ko, D.-H. Shin, and T.-E. Kam. “Attention-based spatio-temporal-spectral feature learning for subject-specific EEG classification”. In: *Proc. BCI*. 2021.

[71] N. Kos’myna and F. Tarpin-Bernard. “Evaluation and Comparison of a Multimodal Combination of BCI Paradigms and Eye Tracking With Affordable Consumer-Grade Hardware in a Gaming Context”. In: *IEEE Trans. Comput. Intell. AI Games* 5.2 (2013), pp. 150–154.

[72] J.-M. López-Gil, J. Virgili-Gomá, R. Gil, T. Guilera, I. Batalla, J. Soler-González, and R. García. “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”. In: *Front. Comput. Neurosci.* 10 (2016), p. 85.

[73] P. Wang et al. “Application of Combined Brain Computer Interface and Eye Tracking”. In: *Proc. BCI.* 2021.

[74] K. Latifzadeh and L. A. Leiva. “Gustav: Cross-device cross-computer synchronization of sensory signals”. In: *Adjunct Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology.* 2022, pp. 1–3.

[75] K. D. Tzimourta, N. Giannakeas, A. T. Tzallas, L. G. Astrakas, T. Afrantou, P. Ioannidis, N. Grigoriadis, P. Angelidis, D. G. Tsalikakis, and M. G. Tsipouras. “EEG Window Length Evaluation for the Detection of Alzheimer’s Disease over Different Brain Regions”. In: *Brain Sci.* 9.4 (2019).

[76] D. Blanco-Mora., A. Aldridge., C. Jorge., A. Vourvopoulos., P. Figueiredo., and S. Bermúdez i Badia. “Finding the Optimal Time Window for Increased Classification Accuracy during Motor Imagery”. In: *Proc. BIODEVICES.* 2021, pp. 144–151.

[77] N. Singh Malan and S. Sharma. “Time window and frequency band optimization using regularized neighbourhood component analysis for Multi-View Motor Imagery EEG classification”. In: *Biomed. Signal Process. Control* 67 (2021), p. 102550.

[78] Z. Mohammadi, J. Frounchi, and M. Amiri. “Wavelet-based emotion recognition system using EEG signal”. In: *Neural Comput. Appl.* 28.8 (2017), pp. 1985–1990.

[79] J. Zhang, M. Chen, S. Zhao, S. Hu, Z. Shi, and Y. Cao. “ReliefF-based EEG sensor selection methods for emotion recognition”. In: *Sensors* 16.10 (2016), p. 1558.

[80] M. Li, H. Xu, X. Liu, and S. Lu. “Emotion recognition from multichannel EEG signals using K-nearest neighbor classification”. In: *Technol. Health Care* 26.S1 (2018), pp. 509–519.

[81] D. Ouyang, Y. Yuan, G. Li, and Z. Guo. “The effect of time window length on EEG-based emotion recognition”. In: *Sensors* 22.13 (2022), p. 4939.

[82] R. N. Henson. “Neuroimaging studies of priming”. In: *Prog. Neurobiol.* 70.1 (2003), pp. 53–81.

[83] T. Mullen, C. Kothe, Y.-M. Chi, A. Ojeda, T. Kerth, S. Makeig, G. Cauwenberghs, and T.-P. Jung. “Real-time modeling and 3D visualization of source dynamics and connectivity using wearable EEG”. In: *Proc. EMBC*. 2013, pp. 2184–2187.

[84] O. Bertrand, F. Perrin, and J. Pernier. “A theoretical justification of the average reference in topographic evoked potential studies”. In: *Electroencephalogr. Clin. Neurophysiol.* 62.6 (1985), pp. 462–464.

[85] S. Bagherzadeh, K. Maghooli, A. Shalbaf, and A. Maghsoudi. “Emotion recognition using effective connectivity and pre-trained convolutional neural networks in EEG signals”. In: *Cogn. Neurodynamics* 16.5 (2022), pp. 1087–1106.

[86] S.-H. Kim, H.-J. Yang, N. A. T. Nguyen, S. K. Prabhakar, and S.-W. Lee. “WeDea: A new EEG-based framework for emotion recognition”. In: *IEEE J. Biomed. Health Inform.* 26.1 (2021), pp. 264–275.

[87] B. Hjorth. “EEG analysis based on time domain properties”. In: *Electroencephalogr. Clin. Neurophysiol.* 29.3 (1970), pp. 306–310.

[88] D. Devi, S. Sophia, and S. Boselin Prabhu. “Chapter 4 - Deep learning-based cognitive state prediction analysis using brain wave signal”. In: *Cognitive Computing for Human-Robot Interaction*. Ed. by M. Mittal, R. R. Shah, and S. Roy. 2021, pp. 69–84.

[89] O. Fasil and R. Rajesh. “Time-domain exponential energy for epileptic EEG signal classification”. In: *Neurosci. Lett.* 694 (2019), pp. 1–8.

[90] A. M. Bhatti, M. Majid, S. M. Anwar, and B. Khan. “Human emotion recognition and analysis in response to audio music using brain signals”. In: *Comput. Hum. Behav.* 65 (2016), pp. 267–275.

[91] L. Shaw and A. Routray. “Statistical features extraction for multivariate pattern analysis in meditation EEG using PCA”. In: *Proc. ISC*. 2016, pp. 1–4.

[92] X. Li, D. Song, P. Zhang, Y. Zhang, Y. Hou, and B. Hu. “Exploring EEG features in cross-subject emotion recognition”. In: *Front. Neurosci.* 12 (2018), p. 162.

[93] K. P. Wagh and K. Vasanth. “Performance evaluation of multi-channel electroencephalogram signal (EEG) based time frequency analysis for human emotion recognition”. In: *Biomed. Signal Process. Control* 78 (2022), p. 103966.

[94] M. A. Rahman, M. F. Hossain, M. Hossain, and R. Ahmmed. “Employing PCA and t-statistical approach for feature extraction and classification of emotion from multichannel EEG signal”. In: *Egypt. Inform. J.* 21.1 (2020), pp. 23–35.

[95] E. Brochu, V. M. Cora, and N. De Freitas. “A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning”. In: *arXiv preprint arXiv:1012.2599* (2010).

[96] V. Navalpakkam, L. Jentzsch, R. Sayres, S. Ravi, A. Ahmed, and A. Smola. “Measurement and Modeling of Eye-Mouse Behavior in the Presence of Nonlinear Page Layouts”. In: *Proceedings of the 22nd International Conference on World Wide Web*. WWW ’13. Rio de Janeiro, Brazil: Association for Computing Machinery, 2013, pp. 953–964. DOI: 10.1145/2488388.2488471.

[97] E. Ostertagova and O. Ostertag. “Methodology and Application of Savitzky-Golay Moving Average Polynomial Smoother”. In: *Global Journal of Pure and Applied Mathematics* 12.4 (Aug. 2016), pp. 3201–3210.

[98] S. Medjden, N. Ahmed, and M. Lataifeh. “Adaptive user interface design and analysis using emotion recognition through facial expressions and body posture from an RGB-D sensor”. In: *PLOS ONE* 15.7 (July 2020), pp. 1–37. DOI: 10.1371/journal.pone.0235908.

[99] J. Vanderdonckt and A. Beirekdar. “Automated Web Evaluation by Guideline Review”. In: *J. Web Eng.* 4.2 (2005), pp. 102–117.

[100] A. Bojko. “Using Eye Tracking to Compare Web Page Designs: A Case Study”. In: *Journal of Usability Studies* 1.3 (May 2006), pp. 112–120.

[101] R. A. Fernandez, J. A. Deja, and B. P. V. Samson. “Automating Heuristic Evaluation of Websites Using Convolutional Neural Networks”. In: *Proceedings of the Asian HCI Symposium’18 on Emerging Research Collection*. Asian HCI Symposium’18. Montreal, QC, Canada: Association for Computing Machinery, 2018, pp. 9–12. DOI: 10.1145/3205851.3205854.

[102] Z. Liang, S. Oba, and S. Ishii. “An Unsupervised EEG Decoding System for Human Emotion Recognition”. In: *Neural Netw.* 116.C (Aug. 2019), pp. 257–268. DOI: 10.1016/j.neunet.2019.04.003.

[103] K. Z. Gajos and K. Chauncey. “The Influence of Personality Traits and Cognitive Load on the Use of Adaptive User Interfaces”. In: *Proceedings of the 22nd International Conference on Intelligent User Interfaces, IUI 2017, Limassol, Cyprus, March 13-16, 2017*. Ed. by G. A. Papadopoulos, T. Kuflik, F. Chen, C. Duarte, and W. Fu. ACM, 2017, pp. 301–306. DOI: 10.1145/3025171.3025192.

[104] M. Zen and J. Vanderdonckt. “Towards an evaluation of graphical user interfaces aesthetics based on metrics”. In: *IEEE 8th International Conference on Research Challenges in Information Science, RCIS 2014, Marrakech,*

Morocco, May 28-30, 2014. Ed. by M. Bajec, M. Collard, and R. Deneckère. IEEE, 2014, pp. 1–12. DOI: 10.1109/RCIS.2014.6861050.

[105] T. Holmes and J. M. Zanker. "Using an Oculomotor Signature as an Indicator of Aesthetic Preference". In: *i-Perception* 3.7 (2012). PMID: 23145294, pp. 426–439. DOI: 10.1068/i0448aap.

[106] N. Burny and J. Vanderdonckt. "(Semi-)Automatic Computation of User Interface Consistency". In: *EICS '22: ACM SIGCHI Symposium on Engineering Interactive Computing Systems, Sophia Antipolis, France, June 21 - 24, 2022, Companion Volume*. Ed. by M. Winckler and A. Quigley. ACM, 2022, pp. 5–13. DOI: 10.1145/3531706.3536448.

[107] L. A. Leiva and R. Vivó. "Web Browsing Behavior Analysis and Interactive Hypervideo". In: *ACM Transactions on the Web* 7.4 (2013).

[108] A. Whitefield, F. Wilson, and J. Dowell. "A framework for human factors evaluation". In: *Behaviour & Information Technology* 10.1 (1991), pp. 65–79. DOI: 10.1080/01449299108924272.

[109] F. Z. Canal, T. R. Müller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon, and A. C. Sobieranski. "A survey on facial emotion recognition techniques: A state-of-the-art literature review". In: *Information Sciences* 582 (2022), pp. 593–617. DOI: https://doi.org/10.1016/j.ins.2021.10.005.

[110] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. "DEAP: A Database for Emotion Analysis ;Using Physiological Signals". In: *IEEE Transactions on Affective Computing* 3.1 (2012), pp. 18–31. DOI: 10.1109/T-AFFC.2011.15.

[111] A. Savitzky and M. J. E. Golay. "Smoothing and Differentiation of Data by Simplified Least Squares Procedures." In: *Analytical Chemistry* 36.8 (1964), pp. 1627–1639. DOI: 10.1021/ac60214a047.

[112] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic. "AFEW-VA database for valence and arousal estimation in-the-wild". In: *Image and Vision Computing* 65 (2017). Multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing, pp. 23–36. DOI: https://doi.org/10.1016/j.imavis.2017.02.001.

[113] J. Russell. "A Circumplex Model of Affect". In: *Journal of Personality and Social Psychology* 39 (Dec. 1980), pp. 1161–1178. DOI: 10.1037/h0077714.

[114] J. R. Bergstrom and A. Schall. *Eye Tracking in User Experience Design*. 1st. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2014.

[115] P. van Schaik and J. Ling. "The role of context in perceptions of the aesthetics of web pages over time". In: *International Journal of Human-Computer Studies* 67.1 (2009), pp. 79–89. DOI: https://doi.org/10.1016/j.ijhcs.2008.09.012.

[116] S. Gwak and K. Park. "Designing Effective Visual Feedback for Facial Rehabilitation Exercises: Investigating the Role of Shape, Transparency, and Age on User Experience". In: *Healthcare* 11 (June 2023), p. 1835. DOI: 10.3390/healthcare11131835.

[117] A. Mollahosseini, B. Hasani, and M. H. Mahoor. "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild". In: *IEEE Transactions on Affective Computing* 10.1 (Jan. 2019), pp. 18–31. DOI: 10.1109/taffc.2017.2740923.

[118] A. P. O. S. Vermeeren, E. L.-C. Law, V. Roto, M. Obrist, J. Hoonhout, and K. Väänänen-Vainio-Mattila. "User Experience Evaluation Methods: Current State and Development Needs". In: *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*. NordiCHI '10. Reykjavik, Iceland: Association for Computing Machinery, 2010, pp. 521–530. DOI: 10.1145/1868914.1868973.

[119] J. Bölte, T. Hösker, G. Hirschfeld, and M. Thielsch. “Electrophysiological correlates of aesthetic processing of webpages: A comparison of experts and laypersons”. In: *PeerJ* 5 (June 2017), e3440. DOI: 10.7717/peerj.3440.

[120] M. Moshagen and M. T. Thielsch. “Facets of visual aesthetics”. In: *International Journal of Human-Computer Studies* 68.10 (2010), pp. 689–709. DOI: https://doi.org/10.1016/j.ijhcs.2010.05.006.

[121] N. Tractinsky, A. Katz, and D. Ikar. “What is beautiful is usable”. In: *Interacting with Computers* 13.2 (2000), pp. 127–145. DOI: https://doi.org/10.1016/S0953-5438(00)00031-X.

[122] A. N. Tuch, S. P. Roth, K. Hornbæk, K. Opwis, and J. A. Bargas-Avila. “Is beautiful really usable? Toward understanding the relation between usability, aesthetics, and affect in HCI”. In: *Computers in Human Behavior* 28.5 (2012), pp. 1596–1607. DOI: https://doi.org/10.1016/j.chb.2012.03.024.

[123] A. Miniukovich and A. De Angeli. “Computation of Interface Aesthetics”. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI ’15. Seoul, Republic of Korea: Association for Computing Machinery, 2015, pp. 1163–1172. DOI: 10.1145/2702123.2702575.

[124] N. Tractinsky, A. Cokhavi, M. Kirschenbaum, and T. Sharfi. “Evaluating the consistency of immediate aesthetic perceptions of web pages”. In: *International Journal of Human-Computer Studies* 64.11 (2006), pp. 1071–1083. DOI: https://doi.org/10.1016/j.ijhcs.2006.06.009.

[125] J. Wang, Y. Liu, Y. Wang, J. Mao, T. Yue, and F. You. “SAET: The Non-Verbal Measurement Tool in User Emotional Experience”. In: *Applied Sciences* 11.16 (2021). DOI: 10.3390/app11167532.

[126] K. Reinecke and K. Z. Gajos. "Quantifying Visual Preferences around the World". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '14. Toronto, Ontario, Canada: Association for Computing Machinery, 2014, pp. 11–20. DOI: 10.1145/2556288.2557052.

[127] Y. Cai, X. Li, and J. Li. "Emotion Recognition Using Different Sensors, Emotion Models, Methods and Datasets: A Comprehensive Review". In: *Sensors* 23.5 (2023). DOI: 10.3390/s23052455.

[128] N. Chettaoui and M. S. Bouhlel. "I2Evaluator: An Aesthetic Metric-Tool for Evaluating the Usability of Adaptive User Interfaces". In: *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2017, AISI 2017, Cairo, Egypt, September 9-11, 2017*. Ed. by A. E. Hassanien, K. Shaalan, T. Gaber, and M. F. Tolba. Vol. 639. Advances in Intelligent Systems and Computing. Springer, 2017, pp. 374–383. DOI: 10.1007/978-3-319-64861-3_35.

[129] J. M. Garcia-Garcia, V. M. R. Penichet, and M. D. Lozano. "Emotion Detection: A Technology Review". In: *Proceedings of the XVIII International Conference on Human Computer Interaction*. Interacción '17. Cancun, Mexico: Association for Computing Machinery, 2017. DOI: 10.1145/3123818.3123852.

[130] M. Ninaus, S. Greipl, K. Kiili, A. Lindstedt, S. Huber, E. Klein, H.-O. Karnath, and K. Moeller. "Increased emotional engagement in game-based learning – A machine learning approach on facial emotion detection data". In: *Computers & Education* 142 (2019), p. 103641. DOI: https://doi.org/10.1016/j.compedu.2019.103641.

[131] T. Thanapattheerakul, K. Mao, J. Amoranto, and J. H. Chan. "Emotion in a Century: A Review of Emotion Recognition". In: *Proceedings of the 10th International Conference on Advances in Information Technology*. IAIT

2018. Bangkok, Thailand: Association for Computing Machinery, 2018. DOI: 10.1145/3291280.3291788.

[132] A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic. "Estimation of continuous valence and arousal levels from faces in naturalistic conditions". In: *Nature Machine Intelligence* 3 (Jan. 2021). DOI: 10.1038/s42256-020-00280-0.

[133] S. Minaee, M. Minaei, and A. Abdolrashidi. "Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network". In: *Sensors* 21 (Apr. 2021), p. 3046. DOI: 10.3390/s21093046.

[134] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. 2010, pp. 94–101. DOI: 10.1109/CVPRW.2010.5543262.

[135] D. Aneja, A. Colburn, G. Faigin, L. Shapiro, and B. Mones. "Modeling Stylized Character Expressions via Deep Learning". In: *Computer Vision – ACCV 2016*. Ed. by S.-H. Lai, V. Lepetit, K. Nishino, and Y. Sato. Cham: Springer International Publishing, 2017, pp. 136–153.

[136] S. Cheng and A. K. Dey. "I see, you design: user interface intelligent design system with eye tracking and interactive genetic algorithm". In: *CCF Trans. Perv. Comput. Int.* 1.3 (2019), pp. 224–236.

[137] P. Emami, Y. Yiang, Z. Guo, and L. A. Leiva. "Impact of Design Decisions in Scanpath Modeling". In: *Proceedings of the ACM Symposium on Eye Tracking Research an Applications (ETRA)*. 2024.

[138] Y. M. Hwang and K. C. Lee. "An eye-tracking paradigm to explore the effect of online consumers' emotion on their visual behaviour between desktop screen and mobile screen". In: *Behaviour & Information Technology* 41.3 (2022), pp. 535–546.

[139] J. Z. Lim, J. Mountstephens, and J. Teo. "Emotion recognition using eye-tracking: taxonomy, review and current challenges". In: *Sensors* 20.8 (2020), p. 2384.

[140] T. Partala and V. Surakka. "Pupil size variation as an indication of affective processing". In: *International Journal of Human-Computer Studies* 59.1 (2003). Applications of Affective Computing in Human-Computer Interaction, pp. 185–198. DOI: https://doi.org/10.1016/S1071-5819(03)00017-X.

[141] M. Oliva and A. Anikin. "Pupil dilation reflects the time course of emotion recognition in human vocalizations". In: *Scientific Reports* 8 (Mar. 2018). DOI: 10.1038/s41598-018-23265-x.

[142] W. Klimesch. "EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis". In: *Brain Research Reviews* 29.2 (1999), pp. 169–195. DOI: https://doi.org/10.1016/S0165-0173(98)00056-3.

[143] C. S. Nayak and A. C. Anilkumar. *EEG Normal Waveforms*. StatPearls Publishing, Treasure Island (FL), 2021.

[144] L. S. Mokatren, R. Ansari, A. E. Cetin, A. D. Leow, O. A. Ajilore, H. Klumpp, and F. T. Yarman Vural. "EEG Classification by Factoring in Sensor Spatial Configuration". In: *IEEE Access* 9 (2021), pp. 19053–19065. DOI: 10.1109/ACCESS.2021.3054670.

[145] J. Cheng, M. Chen, C. Li, Y. Liu, R. Song, A. Liu, and X. Chen. "Emotion Recognition From Multi-Channel EEG via Deep Forest". In: *IEEE Journal of Biomedical and Health Informatics* 25.2 (2021), pp. 453–464. DOI: 10.1109/JBHI.2020.2995767.

[146] K. Chengeta. "Comparative Analysis of Emotion Detection from Facial Expressions and Voice Using Local Binary Patterns and Markov Models:

Computer Vision and Facial Recognition". In: *Proceedings of the 2nd International Conference on Vision, Image and Signal Processing*. ICVISP 2018. Las Vegas, NV, USA: Association for Computing Machinery, 2018. DOI: 10.1145/3271553.3271574.

[147] S. Haddad, O. Daassi, and S. Belghith. "Emotion Recognition from Audio-Visual Information based on Convolutional Neural Network". In: *2023 International Conference on Control, Automation and Diagnosis (ICCAD)*. 2023, pp. 1–5. DOI: 10.1109/ICCAD57653.2023.10152451.

[148] S. Luo, Y.-T. Lan, D. Peng, Z. Li, W.-L. Zheng, and B.-L. Lu. "Multimodal Emotion Recognition in Response to Oil Paintings". In: *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 2022, pp. 4167–4170. DOI: 10.1109/EMBC48229.2022.9871630.

[149] P. Viola and M. Jones. "Rapid object detection using a boosted cascade of simple features". In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*. Vol. 1. 2001, pp. I–I. DOI: 10.1109/CVPR.2001.990517.

[150] A. Delorme and S. Makeig. "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis". In: *Journal of neuroscience methods* 134.1 (2004), pp. 9–21.

[151] R. Reber, N. Schwarz, and P. Winkielman. "Processing Fluency and Aesthetic Pleasure: Is Beauty in the Perceiver's Processing Experience?" In: *Personality and Social Psychology Review* 8.4 (2004). PMID: 15582859, pp. 364–382. DOI: 10.1207/s15327957pspr0804_3.

[152] M. Borys and M. Plechawska-Wójcik. "Eye-tracking metrics in perception and visual attention research". In: *EJMT* 3 (2017), pp. 11–23.

[153] J. Kim, P. Thomas, R. Sankaranarayana, T. Gedeon, and H.-J. Yoon. "Eye-tracking analysis of user behavior and performance in web search on large and small screens". In: *Journal of the Association for Information Science and Technology* 66.3 (2015), pp. 526–544.

[154] A. Sutcliffe and A. Namoun. "Predicting user attention in complex web pages". In: *Behaviour & Information Technology* 31.7 (2012), pp. 679–695.

[155] G. Ziv. "Gaze behavior and visual attention: A review of eye tracking studies in aviation". In: *The International Journal of Aviation Psychology* 26.3-4 (2016), pp. 75–104.

[156] I. Arapakis and L. A. Leiva. "Predicting user engagement with direct displays using mouse cursor information". In: *Proc. 39th International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR)*. 2016, pp. 599–608.

[157] J. Huang, R. White, and G. Buscher. "User see, user point: gaze and cursor alignment in web search". In: *Proc. SIGCHI conference on human factors in computing systems (CHI)*. 2012, pp. 1341–1350.

[158] J. Huang, R. W. White, and S. Dumais. "No clicks, no problem: using cursor movements to understand and improve search". In: *Proc. SIGCHI conference on human factors in computing systems (CHI)*. 2011, pp. 1225–1234.

[159] L. A. Leiva and I. Arapakis. "The Attentive Cursor Dataset". In: *Frontiers in Human Neuroscience* 14 (2020), p. 565664.

[160] I. Arapakis and L. A. Leiva. "Learning efficient representations of mouse movements to predict user attention". In: *Proc. 43rd international ACM SIGIR conference on research and development in information retrieval (SIGIR)*. 2020, pp. 1309–1318.

[161] N. Gao, M. Saiedur Rahaman, W. Shao, and F. D. Salim. “Investigating the reliability of self-report data in the wild: The quest for ground truth”. In: *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*. 2021, pp. 237–242.

[162] E. Vraga, L. Bode, and S. Troller-Renfree. “Beyond self-reports: Using eye tracking to measure topic and style differences in attention to social media content”. In: *Communication Methods and Measures* 10.2-3 (2016), pp. 149–164.

[163] L. A. Leiva, I. Arapakis, and C. Iordanou. “My mouse, my rules: Privacy issues of behavioral user profiling via mouse tracking”. In: *Proc. Conference on Human Information Interaction and Retrieval (CHIIR)*. 2021, pp. 51–61.

[164] P. R. Houssel and L. A. Leiva. “User Re-Authentication via Mouse Movements and Recurrent Neural Networks.” In: *ICISSP*. 2024, pp. 652–659.

[165] A. Chuklin and M. de Rijke. “Incorporating clicks, attention and satisfaction into a search engine result page evaluation model”. In: *Proc. 25th ACM international on conference on information and knowledge management (CIKM)*. 2016, pp. 175–184.

[166] Y. Liu, Y. Chen, J. Tang, J. Sun, M. Zhang, S. Ma, and X. Zhu. “Different users, different opinions: Predicting search satisfaction with mouse movement information”. In: *Proc. 38th international ACM SIGIR conference on research and development in information retrieval (SIGIR)*. 2015, pp. 493–502.

[167] S. SadighZadeh and M. Kaedi. “Modeling user preferences in online stores based on user mouse behavior on page elements”. In: *Journal of Systems and Information Technology* 24.2 (2022), pp. 112–130.

[168] J. Schneider, M. Weinmann, J. vom Brocke, and C. Schneider. “Identifying Preferences through mouse cursor movements-Preliminary Evidence.” In: *Proc. ECIS*. 2017, pp. 2546–2556.

[169] J. Arguello. “Predicting search task difficulty”. In: *Proc. European conference on information retrieval (ECIR)*. 2014, pp. 88–99.

[170] V. Navalpakkam and E. Churchill. “Mouse tracking: measuring and predicting users’ experience of web-based content”. In: *Proc. SIGCHI conference on human factors in computing systems (CHI)*. 2012, pp. 2963–2972.

[171] L. Brückner, I. Arapakis, and L. A. Leiva. “When Choice Happens: A Systematic Examination of Mouse Movement Length for Decision Making in Web Search”. In: *Proc. 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 2021, pp. 2318–2322.

[172] T. Katerina, P. Nicolaos, and Y. Charalampos. “Mouse tracking for web marketing: enhancing user experience in web application software by measuring self-efficacy and hesitation levels”. In: *Int. J. Strateg. Innovative Mark* 1 (2014), pp. 233–247.

[173] E. Y. Fu, T. C. Kwok, E. Y. Wu, H. V. Leong, G. Ngai, and S. C. Chan. “Your mouse reveals your next activity: towards predicting user intention from mouse interaction”. In: *Proc. IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*. Vol. 1. 2017, pp. 869–874.

[174] G. Zhang, Z. Hu, M. Bâce, and A. Bulling. “Mouse2Vec: Learning Reusable Semantic Representations of Mouse Behaviour”. In: *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 2024, pp. 1–17.

[175] M. D. Smucker, X. S. Guo, and A. Toulis. “Mouse movement during relevance judging: implications for determining user attention”. In: *Proc. 37th international ACM SIGIR conference on Research & development in information retrieval (SIGIR)*. 2014, pp. 979–982.

[176] Z. Liu, J. Mao, C. Wang, Q. Ai, Y. Liu, and J.-Y. Nie. “Enhancing click models with mouse movement information”. In: *Information Retrieval Journal* 20 (2017), pp. 53–80.

[177] I. Kirsh, M. Joy, and Y. Kirsh. “Horizontal mouse movements (HMMs) on web pages as indicators of user interest”. In: *International Conference on Human-Computer Interaction (HCII)*. 2020, pp. 416–423.

[178] I. Kirsh. “Directions and speeds of mouse movements on a website and reading patterns: a web usage mining case study”. In: *Proc. 10th International Conference on Web Intelligence, Mining and Semantics (WIMS)*. 2020, pp. 129–138.

[179] A. K. Jaiswal, P. Tiwari, and M. S. Hossain. “Predicting users’ behavior using mouse movement information: an information foraging theory perspective”. In: *Neural Computing and Applications* 35.33 (2023), pp. 23767–23780.

[180] A. Johnson, B. Mulder, A. Sijbinga, and L. Hulsebos. “Action as a window to perception: measuring attention with mouse movements”. In: *Applied Cognitive Psychology* 26.5 (2012), pp. 802–809.

[181] P. Boi, G. Fenu, L. D. Spano, and V. Vargiu. “Reconstructing user’s attention on the web through mouse movements and perception-based content identification”. In: *ACM Transactions on Applied Perception* 13.3 (2016), pp. 1–21.

[182] A. Milisavljevic, F. Abate, T. Le Bras, B. Gosselin, M. Mancas, and K. Doré-Mazars. “Similarities and differences between eye and mouse dynamics during web pages exploration”. In: *Frontiers in Psychology* 12 (2021), p. 554595.

[183] A. Milisavljevic, K. Hamard, C. Petermann, B. Gosselin, K. Doré-Mazars, and M. Mancas. “Eye and mouse coordination during task: from behaviour to prediction”. In: *International Conference on Human Computer Interaction Theory and Applications*. 2018, pp. 86–93.

[184] Q. Guo and E. Agichtein. “Towards predicting web searcher gaze position from mouse movements”. In: *Proc. Extended Abstracts on human factors in computing systems (CHI EA)*. 2010, pp. 3601–3606.

[185] J. Mao, Y. Liu, M. Zhang, and S. Ma. “Estimating credibility of user clicks with mouse movement and eye-tracking information”. In: *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer. 2014, pp. 263–274.

[186] Y. Liu, Z. Liu, K. Zhou, M. Wang, H. Luan, C. Wang, M. Zhang, and S. Ma. “Predicting search user examination with visual saliency”. In: *Proc. 39th International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR)*. 2016, pp. 619–628.

[187] Y. Chen, Y. Liu, M. Zhang, and S. Ma. “User satisfaction prediction with mouse movement information in heterogeneous search environment”. In: *IEEE Transactions on Knowledge and Data Engineering* 29.11 (2017), pp. 2470–2483.

[188] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel. “Image-Based Recommendations on Styles and Substitutes”. In: *Proc. 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 2015, pp. 43–52.

[189] R. N. Meghanathan, C. van Leeuwen, and A. R. Nikolaev. “Fixation duration surpasses pupil size as a measure of memory load in free viewing”. In: *Frontiers in human neuroscience* 8 (2015), p. 1063.

[190] S. Negi and R. Mitra. “Fixation duration and the learning process: An eye tracking study with subtitled videos”. In: *Journal of Eye Movement Research* 13.6 (2020).

[191] S. Chakraborty, Z. Wei, C. Kelton, S. Ahn, A. Balasubramanian, G. J. Zelinsky, and D. Samaras. “Predicting Visual Attention in Graphic De-

sign Documents". In: *IEEE Transactions on Multimedia* 25 (2023). DOI: 10.1109/tmm.2022.3176942.

[192] J. Gleason, A. Koeninger, D. Hu, J. Teurn, Y. Bart, S. Knight, R. E. Robertson, and C. Wilson. "Search Engine Revenue from Navigational and Brand Advertising". In: *Proceedings of the Eighteenth International AAAI Conference on Web and Social Media*. Vol. 18. 2024. DOI: 10.1609/icwsm.v18i1.31329.

[193] L. Brückner, I. Arapakis, and L. A. Leiva. "When Choice Happens: A Systematic Examination of Mouse Movement Length for Decision Making in Web Search". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2021. DOI: 10.1145/3404835.3463055.

[194] A. Chuklin and M. de Rijke. "Incorporating Clicks, Attention and Satisfaction into a Search Engine Result Page Evaluation Model". In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. CIKM '16. 2016. DOI: 10.1145/2983323.2983829.

[195] I. Arapakis, A. Penta, H. Joho, and L. A. Leiva. "A Price-per-attention Auction Scheme Using Mouse Cursor Information". In: *ACM Transactions on Information Systems* 38.2 (Jan. 2020). DOI: 10.1145/3374210.

[196] I. Arapakis and L. A. Leiva. "Predicting User Engagement with Direct Displays Using Mouse Cursor Information". In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '16. 2016. DOI: 10.1145/2911451.2911505.

[197] Q. Guo and E. Agichtein. "Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior". In: *Proceedings of the 21st International Conference on World Wide Web*. WWW '12. 2012. DOI: 10.1145/2187836.2187914.

[198] M. Speicher, A. Both, and M. Gaedke. “TellMyRelevance! predicting the relevance of web search results from cursor interactions”. In: *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*. CIKM ’13. 2013. DOI: 10.1145/2505515.2505703.

[199] J. Gwizdka, R. Tessmer, Y.-C. Chan, K. Radhakrishnan, and M. L. Henry. “Eye-Gaze and Mouse-Movements on Web Search as Indicators of Cognitive Impairment”. In: *Information Systems and Neuroscience*. LNISO. 2022. DOI: 10.1007/978-3-031-13064-9_20.

[200] M. Maldonado, E. Dunbar, and E. Chemla. “Mouse tracking as a window into decision making”. In: *Behavior Research Methods* 51.3 (2019). DOI: 10.3758/s13428-018-01194-x.

[201] M. W. Smith, J. Sharit, and S. J. Czaja. “Aging, Motor Control, and the Performance of Computer Mouse Tasks”. In: *Human Factors* 41.3 (1999). DOI: 10.1518/001872099779611102.

[202] D. Soman. “The illusion of delayed incentives: Evaluating future effort-money transactions”. In: *J. Mark. Res.* 35 (1998), pp. 427–438.

[203] P. E. Stillman, X. Shen, and M. J. Ferguson. “How Mouse-tracking Can Advance Social Cognitive Theory”. In: *Trends in Cognitive Sciences* 22.6 (2018). DOI: 10.1016/j.tics.2018.03.012.

[204] G. Zauberman. “The intertemporal dynamics of consumer lock-in”. In: *J. Consum. Res.* 30 (2003), pp. 405–419.

[205] I. Arapakis, M. Lalmas, and G. Valkanas. “Understanding within-content engagement through pattern analysis of mouse gestures”. In: *Proc. 23rd ACM International Conference on conference on Information and Knowledge Management (CIKM)*. 2014, pp. 1439–1448.

[206] Brightfish, Profacts, and Lumen. *From viewable to viewed: using eye tracking to understand the reality of attention to advertising across media*. White paper. Retrieved on October 10, 2019. Available at https://effectiveviews.be/files/White_Paper_From_Viewable_to_viewed.pdf. 2018.

[207] J. H. Goldberg, M. J. Stimson, M. Lewenstein, N. Scott, and A. M. Wichansky. "Eye tracking in web search tasks: design implications". In: *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications*. ETRA '02. 2002. DOI: 10.1145/507072.507082.

[208] Y. Li, P. Xu, D. Lagun, and V. Navalpakkam. "Towards Measuring and Inferring User Interest from Gaze". In: *Proceedings of the 26th International Conference on World Wide Web Companion*. WWW '17 Companion. 2017. DOI: 10.1145/3041021.3054182.

[209] A. Papoutsaki, P. Sangkloy, J. Laskey, N. Daskalova, J. Huang, and J. Hays. "Webgazer: scalable webcam eye tracking using user interactions". In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. IJCAI'16. 2016.

[210] L. A. Leiva, I. Arapakis, and C. Iordanou. "My Mouse, My Rules: Privacy Issues of Behavioral User Profiling via Mouse Tracking". In: *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. CHIIR '21. 2021. DOI: 10.1145/3406522.3446011.

[211] R. Atterer, M. Wnuk, and A. Schmidt. "Knowing the user's every move: user activity tracking for website usability evaluation and implicit interaction". In: *Proceedings of the 15th International Conference on World Wide Web*. WWW '06. 2006. DOI: 10.1145/1135777.1135811.

[212] I. Arapakis, M. Lalmas, and G. Valkanas. "Understanding Within-Content Engagement through Pattern Analysis of Mouse Gestures". In: *Proceedings of the 23rd ACM International Conference on Conference on Information*

and Knowledge Management. CIKM '14. 2014. DOI: 10.1145/2661829.2661909.

[213] Q. Guo and E. Agichtein. "Exploring mouse movements for inferring query intent". In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '08. 2008. DOI: 10.1145/1390334.1390462.

[214] D. Martín-Albo, L. A. Leiva, J. Huang, and R. Plamondon. "Strokes of insight: User intent detection and kinematic compression of mouse cursor trails". In: *Information Processing & Management* 52.6 (2016). DOI: 10.1016/j.ipm.2016.04.005.

[215] A. Diriye, R. White, G. Buscher, and S. Dumais. "Leaving So Soon? Understanding and Predicting Web Search Abandonment Rationales". In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. CIKM '12. 2012. DOI: 10.1145/2396761.2398399.

[216] H. A. Feild, J. Allan, and R. Jones. "Predicting searcher frustration". In: *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '10. 2010. DOI: 10.1145/1835449.1835458.

[217] M. Claypool, P. Le, M. Wased, and D. Brown. "Implicit interest indicators". In: *Proceedings of the 6th International Conference on Intelligent User Interfaces*. IUI '01. 2001. DOI: 10.1145/359784.359836.

[218] B. Shapira, M. Taieb-Maimon, and A. Moskowitz. "Study of the usefulness of known and new implicit indicators and their optimal combination for accurate inference of users interests". In: *Proceedings of the 2006 ACM Symposium on Applied Computing*. SAC '06. 2006. DOI: 10.1145/1141277.1141542.

[219] Q. Guo and E. Agichtein. "Ready to buy or just browsing? detecting web searcher goals from interaction data". In: *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '10. 2010. DOI: 10.1145/1835449.1835473.

[220] Y. Liu, Y. Chen, J. Tang, J. Sun, M. Zhang, S. Ma, and X. Zhu. "Different Users, Different Opinions: Predicting Search Satisfaction with Mouse Movement Information". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '15. 2015. DOI: 10.1145/2766462.2767721.

[221] J. Huang, R. W. White, G. Buscher, and K. Wang. "Improving searcher models using mouse cursor activity". In: *Proc. 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR)*. 2012, pp. 195–204.

[222] D. Lagun, M. Ageev, Q. Guo, and E. Agichtein. "Discovering common motifs in cursor movement data for improving web search". In: *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*. WSDM '14. 2014. DOI: 10.1145/2556195.2556265.

[223] D. Hauger, A. Paramythis, and S. Weibelzahl. "Using browser interaction data to determine page reading behavior". In: *Proceedings of the 19th International Conference on User Modeling, Adaption, and Personalization*. UMAP'11. 2011.

[224] F. Mueller and A. Lockerd. "Cheese: tracking mouse movement activity on websites, a tool for user modeling". In: *CHI '01 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '01. 2001. DOI: 10.1145/634067.634233.

[225] P. Boi, G. Fenu, L. D. Spano, and V. Vargiu. "Reconstructing User's Attention on the Web Through Mouse Movements and Perception-Based

Content Identification". In: *ACM Transactions on Applied Perception* 13.3 (2016). DOI: 10.1145/2912124.

[226] I. Arapakis, L. A. Leiva, and B. B. Cambazoglu. "Know Your Onions: Understanding the User Experience with the Knowledge Module in Web Search". In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. CIKM '15. 2015. DOI: 10.1145/2806416.2806591.

[227] N. Roy, D. Maxwell, and C. Hauff. "Users and Contemporary SERPs: A (Re-)Investigation". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '22. 2022. DOI: 10.1145/3477495.3531719.

[228] Y. Shao, J. Mao, Y. Liu, M. Zhang, and S. Ma. "From linear to non-linear: investigating the effects of right-rail results on complex SERPs". In: *Advances in Computational Intelligence* 2.1 (2022). DOI: 10.1007/s43674-021-00028-2.

[229] F. Diaz, R. White, G. Buscher, and D. Liebling. "Robust models of mouse movement on dynamic web search results pages". In: *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*. CIKM '13. 2013. DOI: 10.1145/2505515.2505717.

[230] Y. Liu, C. Wang, K. Zhou, J. Nie, M. Zhang, and S. Ma. "From Skimming to Reading: A Two-stage Examination Model for Web Search". In: *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. CIKM '14. 2014. DOI: 10.1145/2661829.2661907.

[231] I. Arapakis and L. A. Leiva. "Learning Efficient Representations of Mouse Movements to Predict User Attention". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '20. 2020. DOI: 10.1145/3397271.3401031.

[232] K. Latifzadeh, J. Gwizdka, and L. A. Leiva. “A Versatile Dataset of Mouse and Eye Movements on Search Engine Results Pages”. In: *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2025, pp. 3412–3421.

[233] M. Cognolato, M. Atzori, and H. Müller. “Head-mounted eye gaze tracking devices: An overview of modern devices and recent advances”. In: *Journal of rehabilitation and assistive technologies engineering* 5 (2018), p. 2055668318773991.

[234] C. Hafner, V. Scharner, M. Hermann, P. Metelka, B. Hurch, D. A. Klaus, W. Schaubmayr, M. Wagner, A. Gleiss, H. Willschke, et al. “Eye-tracking during simulation-based echocardiography: a feasibility study”. In: *BMC Medical Education* 23.1 (2023), p. 490.

[235] Z. Huang, X. Duan, G. Zhu, S. Zhang, R. Wang, and Z. Wang. “Assessing the data quality of AdHawk MindLink eye-tracking glasses”. In: *Behavior Research Methods* (2024), pp. 1–17.

[236] J. Meyer, A. Frank, T. Schlebusch, and E. Kasneci. “U-har: A convolutional approach to human activity recognition combining head and eye movements for context-aware smart glasses”. In: *Proceedings of the ACM on Human-Computer Interaction* 6.ETRA (2022), pp. 1–19.

[237] V. Onkhar, D. Dodou, and J. De Winter. “Evaluating the Tobii Pro Glasses 2 and 3 in static and dynamic conditions”. In: *Behavior Research Methods* 56.5 (2024), pp. 4221–4238.

[238] N. V. Valtakari, I. T. Hooge, C. Viktorsson, P. Nyström, T. Falck-Ytter, and R. S. Hessels. “Eye tracking in human interaction: Possibilities and limitations”. In: *Behavior Research Methods* (2021), pp. 1–17.

[239] M. O. Al-Moteri, M. Symmons, V. Plummer, and S. Cooper. “Eye tracking to investigate cue processing in medical decision-making: A scoping review”. In: *Computers in Human Behavior* 66 (2017), pp. 52–66.

[240] D. A. Hofmaenner, A. Herling, S. Klinzing, S. Wegner, Q. Lohmeyer, R. A. Schuepbach, and P. K. Buehler. "Use of eye tracking in analyzing distribution of visual attention among critical care nurses in daily professional life: an observational study". In: *Journal of Clinical Monitoring and Computing* 35 (2021), pp. 1511–1518.

[241] E. Capogna, F. Salvi, L. Delvino, A. Di Giacinto, and M. Velardo. "Novice and expert anesthesiologists' eye-tracking metrics during simulated epidural block: a preliminary, brief observational report". In: *Local and Regional Anesthesia* (2020), pp. 105–109.

[242] H.-E. Chen, R. R. Bhide, D. F. Pepley, C. C. Sonntag, J. Z. Moore, D. C. Han, and S. R. Miller. "Can eye tracking be used to predict performance improvements in simulated medical training? A case study in central venous catheterization". In: *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*. Vol. 8. 1. SAGE Publications Sage CA: Los Angeles, CA. 2019, pp. 110–114.

[243] J. Mercier, O. Ertz, and E. Bocher. "Quantifying Dwell Time With Location-based Augmented Reality: Dynamic AOI Analysis on Mobile Eye Tracking Data With Vision Transformer". In: *Journal of Eye Movement Research* 17.3 (2024).

[244] I. Tanoubi, M. Tourangeau, K. Sodoké, R. Perron, P. Drolet, M.-È. Bélanger, J. Morris, C. Ranger, M.-R. Paradis, A. Robitaille, et al. "Comparing the visual perception according to the performance using the eye-tracking technology in high-fidelity simulation settings". In: *Behavioral Sciences* 11.3 (2021), p. 31.

[245] C. M. Privitera and L. W. Stark. "Algorithms for defining visual regions-of-interest: Comparison with eye fixations". In: *IEEE Transactions on pattern analysis and machine intelligence* 22.9 (2000), pp. 970–982.

[246] N. W. Rim, K. W. Choe, C. Scrivner, and M. G. Berman. “Introducing Point-of-Interest as an alternative to Area-of-Interest for fixation duration analysis”. In: *PLoS One* 16.5 (2021), e0250170.

[247] D. S. Wooding. “Eye movements of large populations: II. Deriving regions of interest, coverage, and similarity using fixation maps”. In: *Behavior Research Methods, Instruments, & Computers* 34 (2002), pp. 518–528.

[248] E. Mauriz, S. Caloca-Amber, and A. M. Vázquez-Casares. “Using Task-Evoked Pupillary Response to Predict Clinical Performance during a Simulation Training”. In: *Healthcare (Basel)* 11.4 (2023), p. 455. DOI: 10.3390/healthcare11040455.

[249] A. Szulewski, N. Roth, and D. Howes. “The Use of Task-Evoked Pupillary Response as an Objective Measure of Cognitive Load in Novices and Trained Physicians: A New Tool for the Assessment of Expertise”. In: *Academic Medicine* 90.7 (2015). DOI: 10.1097/ACM.0000000000000677.

[250] B. J. Jongkees and L. S. Colzato. “Spontaneous eye blink rate as predictor of dopamine-related cognitive function—A review”. In: *Neuroscience & Biobehavioral Reviews* 71 (2016), pp. 58–82. DOI: https://doi.org/10.1016/j.neubiorev.2016.08.020.

[251] R. Paprocki and A. Lenskiy. “What Does Eye-Blink Rate Variability Dynamics Tell Us About Cognitive Performance?” In: *Frontiers in Human Neuroscience* 11 (2017). DOI: 10.3389/fnhum.2017.00620.

[252] L. Kessler, P. Gröpel, H. Aichner, G. Aspalter, L. Kuster, G. M. Schmölzer, A. Berger, M. Wagner, and B. Simma. “Eye-tracking during simulated endotracheal newborn intubation: a prospective, observational multicenter study”. In: *Pediatric research* 94.2 (2023), pp. 443–449.

[253] N. Ahmadi, F. Sasangohar, J. Yang, D. Yu, V. Danesh, S. Klahn, and F. Masud. “Quantifying workload and stress in intensive care unit nurses:

preliminary evaluation using continuous eye-tracking". In: *Human factors* 66.3 (2024), pp. 714–728.

[254] S. Viriyasiripong, A. Lopez, S. H. Mandava, W. R. Lai, G. C. Mitchell, A. Boonjindasup, M. K. Powers, J. L. Silberstein, and B. R. Lee. "Accelerometer Measurement of Head Movement During Laparoscopic Surgery as a Tool to Evaluate Skill Development of Surgeons". In: *Journal of Surgical Education* 73.4 (2016), pp. 589–594. DOI: https://doi.org/10.1016/j.jsurg.2016.01.008.

[255] Z. Zhao, Z. Zhu, X. Zhang, H. Tang, J. Xing, X. Hu, J. Lu, Q. Peng, and X. Qu. "Atypical head movement during face-to-face interaction in children with autism spectrum disorder". In: *Autism Research* 14.6 (2021), pp. 1197–1208.

[256] O. A. Zobeiri, B. Ostrander, J. Roat, Y. Agrawal, and K. E. Cullen. "Loss of peripheral vestibular input alters the statistics of head movement experienced during natural self-motion". In: *The Journal of physiology* 599.8 (2021), pp. 2239–2254.

[257] D. J. Frank, B. Nara, M. Zavagnin, D. R. Touron, and M. J. Kane. "Validating older adults' reports of less mind-wandering: An examination of eye movements and dispositional influences." In: *Psychology and Aging* 30.2 (2015), p. 266.

[258] S. Huette, A. Mathis, and A. Graesser. "Blink durations reflect mind wandering during reading." In: *CogSci*. 2016.

[259] K. Latifzadeh and L. A. Leiva. "Thalamus: A User Simulation Toolkit for Prototyping Multimodal Sensing Studies". In: *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*. 2025, pp. 109–113.

[260] C. de la Torre-Ortiz, M. M. Spapé, L. Kangassalo, and T. Ruotsalo. "Brain relevance feedback for interactive image generation". In: *Proc. UIST*. 2020.

[261] Y. Zhao, B. Li, Y. Li, J. Zhou, J. Cao, Y. Luo, X. Zhou, C. Yao, L. Shi, and G. Wang. “Rope X: Assistance and Guidance on Jumping Rope Frequency, based on Real-time, Heart Rate Feedback During Exercise”. In: *Adj. Proc. UIST*. 2021.

[262] M. Richardson, M. Durasoff, and R. Wang. “Decoding surface touch typing from hand-tracking”. In: *Proc. UIST*. 2020.

[263] L. A. Leiva and R. Vivó. “Interactive hypervideo visualization for browsing behavior analysis”. In: *Proc. WWW Companion*. 2012.

[264] Y. Sugano, X. Zhang, and A. Bulling. “Aggregaze: Collective estimation of audience attention on public displays”. In: *Proc. UIST*. 2016.

[265] X. Zhang. “Evaluating the Effects of Saccade Types and Directions on Eye Pointing Tasks”. In: *Proc. UIST*. 2021.

[266] A. Borawska, J. Duda, and K. Biercewicz. “Best practices of neurophysiological data collection for media message evaluation in social campaigns”. In: *Procedia Comput. Sci.* 192 (2021).

[267] J. C. de Munck, S. I. Gonçalves, R. Mammoliti, R. M. Heethaar, and F. H. L. da Silva. “Interactions between different EEG frequency bands and their effect on alpha–fMRI correlations”. In: *NeuroImage* 47.1 (2009), pp. 69–76.

[268] D. Kahneman and J. Beatty. “Pupil Diameter and Load on Memory”. In: *Science* 154.3756 (1966).

[269] G. M. Notaro and S. G. Diamond. “Simultaneous EEG, eye-tracking, behavioral, and screen-capture data during online German language learning”. In: *Data Brief* 21 (2018).

[270] P. H. Zimmerman, J. E. Bolhuis, A. Willemsen, E. S. Meyer, and L. P. Noldus. “The Observer XT: A tool for the integration and synchronization of multimodal signals”. In: *Behav. Res. Methods* 41.3 (2009).

[271] E. J. Shah, J. Y. Chow, and M. J. Lee. "Anxiety on Quiet Eye and Performance of Youth Pistol Shooters". In: *J. Sport Exerc. Psychol.* 42.4 (2020).

[272] J. Xue, C. Quan, C. Li, J. Yue, and C. Zhang. "A crucial temporal accuracy test of combining EEG and Tobii eye tracker". In: *Medicine* 96.13 (2017).

[273] M. Ragot, N. Martin, S. Em, N. Pallamin, and J.-M. Diverrez. "Emotion Recognition Using Physiological Signals: Laboratory vs. Wearable Sensors". In: *Proc. AHFE*. 2017.

[274] R. Xiao, C. Ding, and X. Hu. "Time Synchronization of Multimodal Physiological Signals through Alignment of Common Signal Types and Its Technical Considerations in Digital Health". In: *Imaging* 8.5 (2022).

[275] F. Wolling, C. D. Huynh, and K. Van Laerhoven. "IBSync: Intra-body synchronization of wearable devices using artificial ECG landmarks". In: *Proc. ISWC*. 2021.

[276] D. Szajerman, P. Napieralski, and J.-P. Lecointe. "Joint analysis of simultaneous EEG and eye tracking data for video images". In: *Proc. IEEE ISEF*. 2018.

[277] P. Bœkgaard, M. K. Petersen, and J. E. Larsen. "In the twinkling of an eye: Synchronization of EEG and eye tracking based on blink signatures". In: *Proc. CIP Workshop*. 2014.

[278] R. Taib, B. Itzstein, and K. Yu. "Synchronising Physiological and Behavioural Sensors in a Driving Simulator". In: *Proc. ICMI*. 2014.

[279] M. I. Posner. "Orienting of attention". In: *Q. J. Exp. Psychol. (Hove)* 32.1 (1980).

[280] N. Dillen, M. Ilievski, E. Law, L. E. Nacke, K. Czarnecki, and O. Schneider. "Keep calm and ride along: Passenger comfort and anxiety as physiological responses to autonomous driving styles". In: *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020, pp. 1–13.

[281] S. Kim, K. A. E. Patra, A. Kim, K.-P. Lee, A. Segev, and U. Lee. "Sensors know which photos are memorable". In: *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2017, pp. 2706–2713.

[282] X. Peng, C. Meng, X. Xie, J. Huang, H. Chen, and H. Wang. "Detecting challenge from physiological signals: A primary study with a typical game scenario". In: *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 2022, pp. 1–7.

[283] N. J. Shoumy, L.-M. Ang, K. P. Seng, D. M. Rahaman, and T. Zia. "Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals". In: *Journal of Network and Computer Applications* 149 (2020), p. 102447.

[284] Y. Sun, F. Guo, F. Kaffashi, F. J. Jacono, M. DeGeorgia, and K. A. Loparo. "INSMA: An integrated system for multimodal data acquisition and analysis in the intensive care unit". In: *Journal of biomedical informatics* 106 (2020), p. 103434.

[285] R. Murray-Smith, A. Oulasvirta, A. Howes, J. Müller, A. Ikkala, M. Bachinski, A. Fleig, F. Fischer, and M. Klar. "What simulation can do for HCI research". In: *Interactions* 29.6 (2022), pp. 48–53.

[286] A. Riegler, A. Riener, and C. Holzmann. "AutoWSD: Virtual reality automated driving simulator for rapid HCI prototyping". In: *Proceedings of Mensch und Computer 2019*. 2019, pp. 853–857.

[287] E. Y. Kimchi, B. F. Coughlin, B. E. Shanahan, G. Piantoni, J. Pezaris, and S. S. Cash. "OpBox: Open source tools for simultaneous EEG and EMG acquisition from multiple subjects". In: *eNeuro* 7.5 (2020).

[288] P. Lopes, L. L. Chuang, and P. Maes. "Physiological I/O". In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–4.

[289] S. Katsigiannis and N. Ramzan. “DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices”. In: *IEEE J. Biomed. Health Inform.* 22.1 (2017), pp. 98–107.

[290] A. Savitzky and M. J. Golay. “Smoothing and differentiation of data by simplified least squares procedures.” In: *Analytical chemistry* 36.8 (1964), pp. 1627–1639.

[291] Q. Li, R. Li, K. Ji, and W. Dai. “Kalman filter and its application”. In: *2015 8th International Conference on Intelligent Networks and Intelligent Systems (ICINIS)*. IEEE. 2015, pp. 74–77.