

NORTHWESTERN UNIVERSITY

Multi-modal wearable camera systems for AI-driven detection and prediction of
health-risk behaviors and outcomes

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Computer Science
TR Number: NU-CS-2026-04

By

Glenn Jacob Fernandes

EVANSTON, ILLINOIS

Dec 2025

© Copyright by Glenn Jacob Fernandes, 2025

All Rights Reserved

Abstract

Wearable and mobile sensing technologies offer a promising path to unobtrusively monitor and intervene on health-risk behaviors such as overeating and smoking in everyday life. Yet real-world deployment faces persistent barriers: reliably detecting health-risk behaviors in the wild, protecting wearer and bystander privacy, and ensuring that AI-driven predictions are interpretable and actionable for clinicians and end users. This work systematically addresses these barriers through a cohesive framework that combines sensing, modeling, privacy, and explanation methods to capture and interpret health-risk behaviors from multimodal wearable data.

To address the challenge of reliably detecting health-risk behaviors in the wild, we present *HabitSense* (Chapter 2), a neck-worn, activity-oriented camera system that integrates RGB, thermal, and inertial sensing (N = 15; 768 h). By orienting its field of view toward the wearer’s mouth and selectively recording hand-to-mouth activity, *HabitSense* captures visual evidence of eating and smoking episodes in everyday settings while maintaining all-day wearability. However, single-modality models trained on *HabitSense* data often misclassified visually similar behaviors, such as eating and smoking. To improve automated detection, we designed *RTFVNet* (Chapter 3), a mid-fusion deep learning architecture that jointly leverages activity-oriented RGB and thermal video to classify smoking, eating, and background activity. On a free-living RGB–thermal dataset (N = 7; 441.9 h), *RTFVNet* achieves robust performance (F1 = 85%) while reducing computational cost relative to dual-model fusion approaches, providing a practical path toward real-time multimodal behavior detection.

Accurate detection alone is insufficient if users are uncomfortable being recorded; we, therefore, designed and developed cartoon-based generative obfuscation techniques

(Chapter 4) that preserve gesture- and context-level information while masking identifiable facial and background details. Across online ($N = 60$) and field ($N = 49$) studies, these methods substantially reduced privacy concerns in high-risk scenarios while maintaining the ability to visually confirm behavioral context in 90% of cases.

While privacy-preserving data collection and accurate behavioral detection provide a necessary foundation, clinical impact depends on whether clinicians can interpret machine-learning (ML) model outputs to guide treatment decisions. Yet many ML models operate as black boxes, limiting clinicians' understanding and trust. To address this gap, we developed *PRIMO* (Chapter 5), an interactive visualization tool for prime-implicant-based explanations of model predictions. In evaluations with weight management experts ($N = 14$), *PRIMO* increased agreement with model predictions and was rated the most helpful method for understanding model behavior.

Collectively, these systems and studies—spanning over 150 participants—demonstrate a pathway from passive multimodal sensing to interpretable, privacy-aware systems for monitoring and supporting health-risk behaviors, advancing the next generation of ethical, transparent, and deployable digital health technologies.

Acknowledgment

I would like to express my deepest gratitude to my advisor, Prof. Nabil Alshurafa, for his exceptional guidance, mentorship, and unwavering support throughout my doctoral journey. His insight, patience, and dedication have shaped not only this dissertation but also my growth as a researcher. I am also sincerely thankful to my committee members, Prof. Aggelos Katsaggelos, Prof. Maia Jacobs and Prof. Karan Ahuja for their thoughtful feedback, encouragement, and meaningful contributions to this work.

I am grateful to my colleagues at the HABits Lab—Farzad Shahbi, Soroush Shahi, Boyang Wei, and Blaine Rothrock—for their collaboration, technical insights, and many thoughtful discussions that shaped this work. I would also like to acknowledge Prof. Mahdi Pedram for his invaluable guidance and support on hardware and firmware development and overall research direction. I am thankful to Bonnie Nolan and Christopher Romano for their assistance with study coordination and data collection, and to Tanmeet Butani for his support with technical infrastructure and server management. I also sincerely appreciate the contributions of Meixi Lu, Jiayi Zhang, and Helen Zhu, whose work as master’s students significantly contributed to this research.

I extend my deepest appreciation to my parents and sister, whose unwavering love, support, and belief in me have been the foundation of all my pursuits. I am also deeply grateful to my extended family for their constant encouragement and support throughout this journey. I am also thankful to my friends and colleagues, including Mara Ullola and Connie Chau, for their kindness, encouragement, and companionship throughout this journey.

List of Abbreviations

AI	Artificial Intelligence
AOC	Activity-Oriented Camera (system)
F1	F1-score
GAN	Generative Adversarial Network
HTM	Hand-to-Mouth
IMU	Inertial Measurement Unit
mHealth	Mobile Health
PRIMO	Prime Implicant Maintenance of Outcome
RQ	Research Question
RGB	Red, Green, Blue (color channels)
RTFVNet	RGB-Thermal Fusion Video Network
SECURE	Sensor-Enabled Control of Ubiquitous Recording and Evaluation
TFLOPs	Tera Floating-Point Operations (per second)

Table of Contents

Abstract	3
Acknowledgment	5
List of Abbreviations	6
Table of Contents	7
List of Figures	10
List of Tables	14
1 Introduction	17
1.1 Motivation	17
1.2 Research Questions and Approach	18
1.3 Integrating Sensing, Modeling, Privacy, and Explainability	23
1.4 Major Contributions	24
2 HabitSense	26
2.1 Introduction and Background	26
2.2 Related Work	31
2.3 Clinician-Informed Design: Integrating Focus Group Feedback into Wear- able Camera Development	38

2.4	Design Implications from Experiential Insights in Wearable Cameras . . .	46
2.5	Methods: Design and Development of the HabitSense System	51
2.6	Results	73
2.7	Discussion and Limitations	86
2.8	Summary	93
2.9	Supplementary	94
3	Modality Fusion	103
3.1	Introduction and Background	103
3.2	Methodology	105
3.3	Experiments	106
3.4	Summary	112
4	Mitigating Privacy Concerns	114
4.1	Introduction and Background	114
4.2	Related Works	117
4.3	Methods	120
4.4	Analysis and Results	124
4.5	Discussion	128
4.6	Summary	130
5	Explainability	132
5.1	Introduction and Background	132
5.2	Methods	137
5.3	Results	151
5.4	Discussion	158
5.5	Summary	162

6	Discussion	166
6.1	Synthesis of Findings	166
6.2	Practical Implications	167
6.3	Limitations	168
6.4	Future Directions	169
7	Conclusion	171
	References	173

List of Figures

2.1	The HabitSense device (left) and its applications (right). HabitSense is a wearable multi-modal system (thermal, RGB, accelerometer) optimized for privacy, enabling all-day recording and on-device processing to detect health-risk behaviors associated with smoking and eating.	26
2.2	Mean reported interest in using the HabitSense system by participant group.	44
2.3	Iterations through different generations (devices and magnetic back plates). Showcasing the transition from Gen1 to Gen2 to HabitSense, this figure reflects the design process of HabitSense, guided by expert insights and user feedback, emphasizing key considerations such as privacy, affordability, and real-time detection. Pink regions indicate parts used for orienting the sensor lenses with the activity of interest. All dimensions are in mm. . . .	47
2.4	Internal design schematic and electronic components used in HabitSense device.	51
2.5	Iterative Design of HabitSense Enclosure. (A) Slim initial 3D design, (B) size comparison with ID card, (C) two-piece hinge-connected enclosure, (D) rotatable internal mechanism for sensor adjustment, (E) revised enclosure with optimized dimensions, (F) final product-quality design.	55
2.6	Sensor-Enabled Activation Algorithm. Outlines the process from power-on to data recording, starting with deep sleep mode, followed by accelerometer checks for the device being worn (level 1), thermal frame sampling for hand-to-mouth gesture detection (level 2), and concluding with obfuscated RGB data recording, designed to record only essential data to optimize battery life and memory use.	59

2.7	HabitSense gesture characterization. Illustrates smoking and drinking gestures in RGB and thermal data, with RGB frames confirming gestures, normalized thermal frames showing temperature distribution, and flattened thermal histograms highlighting the heat signatures during eating and smoking.	61
2.8	SECURE pipeline energy profiling current consumption during different levels.	63
2.9	Sensor Calibration and RGB Obfuscation Process. Depicts calibration via a screen with heated and colored sections for sensor alignment and shows how calibrated thermal data enables obfuscation in RGB sequences, keeping heads, hands, and objects in hands clear while obfuscating the background.	64
2.10	Obfuscation threshold determination.	64
2.11	HabitSense Exploration and Labeling Platform (HELP) web application (left) and the Smoking, Nutrition, and Personal Observation Tool (SNAPSHOT) mobile application (right).	67
2.12	Luminosity threshold determination	71
2.13	(A) Head movement range and fields of view for RGB and thermal cameras, (B) Head movement characterization, including extension and flexion (top), lateral flexion (middle), and rotation (bottom)	73
2.14	SECURE pipeline parameter determination. This figure shows the effects of varying parameter settings on data collection. The graph displays the effectiveness of choosing a threshold of one hand-to-mouth gesture across five segments and a recording duration of 360 frames (1.2 minutes), achieving 89.8% capture of eating and smoking gestures with only 52.4% data recorded.	78
2.15	Performance comparison of TimeSformer models (offline). H2M, hand-to-mouth.	81
2.16	Performance comparison of 2-Layer Neural Network, on-device models. H2M, hand-to-mouth.	81
2.17	Performance of eating and smoking detection TimeSformer models trained on RGB data in different circumstances to highlight robustness	82

2.18	This figure illustrates the results of our incremental data addition experiment designed to evaluate the impact of incorporating additional participant data into our model training.	83
2.19	Participant responses to burden metrics concerning (A) privacy, (B) obtrusiveness, (C) willingness-to-wear, and (D) physical comfort.	87
2.20	Heat signature, cigarette vs e-cigarettes	92
2.21	Distribution of number of frames comprising one eating gesture (mean=10), smoking gesture (mean=12), and hand-to-mouth gestures (mean=10)	96
2.22	Attention maps	97
3.1	RTFVNet Model Architecture	105
3.2	Early fusion inputs for experiment 1 and 2	108
4.1	(a) Study design for the survey. (b) Illustration of a person wearing a necklace form-factor activity-oriented camera with lens oriented upward. The dashed lines represent the field of view of the camera.	117
4.2	Obfuscation Pipeline: The process of generating a cartoonized obfuscated frame from a raw frame.	118
4.3	(a) The mean (95% confidence interval) concern level of participants in each of the conditions (raw, cartoon) given low-risk or high-risk scenarios with questions focused on either the wearer or bystander. The y-axis represents the Likert scale (0, not concerned at all; 4, extremely concerned) (b) The percentage of participants that got the context (of the video shown to them) correct, partially correct, or incorrect. Each column represents the scenario described.	120
4.4	Ranking Question for (a) prolific study and (b) in-lab study. The distribution are displayed for participants' rankings for each of the obfuscation techniques in response to the given video scenarios (range, 1-4).	124
5.1	(a) Tradeoff between the performance of a model (accuracy) and the explainability. (b) Black box machine learning model.	137

5.2	Generating tractable representation of machine learning model to compute prime implicant explanations.	141
5.3	In the 2 PRIMO visuals, the red point indicates the patient profile for basic and complex weight loss archetypes. The complex weight loss archetype tends to be difficult because the red points fall at the edges of the PRIMO-generated prime implicant ranges.	143
5.4	PRIMO software interface.	144
5.5	Study Design	146
5.6	Monotonicity visual	146
5.7	Best-worst case visual	147
5.8	PRIMO software UI	148
5.9	Hypothetical archetypes shown to participant types.	149
5.10	Likert scale analysis for intelligibility questionnaire (mean value and standard error)	154

List of Tables

2.1	Comparison of HabitSense and existing systems	32
2.2	Themes and associated design considerations	40
2.3	Specifications of different generations of devices.	47
2.4	HabitSense hardware design space	52
2.5	Demographic information of participants.	74
2.6	Daytime-Nighttime table	75
2.7	Luminosity table	76
2.8	Head Motion Table	77
2.9	Performance of each SECURE pipeline level of the system based on the leave-one-participant-out evaluation.	79
2.10	Percent of data recorded versus percent of captured gestures and result- ing storage and energy savings using the SECURE pipeline. Utilizing the SECURE pipeline, we captured approximately 89.8% of hand-to-mouth ges- tures while recording only 52.4% of the data, thereby conserving about 47.6% of storage space and achieving 30.0% energy savings.	79
2.11	Energy requirements for each level of the SECURE pipeline.	80
2.12	On-device smoking detection model (thermal data), participant-level anal- ysis.	86
2.13	Hand-to-mouth (H2M) statistics across participants.	98

2.14 eating offdevice rgb obfuscated mask	98
2.15 eating offdevice rgb obfuscated blur	98
2.16 eating-offdevice-rgb	99
2.17 eating-offdevice-thermal	99
2.18 eating-ondevice-thermal	99
2.19 h2m-offdevice-rgb-obfuscated-mask	99
2.20 h2m-offdevice-rgb-obfuscated-blur	100
2.21 h2m-offdevice-rgb	100
2.22 h2m-offdevice-thermal	100
2.23 h2m-ondevice-thermal	100
2.24 smoking-offdevice-rgb-obfuscated-blur	101
2.25 smoking-offdevice-rgb-no-obfuscation	101
2.26 smoking-offdevice-thermal	101
2.27 smoking-ondevice-thermal	101
2.28 smoking-offdevice-rgb-obfuscated-mask	101
2.29 multiclass-offdevice-rgb-obfuscated-mask (weighted-scores)	101
2.30 multiclass-offdevice-rgb-obfuscated-blur (weighted-scores)	102
2.31 multiclass-offdevice-rgb (weighted-scores)	102
2.32 multiclass-offdevice-thermal (weighted-scores)	102
2.33 multiclass-ondevice-thermal (weighted-scores)	102

3.1	Performance metrics (% , mean $\pm$ standard error) for different models, and computational cost (TFLOPs).	107
3.2	Performance metrics of decision fusion experiments.	109
3.3	Performance metrics of decision fusion experiments.	109
3.4	Performance metrics of experiments on cross-modal connections.	110
3.5	Performance metrics of experiments on attention mechanism.	111
3.6	Performance metrics with early and late attention experiments.	112
4.1	U-tests (Wilcoxon rank-sum test). Four hypothesis tests are being conducted, therefore, p-values are set to $\alpha = 0.0125$	120
4.2	Recurring themes captured during the in-lab semi-structured interviews .	126
5.1	Feature Details:	139
5.2	Participant Demographics (Interviewees)	145
5.3	Intelligibility questionnaire. The 5 quantitative questions were presented as adapted from Cahour et al [177].	151
5.4	GLMM Results.	153
5.5	Comparison between overall agree, switch to disagree and switch to agree groups.	154

Chapter 1

Introduction

1.1 Motivation

In the United States, obesity affects approximately 40% of adults [1], while cigarette smoking causes nearly 480,000 deaths annually—approximately one in every five deaths in the United States [2]. Together, overeating and smoking account for a substantial portion of avoidable morbidity and mortality [3]. These health-risk behaviors, rooted in daily routines and habits, highlight the urgency of understanding, monitoring, and intervening on eating and smoking as they unfold in everyday life. Understanding how everyday contexts, routines, and internal states interact to drive these behaviors in real-world settings is crucial for designing timely and personalized health interventions [4].

Despite decades of behavioral and clinical research, much of what we know about eating and smoking is still based on self-report, retrospective surveys, or lab-based studies [5]. These methods are constrained by recall bias and limited ecological validity: individuals may underreport their behaviors or forget key contextual details, and laboratory environments fail to capture the complexity of real-world experiences. As a result, the very moments in which behavior change could be most impactful—when individu-

als decide to eat, smoke, or resist temptation—often remain invisible to researchers and clinicians.

This limitation motivates a new generation of research that leverages wearable and mobile sensing to observe, interpret, and explain health-risk behaviors in the wild. Activity-oriented wearables and body-worn cameras can provide rich, continuous streams of behavioral and contextual signals, enabling objective, fine-grained measurement of behaviors such as eating and smoking in free-living environments [6]. However, turning these raw data streams into practical tools for clinical and public health use raises three intertwined challenges: reliable behavioral detection in the wild [7], protection of wearer and bystander privacy [6], and interpretability and clinical actionability of AI-driven predictions [8].

1.2 Research Questions and Approach

To address these behavioral, privacy, and AI-related challenges, this dissertation focuses on unobtrusively capturing and interpreting health-risk behaviors associated with overeating and smoking in the wild. It does so through four research questions, each corresponding to a major area of contribution:

RQ1: How can wearable sensing systems be designed to visually confirm health-risk behaviors and objectively capture health-risk behaviors in real-world settings?

RQ2: How can activity-oriented RGB and thermal data from HabitSense be fused using deep learning to address the multi-class problem?

RQ3: How do we best capture the wearer’s activity of interest while reducing both wearer and bystander privacy concerns?

RQ4: How can AI-driven predictions derived from passively sensed health-risk behaviors be made interpretable, transparent, and actionable for clinicians and end users?

The following subsections motivate each question in depth and outline the corresponding technical and empirical contributions.

1.2.1 Designing for Visual Confirmation and Objective Capture (RQ1)

Overcoming the limitations of self-report and recall-based methods requires objective, continuous tools that can capture health-related behaviors as they unfold in daily life. Wearable sensing systems provide such an opportunity, enabling unobtrusive, real-time capture of behavioral and contextual signals from daily activities such as eating and smoking [9], [10]. However, most traditional wearables rely primarily on inertial sensing (e.g., accelerometers or gyroscopes) and are well-suited for coarse activity recognition—such as walking, running, or step counting—but lack the specificity to identify fine-grained gestures and contextual cues associated with health-risk behaviors such as overeating and smoking [11]. Detecting eating and smoking episodes in real-world settings requires visual confirmation of hand-to-mouth activity and the surrounding context [12], [13]. Activity-oriented wearable cameras—devices whose field of view is explicitly oriented toward a region of interest, such as the mouth—offer a promising approach [7]. Yet introducing such cameras into everyday life raises practical and ethical challenges: devices must operate reliably over full days of wear, limit capture to brief behavior-triggered views of the mouth region to conserve battery and storage, and ensure that any visual data suitable for continuous capture can be processed in a privacy-preserving manner [6]. These considerations motivate the first research question:

RQ1: How can wearable sensing systems be designed to visually confirm health-risk behaviors and objectively capture health-risk behaviors in real-world settings?

To address RQ1, this dissertation introduces *HabitSense* [14], a lightweight, neck-worn, activity-oriented camera system that integrates RGB, thermal, and inertial sensing and orients its field of view toward the wearer’s mouth. HabitSense uses an activation pipeline that selectively records hand-to-mouth activity, enabling visual confirmation of eating and smoking episodes while maintaining all-day wearability and reducing unnecessary data capture. The design of HabitSense is informed by qualitative work with domain experts and focuses on balancing user comfort, clinical relevance, and data utility for downstream modeling.

1.2.2 Fusing RGB and Thermal Signals for Multi-Class Detection (RQ2)

Designing a wearable system capable of capturing relevant behaviors is only the first step. The next challenge is to use the captured data to accurately detect and distinguish behaviors of interest. In the case of eating and smoking, this is particularly challenging: both behaviors involve similar hand-to-mouth gestures and overlapping postures, and they occur in diverse environments with variable lighting, backgrounds, and occlusions.

Models trained on a single modality often struggle with these complexities. For example, RGB-only models may fail in low-light conditions or when the face is partially obscured, while thermal-only models may be confused by heat sources unrelated to the behavior [14]. Inertial signals, while useful for motion dynamics, cannot on their own capture object-level context or visual details. These limitations highlight the need for multimodal modeling strategies that combine complementary information across sen-

sors. This leads to the second research question:

RQ2: How can activity-oriented RGB and thermal data from HabitSense be fused using deep learning to address the multi-class problem?

Addressing RQ2, this dissertation develops *RTFVNet* [15], a mid-fusion deep learning architecture that jointly leverages activity-oriented RGB and thermal video to classify smoking, eating, and background activity. Trained on a free-living RGB–thermal dataset collected with HabitSense, RTFVNet achieves robust performance while reducing computational cost relative to dual-model decision fusion baselines, offering a practical path toward real-time multimodal behavior detection on wearable platforms.

1.2.3 Preserving Privacy (RQ3)

Even with accurate multimodal detection, widespread deployment of activity-oriented camera systems is unlikely unless users and bystanders trust that their privacy will be protected [14]. Camera-based sensing can capture sensitive information about faces, environments, and bystanders who have not consented to being recorded [6], [7]. These concerns are not merely technical; they shape perceptions of surveillance, social comfort, and fairness, and can directly affect willingness to adopt and wear such devices.

At the same time, the very visual information that raises privacy concerns often carries important contextual cues for behavioral interpretation. Naively eliminating context—for example, by aggressive blurring or masking—can undermine analytic utility and make it difficult for clinicians and researchers to interpret recorded episodes. This tension between analytic utility and perceived privacy motivates the third research question:

RQ3: How do we best capture the wearer’s activity of interest while reducing

both wearer and bystander privacy concerns?

To address RQ3, this dissertation investigates human-centered and generative approaches to visual obfuscation [16]. In particular, it develops cartoon-based generative methods that transform raw video into stylized representations that preserve gesture- and context-level information while masking identifiable facial and background details. Through a combination of online and field studies, these methods are evaluated for both perceived privacy and retained interpretability, with the goal of enabling socially acceptable capture of behavior-relevant visual evidence.

1.2.4 From Opaque Models to Clinical Insight (RQ4)

Finally, the utility of wearable sensing and predictive modeling depends on whether their outputs can be meaningfully integrated into clinical decision-making. Clinicians and other health professionals are often the end users of these systems: they must interpret model predictions, assess their reliability, and translate them into intervention strategies or treatment adaptations [17]. Yet many machine learning models operate as black boxes, providing predictions without clear explanations of the underlying reasoning [18].

In digital health, predictive models can flag patients who may be at risk of poor outcomes—such as non-response to a weight loss intervention—based on passively sensed behavioral patterns [19]. However, if clinicians cannot understand why a particular patient is flagged, they may be unable to connect predictions to actionable behavioral targets or may distrust the model entirely. This gap between algorithmic prediction and human understanding motivates the fourth research question:

RQ4: How can AI-driven predictions derived from passively sensed health-risk behaviors be made interpretable, transparent, and actionable for clinicians

and end users?

To address RQ4, this dissertation develops *PRIMO* [20], an interactive visualization tool that explains model predictions using prime-implicant-based summaries. *PRIMO* is designed to align with clinicians’ reasoning processes by highlighting combinations of behavioral features that are sufficient to support a prediction and allowing experts to explore alternative scenarios. Evaluations with clinical experts examine whether *PRIMO* increases agreement with model predictions, improves understanding of model behavior, and is preferred over common explanation methods such as feature importance rankings.

1.3 Integrating Sensing, Modeling, Privacy, and Explainability

Together, these four research questions define a cohesive agenda for unobtrusive health behavior sensing that spans device design, multimodal modeling, privacy, and interpretability. RQ1 focuses on activity-oriented wearable system design for capturing eating and smoking episodes in the wild. RQ2 extends this by developing deep learning methods that fuse RGB and thermal data from HabitSense to tackle a challenging multi-class detection problem. RQ3 addresses the privacy and social acceptability of camera-based capture, proposing generative visual obfuscation techniques that protect identities while preserving behavioral context. RQ4 advances interpretability and clinical translation through interactive explanation tools that make model predictions more transparent and actionable.

Collectively, these strands form a framework that links sensing, modeling, privacy, and explainability into an end-to-end pipeline: from capturing episodes of eating and

smoking in everyday life, to recognizing and disambiguating behaviors, to making captured and modeled information both privacy-aware and clinically interpretable.

1.4 Major Contributions

This dissertation makes four major contributions to the field of mobile and ubiquitous health sensing, each supported by empirical evidence from studies spanning more than 150 participants and hundreds of hours of multimodal data.

- **Design and deployment of HabitSense.** We developed *HabitSense*, a lightweight, neck-worn, activity-oriented camera system that integrates RGB, thermal, and inertial sensing to capture eating and smoking behaviors in naturalistic environments. A 15-participant, 768-hour field deployment demonstrated that our activity-triggered capture pipeline can substantially reduce stored video and sensing duty cycle while maintaining reliable capture of eating and smoking episodes. HabitSense orients its field of view toward the mouth and selectively records hand-to-mouth activity, enabling visual confirmation of eating and smoking episodes while limiting continuous recording and overall data volume.
- **RTFVNet for multimodal behavior recognition.** To address the limitations of single-modality sensing, we introduced *RTFVNet*, a mid-fusion deep learning architecture that jointly leverages activity-oriented RGB and thermal video to classify smoking, eating, and background activity. Trained on 441.9 hours of free-living RGB–thermal data from seven participants, RTFVNet achieved strong performance ($F1 = 85\%$) while requiring less computation than dual-model decision fusion baselines. This contribution demonstrates that carefully designed multimodal integration can improve efficiency and robustness for detecting complex, overlapping behaviors.

- **Generative privacy-preserving methods for wearable video.** To mitigate privacy risks in wearable vision systems, we developed and evaluated generative cartoonization techniques that transform raw video into obfuscated yet contextually meaningful representations. Across an online study ($N = 60$) and a field deployment ($N = 49$), these methods substantially reduced privacy concerns among both wearers and bystanders while maintaining the ability to visually confirm behavioral context in the majority of cases (approximately 90%). These results provide concrete evidence that privacy and analytic utility can be jointly optimized in the design of wearable camera systems.
- **Explainable AI for clinical insight (PRIMO).** Finally, we created *PRIMO*, an explainable AI interface that presents prime-implicant-based explanations of model predictions through interactive visualizations. In a study with 14 clinical experts in weight management, PRIMO increased agreement with model predictions and was rated the most helpful method for understanding model behavior compared to baseline explanation techniques. This contribution shows how carefully designed explanation interfaces can bridge the gap between complex predictive models and the needs of clinicians in real-world decision-making.

Chapter 2

HabitSense

2.1 Introduction and Background

Research Question (RQ1). *How can wearable sensing systems be designed to visually confirm health-risk behaviors and objectively capture health-risk behaviors in real-world settings?*

Guided by RQ1, this chapter synthesizes clinical input, privacy-preserving camera design, and smart activation to motivate and present *HabitSense*.

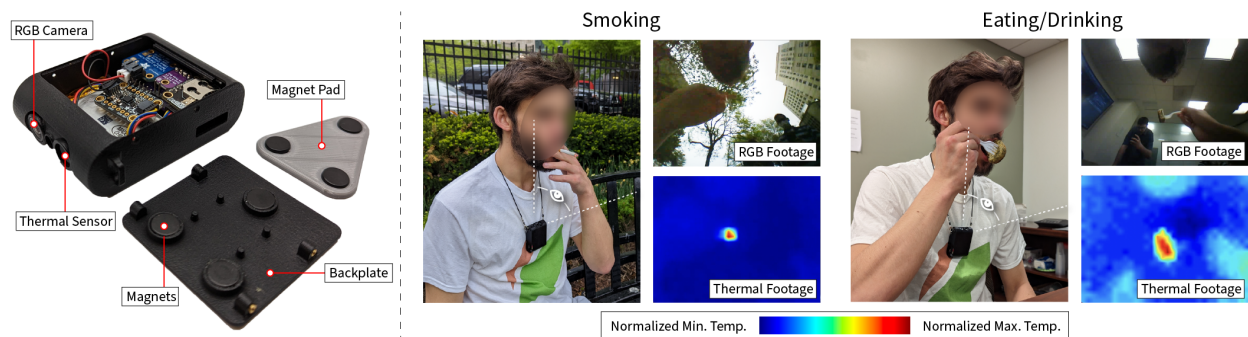


Figure 2.1: The HabitSense device (left) and its applications (right). HabitSense is a wearable multi-modal system (thermal, RGB, accelerometer) optimized for privacy, enabling all-day recording and on-device processing to detect health-risk behaviors associated with smoking and eating.

In the United States, obesity represents a significant public health concern, affecting approximately 40% of adults [21], [22]. Obesity is closely associated with an elevated risk of various chronic diseases, including heart disease, certain types of cancer, and other long-term health issues [21], [22]. Similarly, cigarette smoking remains the leading cause of preventable death [23], [24], [25]. Health-risk behaviors such as overeating, excessive food consumption beyond physiological needs, and tobacco smoking are major contributors to premature morbidity and mortality worldwide. [26], [27], [28]. Eliminating these behaviors can help prevent several chronic diseases. However, this requires understanding and modifying what individuals put into their mouths. Accurate and efficient detection of consumption patterns will improve our knowledge of these behaviors and facilitate the development of effective interventions to prevent them.

In current approaches to understanding the factors influencing the daily habits of individuals, clinicians often rely on self-report methods to document food consumption, timing, and psychological factors related to the behavior of interest. Although self-report methods effectively capture psychological factors, these methods may introduce bias, impose a significant burden on participants, and are frequently susceptible to errors from forgetfulness [29], [30], [31], [32]. Integrating automatic, objective measures would significantly enhance the accuracy in documenting the timing, patterns, and context surrounding consumption behaviors [33], [34]. The increasing popularity of wearable technology, such as smartwatches, offers the potential for automated behavior monitoring, but these devices lack essential visual confirmation capabilities for real-life applications. *Wearable cameras offer a promising solution by providing visual evidence that enhances contextual understanding and offers more reliable approaches to automated confirmation of health-risk behaviors.* Therefore, there is a need for objective, unobtrusive wearable camera systems that can enhance the diagnosis of health-risk behaviors, allowing for personalized treatments

and real-time triggers to test interventions that improve treatment outcomes [11], [35].

Wearable cameras developed by the ubiquitous computing community have shown promise in monitoring behaviors such as overeating and smoking. However, their adoption, particularly in clinical settings, faces significant challenges, primarily due to the lack of clinical involvement in the design process [36]. This oversight often results in systems that do not align with the specific needs and workflows of healthcare providers, preventing their integration into routine clinical practice [36]. Engaging clinicians early in the development process is essential to ensure that the physical design and capabilities of the system are practical and valuable in real-world healthcare settings. [37]. However, the role of clinician involvement in designing wearable cameras and qualitatively analyzing potential obstacles that might impede their acceptance and feasibility in healthcare settings has not been extensively investigated.

Based on feedback from our focus groups with clinicians (N=36) and experience from user studies involving wearable camera systems (N=105, across studies), privacy concerns continue to be a significant barrier to adopting wearable cameras. To address user privacy concerns, recent studies have explored the concept of activity-oriented cameras (AOCs) [38]. In contrast to traditional surveillance and egocentric cameras, AOCs are designed to focus only on capturing the intended activity of interest, inspired by the principle of least privilege (PoLP)[39], [40]. For example, one approach to capturing hand-to-mouth gestures involves orienting the lens of a neck-worn camera toward the mouth of the *wearer* (the person wearing the camera) and obfuscating background pixels to prevent capturing *bystanders* (individuals other than the wearer) or unintended behaviors in the video.[12]. Research has demonstrated a significant increase in system acceptability among both AOC wearers and bystanders while simultaneously maintaining the utility

of visually confirming hand-to-mouth behavior [41]. However, there currently exists no AOC-based wearable system capable of running such obfuscation algorithms on-device in real time, while also detecting health-risk behaviors.

Developing an AOC-based wearable camera system that can capture eating and smoking gestures while running obfuscation algorithms in real time presents two challenges. First, the device must operate continuously for a full, 16-hour day [13]. Second, the camera must be computationally capable of running real-time obfuscation algorithms with the least possible impact on battery life. Smart activation mechanisms can improve battery life by using a low-energy sensor to detect preliminary signs of specific behaviors, activating more power-intensive tasks only as needed [42]. Given that eating and smoking are intermittent activities, this approach will allow selective activation of obfuscated recordings only *when a sensor detects preliminary gestures of these activities*. This will not only conserve battery life but also enhance privacy by ensuring that only detected health-risk behaviors are recorded.

Given the need for a sensor that can detect preliminary gestures of interest to activate high-energy processes and considering the necessity to obfuscate the user's background for privacy, we decided to build a wearable camera that incorporates an infrared (IR) thermal sensor alongside RGB. The choice of an IR thermal sensor is strategic because it allows the detection of the wearer and any objects they might handle by exploiting temperature differentials. Because the wearer is closest to the camera lens, their body temperature will create a distinct thermal signature, effectively distinguishing them from the background, enabling methods that distinguish the wearer and foreground from the background [12].

Our goal was to synthesize diverse research strands surrounding wearable cameras

into a comprehensive whole, providing clinicians and users with a privacy-preserving, unobtrusive wearable camera platform capable of providing visual evidence and detecting health-risk behaviors that users are willing to wear throughout the day. In doing so, HabitSense can assist clinicians in improving treatment efficacy by identifying the contextual factors that drive these behaviors when they occur to aid in personalizing treatment strategies.

To answer RQ1, we (i) align with clinical workflows, (ii) reduce privacy risk via activity-oriented capture and on-device obfuscation, and (iii) ensure data utility through smart activation and robust modeling. Our contributions are:

1. *To increase clinical adoption and inform design*, we report on insights from a qualitative analysis of focus group discussions involving 36 weight management and smoking treatment specialists to inform the need, design, and functionality of wearable cameras.
2. *To increase user adoption*, we leveraged our extensive iterative experience (N=105, in all studies) in wearable camera design to develop HabitSense (illustrated in Figure 2.1), an RGB thermal (RGB-T) wearable camera that is designed to be easy to use, unobtrusive, reproducible, lightweight, and capable of operating all day (16 hours). HabitSense is also designed to detect eating and smoking behaviors while incorporating on-device obfuscation to enhance privacy.
3. *To enhance user privacy and reduce burden*, we developed the Sensor-Enabled Control of Ubiquitous Recording and Evaluation (SECURE) algorithm. This algorithm efficiently activates obfuscated RGB video recording, successfully capturing 90% of hand-to-mouth gestures associated with eating and smoking while turning on the

camera only 52.4% of the time. As a result, it achieved a 48% reduction in data storage and a 30% reduction in battery usage, significantly optimizing device efficiency.

4. *To evaluate our device*, we tested HabitSense in a natural setting with 15 participants (8 with obesity, 7 smokers), yielding 768 hours of footage, and reported on the high user acceptability of the HabitSense device.
5. *To validate the utility of our device*, we trained machine learning models on RGB, thermal, and obfuscated video data to run offline and in real time. We then rigorously evaluated their ability to recognize eating, smoking, and hand-to-mouth gestures and reported on the performance of our models. Our comprehensive assessments demonstrated a 92% F1-score in hand-to-mouth gesture recognition, highlighting the effectiveness and versatility of our models. Additionally, our eating and smoking detection models proved to be robust across various conditions, including night time, low-light settings, and during intense head movements.

2.2 Related Work

2.2.1 Current Wearables for Detection of Eating and Smoking Activities

This section highlights a few of the existing wearables for eating and smoking detection and their limitations.

2.2.1.1 IMU-based sensing

Wearable technologies like smartwatches, commonly equipped with inertial measurement units (IMUs), have gained popularity for automatically detecting eating and smoking activities. This functionality helps overcome certain limitations inherent in self-reporting

Work	Device Name	Popular	IMU	RGB Camera	Form Factor	Privacy	Participants	Meals	Smoking Events	Free-living	In-lab vs	Free-living Days	Clinician Study	Hours of Data	Detection Targets
Eating	Bedri et al. 2017	EarBit	●			👤	10	16		free-living	NR	45			chewing sounds, chewing motion
	Bi et al. 2018	Auracle				👤	14	26		both	NR	2+32			chewing sounds
	Bedri et al. 2020	FitByte	●		👤	👤	23	69		both	NR	91			chewing and swallow motion, h2m gestures
	Morshed et al. 2020		●			👤	28	1,305		free-living	21	NR			wrist motion
	Nyamukuru et al. 2020	Tiny Eats				👤	20	20		in-lab		11			chewing sounds
	Zhang et al. 2020	NeckSense	●			👤	20	40		free-living	14,2	470			chewing motion, forward-lean angle
	Bedri et al. 2022	FitNibble	●			👤	12	NR		free-living	18	NR			h2m gestures, chewing motion
	Bi et al. 2022				👤	👤	10	NR		free-living	NR	55			h2m gestures
	Morshed et al. 2022		●			👤	18	54		in-lab		11.5			chewing motion
	Shin et al. 2022	MyDJ	●			👤	24	94		free-living	7	237			chewing motion
Smoking	Sazonov et al. 2013	PACT				👤	20		40	in-lab		20			breathing motion, h2m gestures
	Lopez-Meyer et al. 2015	PACT				👤	20		40	in-lab		20			breathing motion, h2m gestures
	Echebarria et al. 2017					👤	2	6		in-lab		NR			breathing sounds
	Añazco et al. 2018		●			👤	4	4		in-lab		NR			h2m gestures
	Senyurek et al. 2019		●			👤	35	443		both	1	821			h2m gestures
	Senyurek et al. 2019	PACT2.0	●			👤	35	443		both	1	871			h2m gestures
	Tai et al. 2020					👤	NR	NR		in-lab		NR			nicotine biomarkers
	Imtiaz et al. 2020				👤	👤	10	50		free-living	1	108			h2m gestures
	Alharbi et al. 2022	SmokeMon	●			👤	19	115		free-living	NR	110			heat signature
Privacy	Chattopadhyay & Boulton 2013					👤									
	Wrinkler & Rinner 2010					👤									
	Wrinkler et al. 2014					👤									
	Fernandes et al. 2024	HabitSense	●	●	●	👤	15	522	283	free-living	7	✓	768		h2m gestures, heat signature

Key
NR
Not Reported
Smartwatch
Earbuds
Necklace
Microphone
Eyeglasses
Hat
Vest
Nasal Cannula
Sweatband

Table 2.1: Comparison of HabitSense and existing systems. This table highlights the distinctions between HabitSense and other systems in terms of sensor modalities, form factor, privacy protection, and the nature of studies conducted

methods. IMU-based approaches to eating detection largely focus on extracting hand-to-mouth gestures from motion data, typically collected at the wrist [11]. For instance, Morshed et al. [43] achieved real-time eating detection at the meal level by training a classifier to recognize eating gestures using smartwatch accelerometer data. Other on-body IMU sensor locations have also been explored. One example is the earbud accelerometer in the study by Morshed et al. [44], designed to detect chewing motions. These earbud sensors are less influenced by confounding hand-to-mouth gestures typical of wrist-worn IMUs, enabling more precise differentiation between eating activities such as meal consumption and snacking.

IMUs have also supported multimodal systems where non-IMU sensors primarily detect eating activities, such as in a study by Zhang et al. [45]. Their necklace device employs proximity and ambient light sensors aimed at the jaw to measure chews directly,

while the IMU supplements by estimating the lean-forward angle, enhancing the prediction model. Bedri et al. [46] embedded an IMU and proximity sensor into eyeglasses. This design enables detecting eating gestures through hand-to-mouth proximity and captures chews from IMU-registered movements of the facial muscles. Similarly, recent developments feature the MyDJ device [47], which integrates both IMU and piezoelectric sensors onto eyeglasses for eating detection.

IMU-based systems have also been explored for smoking detection. Añazco et al. [48] demonstrated the efficacy of wrist-worn IMUs in detecting smoking-related hand gestures, successfully differentiating them from confounding activities such as eating and drinking. However, due to variations in gestures across individuals, their method lacks generalizability. To address this, Senyurek et al. [49] focused on analyzing gesture regularity from wrist-based IMUs, although performance significantly decreased in free-living settings. Further attempts to enhance performance led Senyurek et al. [50] to integrate a smart lighter with a wrist-worn three-axis accelerometer. Still, inconsistencies arose due to self-report errors and instances of participants using their lighters instead of the provided smart lighter.

To summarize, although wrist-worn IMU sensing modalities are recognized for their ability to detect eating and smoking activities, they face challenges in practical application. These include false positives from unrelated movements, issues with sensor positioning (i.e., dependency on the dominant hand), and difficulties in generalizing across different individuals [11]. However, IMU-based sensors continue to supplement multi-modal devices [51].

2.2.1.2 Acoustic-based sensing

Another area explored uses acoustics to capture sonic features of chews, swallows, and jaw movements. Bi et al. [52] found that chewing sounds could be retrieved from audio collected with a contact microphone on the mastoid bone. Nyamukuru et al. [53] improved on this approach by implementing a shallow gated recurrent unit neural network optimized for execution on a low-power microcontroller, detecting mastoid chewing sound with greater accuracy and efficiency. Acoustic sensors have been employed in multi-modal systems, such as in a study by Bedri et al. [54] where a neck-positioned microphone detects swallowing sounds. Additionally, an earbud equipped with an IMU and a proximity sensor was used to identify chewing. Acoustic sensing modalities have also been explored to detect smoking activities. Echevarria et al. [55] demonstrated the potential of wearable acoustic sensors, designed for respiratory monitoring, to identify smoking behavior by capturing the distinct acoustic properties of smoking inhalation. However, microphones are susceptible to noise interference, particularly in free-living environments. Moreover, using microphones raises significant privacy concerns in everyday settings [38].

2.2.1.3 Alternate sensing modalities

Several studies have explored alternate sensing modalities and methodologies to monitor cigarette smoking behavior precisely. One approach involved the development of the PACT system [56], which integrated a comprehensive sensor suite consisting of a respiratory inductance plethysmograph and hand gesture sensor capable of capturing multiple behavioral indicators associated with cigarette smoking. Using support vector machines trained on data from the PACT system, Lopez-Meyer et al. [57] highlighted the feasibility of autonomously identifying smoke inhalations by continuously monitoring respira-

tory patterns and hand-to-mouth gestures. Although this methodology holds promise for delivering precise and objective evaluations of smoking behaviors, the data collection and assessment of the PACT system were conducted within a controlled environment, limiting generalization to free-living contexts. Studies have also explored the feasibility of using wearable sweatbands to monitor nicotine levels in the sweat, thereby inferring smoking behavior [58]. Although the prospects appear promising, the frequent need for calibration, coupled with potential contamination from passive smokers' smoke deposition, could compromise the accuracy of readings. Battalio et al. conducted a micro-randomized trial using wrist-worn and ECG sensors, intending to refine a just-in-time-adaptive stress management intervention to mitigate smoking relapse [59]. This research emphasized the potential of wearable sensors in facilitating real-time data acquisition, thereby supporting personalized interventions.

Although current methods for automated detection of eating and smoking are promising, deploying these methods in free-living settings often relies on self-reporting or external tools, which can introduce inaccuracies and biases into the ground truth data. Additionally, there is no visual evidence to confirm the occurrence of the events.

2.2.2 The Need for Wearable Camera Systems for Detecting Eating and Smoking

To circumvent problems of bias in self-reporting, researchers have turned to wearable camera devices that continuously record activities, providing rich contextual information and reliable visual confirmation for ground truth data. The introduction of SenseCam marked a notable stride in this direction—a wearable digital camera capable of capturing a comprehensive record of the wearer's daily activities through still images and

sensor data [60]. Building on this concept, Bedri et al. [46] integrated a camera into an eyeglasses-based system that used IMUs and proximity sensors to detect eating behaviors. Similarly, head-mounted video cameras [61] have been developed specifically for video-based eating detection. In both cases, the addition of cameras provides visual ground truth from the wearer’s perspective, enhancing data reliability. Wearable camera systems in the form of eyeglasses have also been used for smoking detection in free-living settings [62]. Such eyeglasses-based systems, outfitted with multi-modal sensing capabilities, excel in gesture recognition for eating and smoking detection tasks. However, the eyeglass form factor impacts its generalizability. Privacy concerns, practical issues such as hair obstructing the camera, using a single eyeglass frame type, and discomfort or reluctance to wear glasses among some individuals further impede adoption [62]. Addressing some of these issues, recent research introduced a thermal camera in a necklace form factor for smoking detection [13]. The authors conducted a free-living study with the device, collecting thermal data as participants self-reported their smoking activities. However, the reliance on potentially inconsistent and inaccurate self-reported logs posed a challenge due to the lack of visual evidence for confirmation. Building on earlier discussions of multi-modal systems with IMUs and camera-based systems using RGB and thermal technologies, there is potential for developing a multi-modal platform that integrates RGB, thermal imaging, and IMU sensors.

2.2.3 The Need for Clinical Feedback in Informing the Design of Wearable Systems

Wearable technology is increasingly recognized as a vital component in healthcare, offering the potential to passively monitor health, assist in diagnosis, and enhance patient care [63], [64], [65]. However, several challenges impede the adoption of wearable tech-

nology in clinical practice, such as device accuracy, privacy concerns, and cost, as well as limited clinical involvement in design and technical and interoperability issues [63], [66], [67], [68], [69], [70], [71]. Integrating wearable devices into medical practice necessitates clinician input to ensure device effectiveness and relevance in clinical settings. Qualitative analysis with clinicians has shown potential in integrating clinician feedback to improve mobile health (mHealth) technology and increase adoption [20], [36], [72].

2.2.4 Sensing Multiple Health Behaviors with Privacy-Preserving Wearable Cameras

Using AOCs along with obfuscation algorithms to obfuscate backgrounds—including bystanders and sensitive activities—has proven effective in mitigating privacy concerns [38], [41]. Surveys [41] have examined the balance between privacy concerns and data utility using various obfuscation techniques, including blurring, applying a canny edge algorithm, entirely blacking out the background, or using Generative Adversarial Network (GAN) -based cartoon obfuscation [16] in activity-oriented videos captured using neck-worn cameras. User studies have also demonstrated that these techniques effectively reduce privacy concerns [16], [41]. However, implementing these algorithms in real time on resource-constrained devices is yet to be tested and presents a challenge.

Several camera systems employ on-device obfuscation algorithms to improve privacy [73], [74], [75]. However, these systems are not designed as wearable cameras that can be worn all day on a single charge, and they lack comprehensive user studies. Recent research has highlighted the potential of thermal cameras for obfuscation [12]. Thermal data can isolate the wearer’s foreground activities from the background. Although this approach is promising, it was tested only in offline scenarios. Additionally, it lacks a

comprehensive assessment of how well obfuscation and the use of thermal-only versus RGB-only data impact detecting activities such as hand-to-mouth, smoking, and eating gestures. Thus, further extensive investigations are necessary to refine thermal camera efficiency for real-time, on-device applications in real-world settings. This collective evidence suggests significant potential for developing an AOC RGB-T wearable camera system that uses the thermal sensor for obfuscation over the need for complex computer vision or deep learning algorithms on the device.

Based on these motivations, the novelty of HabitSense stems from integrating four key aspects: (1) incorporation of feedback from clinician focus groups and user studies into its development, (2) on-device obfuscation capabilities, (3) intelligent activation of obfuscated recording to enhance privacy and conserve battery life, and (4) a comprehensive evaluation framework that assesses the system’s performance across a variety of real-world scenarios. Table 2.1 compares our innovative features to existing methods for detecting health-risk behaviors associated with eating and smoking.

2.3 Clinician-Informed Design: Integrating Focus Group Feedback into Wearable Camera Development

To learn about the importance and feasibility of using privacy-conscious RGB-T wearables for monitoring various health-risk behaviors, we engaged in focused discussions with healthcare professionals experienced in managing eating and smoking habits. We recruited clinical dietitians (N=18) and tobacco treatment specialists (N=18) to ascertain problems faced in their practices, to inform design priorities from their experiences, to assess their willingness to use and adopt our proposed wearable camera system, and to identify practical challenges in implementing the system into treatment programs. Each

session included either dietitians or tobacco treatment specialists; none included participants from both groups. All participants worked in the Chicago area and had at least two years of professional experience. The focus groups took place during the early stages of the development of the HabitSense system.

In the focus group sessions, the moderator used a topic guide to ask questions targeting four user-centered design elements: **Motivation:** What measurement tools do the clinicians have, and what are their limitations?; **Desirables:** What additional measures do clinicians want to improve their understanding of their patients' behaviors?; **Advantages:** How could a passive-sensing visual/thermal wearable with real-time activity recognition benefit clinical practice?; **Barriers:** What challenges do practitioners foresee in utilizing a system like ours? To avoid biasing responses to questions about the clinicians' experiences with contemporary measures, no information about the proposed system was given until after addressing the Desirables portion of the focus group. The proposed system shown to participants was an early prototype that incorporated design lessons learned across Gen 1 and Gen 2 deployments. In addition to planned questions, the moderator asked follow-up questions and encouraged participants to elaborate on their answers and comment on other participants' answers.

We structured our analysis to reflect the role of our participants as administrators, rather than end-users, of the proposed system and to account for the value of their domain expertise in our design process. First, we separated focus group transcripts by topic and labeled the participants. Two authors then independently analyzed and thematically coded the transcripts. Once independent code lists were generated, the authors met to compare lists, resolve discrepancies, and produce and iterate upon a common set of themes. All unique responses that were "seconded" (i.e., repeated or affirmed) by another

participant in the same session were included as themes. Additional themes were derived inductively and proposed by the authors in keeping with standard thematic analysis practices [76]. Proposed themes that achieved consensus between authors were added to the final list of themes, detailed in Table 2.2.

Unsurprisingly, there was a high similarity in how dietitians and tobacco treatment specialists answered the moderator's questions. Both types of practice hinge on providers' understanding of habitual health-risk behaviors for any given patient. Both providers rely on similar measurement tools and, though measuring different behaviors, experience similar measurement limitations.

Motivation (what do they have?)	Themes	Design Considerations
Existing measures	Self-report Memory recall	- -
Existing measures limitations	<i>self-report</i> : Inadherence <i>self-report</i> : Inaccuracy/dishonesty <i>self-report</i> : Dietary knowledge (<i>dietitians only</i>) <i>memory recall</i> : Forgetting/dishonesty	Low burden Real-time detection RGB, All-day wear, RGB buffer* Real-time detection, All-day wear
Desirables (what do they want?)		
	Behavior Context Other substance-use (<i>tobacco TS only</i>) Food budget/Income (<i>dietitians only</i>)	RGB, FOV, Orientation, Resolution RGB, Thermal, FOV, Orientation Cost
Advantages (how could wearables help?)		
	Real-time treatment monitoring Behavior trigger identification Improved reporting accuracy Fine-grain features of behavior Food temperature (<i>dietitians only</i>)	Real-time detection RGB, RGB buffer* Real-time detection, All-day wear RGB, Thermal, Model robustness Thermal
Barriers (what challenges do they predict?)		
	Detection failure cases Dishonesty Operational conditions Social stigma Privacy Cost (<i>dietitians only</i>)	Model robustness - Waterproofing*, Drop-safety Form-factor, Acceptability FOV, Obfuscation Cost

Table 2.2: Themes and associated design considerations. Themes are categorized based on those identified in focus group sessions and are organized by participant group and linked to relevant design considerations. *Retained for future work.

2.3.1 Motivation: Existing Measures and Their Limitations

Treatment providers in both groups reported that eating and smoking are traditionally measured by self-report, in which patients either estimate their dietary/smoking activities over a given time range or record them in real time via a paper journal or software application. Providers further confirmed that real-time self-reporting is burdensome and thus achieves limited adherence. Several participants from both groups indicated difficulty obtaining patient self-reports that are timely, accurate, and consistent over time. The participants reported that comprehensive calorie or tobacco intake accounting requires too much work for the typical patient. Additionally, dietitians reported that the quality of self-reported eating data correlates to the dietary knowledge of the patient. For instance, patients with low dietary knowledge frequently neglect to report sauces and their ingredients because they are unaware of their importance as food items. In both groups, some providers reported encouraging patients to use mobile applications for self-reporting. These applications can modestly improve reporting accuracy for some patients but ultimately suffer from the same adherence problems stemming from the burden involved in any form of continuous self-reporting.

Providers reported that memory recall approaches (where a trained specialist guides memory exercises to prompt maximum recollection of events) to smoking and dietary monitoring address the burden problem by combining the reporting of multiple smoking or eating events into a single recall event. However, both groups of providers reported that memory recall suffers its own limitations, namely forgetting and dishonesty, both of which variably bias patient data.

2.3.2 Desirables: Desired Information About Patients

Because we are designing new modalities of behavior measurement, we asked participants not just what problems they experience but also what information they would want about their patients in a hypothetical "perfect world" in which providers can magically obtain any data they want about their patients. Across both participant groups, responses converged around two key points: (1) context: environmental or social factors surrounding a given smoking or eating event (e.g., location, presence/absence of others, presence/absence of screens); and (2) secondary activities: behaviors that co-occur with smoking or eating (e.g., reading, talking on the phone, drinking coffee). With reliable detection of eating or smoking, our device could trigger timely smartwatch or smartphone notifications to obtain context or secondary activities.

2.3.3 Advantages: Benefits of Passive Sensing

After explaining the premise of the HabitSense system to participant groups, we sought their input on the benefits of an RGB-T passive-sensing system in contrast to the tools currently used in their practice. In linking the system's design to the constraints of existing measurement methods, both groups agreed that the proposed system could enhance the precision of dietary and smoking data accessible to healthcare professionals.

In both participant groups, the automatic delivery of real-time behavioral data to clinicians was viewed as a strong benefit. Specifically, one clinician stated that *"In therapies ... we could check in during the day-to-day through the device to actually know about the improvement or if they are lagging behind."* Another clinician expanded on this perspective, emphasizing that *"... the disparity of information we are trying to close can be achieved if we get live statements and summaries from this sensing device."* Beyond monitoring treatment

efficacy, clinicians reported great interest in using data reported by the system to infer behavioral triggers from the context and/or secondary activities associated with eating or smoking episodes. An example provided by a tobacco treatment specialist illustrates this point: *"If the day after tomorrow this person took 15 cigarettes, and maybe today this person took 9, we can say what happened in that particular day.... Maybe this person was engaging in very tedious activity, and that craving was a little bit higher."* Another tobacco specialist observed that *"Using the video, you'd be able to see if the individual is in a social environment, or among friends... is it peer pressure? Or is it such kinds of things where the individual is being influenced in a social environment?"*

This indicates that these providers are not solely concerned with the observable behaviors but also with the underlying causes, which are more readily determined when rich contextual information is present before the event takes place. One dietitian stated, *"... the device would be really helpful because we would be able to get accurate data on what this individual usually does in their normal life, in their normal environment, so that we'll be able to understand specifically the activities which usually surround their behavior."* In the interest of identifying these patterns, both groups saw value in fine-grained behavioral measurements. Dietitians expressed interest in episode-level metrics, including meal duration, number of bites, and number of chews. Likewise, tobacco treatment specialists expressed interest in smoking topography metrics such as smoking episode duration, number of puffs, puff duration, inter-puff interval, and puff volume. Dietitians also expressed interest in food temperature as an indicator of food type and preparation methods.

2.3.4 Barriers: Challenges in System Implementation

We prompted participants to predict the challenges of implementing the HabitSense system with real patients. Both groups shared concerns regarding potential instances in which the system fails to detect a smoking/eating episode (false negative) or erroneously detects an episode when none occurred (false positive). Similarly, participants raised concerns over operational conditions the de-

vice may encounter in practical use, potentially impacting its effectiveness, such as water exposure, physical damage, or unusual environments. Participants speculated that issues related to privacy and social stigma might affect the willingness of patients to use the device, citing concerns over the collection of sensitive data and how the device's appearance might be perceived by individuals with whom the wearer interacts. Participants highlighted that dishonesty remains relevant because patients who aim to conceal their eating or smoking behaviors from healthcare providers could still do so if they know how the system operates. Lastly, dietitians reiterated that cost is a major factor patients and providers will consider when determining whether to adopt the device.

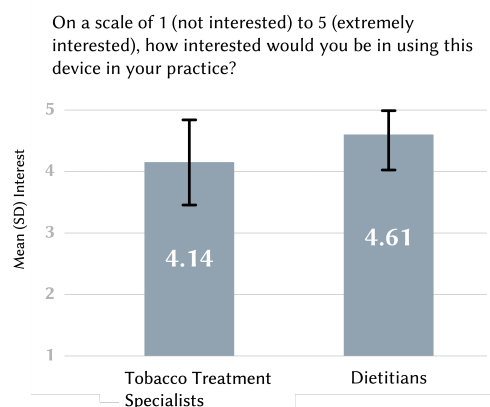


Figure 2.2: Mean reported interest in using the HabitSense system by participant group.

2.3.5 Treatment Provider Interest in Using the HabitSense Wearable Camera

At the end of each session, all participants were asked to rate their interest in using the proposed HabitSense system in their practice on a scale of 1 to 5, with 1 indicating "not interested" and 5 indicating "extremely interested." A two-sample t-test was performed

to compare interest in using HabitSense in clinical dietitians and treatment providers. There was no significant difference in interest, $t(34)=1.67$, $p=0.11$, despite 18 clinical dietitians reporting higher levels of interest ($M=4.61$, $SD=0.60$), compared with the 18 tobacco treatment specialists ($M=4.14$, $SD=0.69$). Certain participants specified that their rating was contingent on the HabitSense system "working as advertised" (i.e., performing at the levels of accuracy, generalizability, reliability, and efficiency stipulated in the design). These findings suggest a significant level of interest among both groups in wearable intake-monitoring systems and reflect their high expectations regarding the design and performance of these systems. The results of this focus group question are shown in Figure 2.2.

Although the difference was not significant, several factors may account for the slightly lower interest in the HabitSense device among tobacco treatment specialists. More dietitians reported using mobile applications and related technologies to support diet self-tracking than tobacco treatment specialists, which may contribute to dietitians having a slightly higher interest in adopting new technologies. Additionally, tobacco treatment specialists cited slightly greater concern over conditions interfering with the device's ability to capture contextual information since cigarettes are smoked in less favorable conditions (outdoors, dark, in motion) compared with eating (well lit, indoors, stationary plate). Lastly, a picture of a meal may provide slightly greater informational value than an image of a lit cigarette.

2.4 Design Implications from Experiential Insights in Wearable Cameras

Drawing on insights from dietitians and tobacco treatment specialists, we recognized the key design considerations outlined in the previous section as including privacy, RGB-T capabilities, sufficient resolution and field of view, affordability, minimal burden for users, and real-time detection capabilities during all-day wear. Refining our design in response to feedback and identified considerations, in addition to clinician input, we also uncovered hardware and firmware challenges through free-living studies with our evolving devices (see Supplementary Section 9.2 for Gen 1 and Gen2 study protocols). This experiential knowledge also informed the design and development of HabitSense. The subsequent section details our progressive evolution from Gen 1 to the culmination of HabitSense (see Figure 2.3 and Table 2.3).

2.4.1 Iterative Design

2.4.1.1 Gen 1

To overcome challenges encountered in neck-worn [45], [77], [78] and wrist-worn devices [79] designed to detect hand-to-mouth gestures related to eating and smoking activities, we set out to develop wearable cameras as a viable solution. We first developed Gen 1, which integrated a thermal sensor (GridEye 8x8 IR array) and RGB camera (OV2640) onto a printed circuit board (PCB). The addition of thermal imaging as a secondary sensing modality complemented the RGB data, particularly in detecting the wearer through the heat signature of pixels in the foreground. Privacy concerns were addressed by orienting the sensors toward the user’s face and upper torso. Through Gen 1, we improved data

collection while reducing user discomfort and privacy concerns.

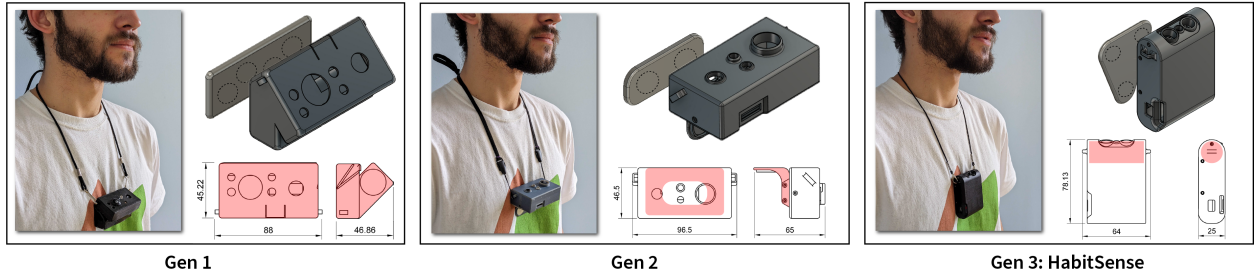


Figure 2.3: Iterations through different generations (devices and magnetic back plates). Showcasing the transition from Gen1 to Gen2 to HabitSense, this figure reflects the design process of HabitSense, guided by expert insights and user feedback, emphasizing key considerations such as privacy, affordability, and real-time detection. Pink regions indicate parts used for orienting the sensor lenses with the activity of interest. All dimensions are in mm.

Table 2.3: Specifications of different generations of devices.

Device	Sensors	Weight (g)	Volume (cm ³)	Battery life (h)	Connectivity	Applications
Gen 1	Grid-EYE, OV2640	106	102.18	12	BLE	Eating Detection
Gen 2	MLX9640, OV2640	132	114.75	14	BLE	Eating Detection
Gen 3	LIS3DH, MLX9640, OV2640	87	93.75	17.6	BLE, WiFi	Eating/Smoking Detection and EE Estimation

BLE: Bluetooth Low Energy, EE: Energy Expenditure.

Challenges and lessons learned: Although the device was successfully deployed in free-living settings (see Supplementary Section 9.2.1 for study protocol), the research team encountered three challenges during the study: (1) the device being warm to touch while charging, (2) time synchronization-related inaccuracies, and (3) limited field of view of the GridEye sensor. We addressed these challenges in the second iteration of our device.

2.4.1.2 Gen 2

To mitigate the problem of elevated device temperature during charging, we reduced the charging current of the onboard Li-Po charger in Gen 2, trading increased charging time for enhanced user comfort. To remedy time synchronization inaccuracies, we embedded a real-time clock (RTC) within the device powered by the onboard Li-Po battery. Concur-

rently, to prevent time loss due to battery depletion, we refined the device's firmware to include a sleep mode feature; this mode deactivates power-intensive functions when the battery falls below a set threshold, thus prioritizing RTC power supply and prolonging the device's operational longevity up to 20 days in sleep mode. To address the constraints posed by the limited 90° field of view and 8x8 pixel resolution of the GridEye sensor, we transitioned to the MLX90640 thermal camera, which offered a wider 110° field of view and a higher 32x24 pixel infrared array resolution.

Challenges and lessons learned: In the post-COVID landscape, the research and design team faced significant assembly challenges due to a marked shortage of STM chips and other key components. Despite an improved time-keeping mechanism extending battery life, time-drift issues led to unforeseen synchronization challenges, prompting repeated manual time corrections and reprogramming efforts. The device's non-modular architecture, based on a fixed PCB layout, posed challenges to incorporating new sensors or computational modules, such as machine learning accelerators, which would require a PCB redesign specific to the chosen accelerator. Feedback from participants again highlighted the need for a device that was smaller, lighter, and more ergonomically designed for comfort; that is, despite improvements from the first to the second generation, the device's form factor still did not achieve the level of unobtrusiveness necessary. Finally, an unfavorable center of gravity (see Fig. 2.3) and an inadequately designed magnetic pad failed to maintain the device's position consistently, leading to frequent displacements and consequently shifts in lens alignment that affected the integrity of the data collected.

2.4.2 Design Implications

Drawing on insights from focus group discussions and our iterative design experience in Gen 1 and 2, we delineate the guiding principles and objectives that informed the design

process of the third-generation HabitSense device.

2.4.2.1 Prioritizing user privacy through on-device obfuscation algorithms

Due to burden and biased data, our thematic analysis and participant feedback from wearable camera studies indicate that health professionals and users favor camera usage over traditional self-report methods; however, they caution against the potential invasiveness and extensive field of view of cameras that may capture more than intended. Researchers have addressed privacy concerns in wearable camera technology by proposing two key privacy-enhancing approaches: first, orienting cameras toward the activity of interest to ensure focused capture [38] and second, applying obfuscation algorithms to protect the confidentiality and integrity of the data collected [41]. Both the Gen 1 and Gen 2 cameras were designed to orient toward the activity of interest; however, they lacked the capability to execute obfuscation algorithms on-device due to the PCB's configuration. Executing obfuscation algorithms offline exposes sensitive information to risks that could jeopardize user privacy and the integrity of the data. Section 5.5 details the design and execution of our proposed on-device obfuscation algorithm, which serves as an effective measure to protect user privacy by curtailing the exposure of sensitive data. Developing on-device obfuscation algorithms necessitates preserving the data's context to ensure that health professionals can accurately interpret user activities, thereby balancing user privacy with the practical value of the data for behavioral analysis.

2.4.2.2 Acquiring minimal data for maximal utility in activity detection

The principle of least privilege (PoLP) represents a foundational concept in secure design, advocating for minimal privilege allocation to programs and users to fulfill their roles [40], [80]. This principle guides our approach to data acquisition in wearable camera research. Continuous video recording in free-living studies generates vast, oftentimes

irrelevant data, burdens storage media, and depletes device battery. Moreover, acquiring unneeded data, even if obfuscated, raises significant security concerns. We introduce the SECURE algorithm in Section 5.3 to navigate these issues. This algorithm strategically activates sensors only upon detecting a high likelihood of the targeted behaviors, thus limiting the data recorded to that which are pivotal for researchers and health professionals.

2.4.2.3 Modularity, reliability, reproducibility: pillars of system integrity and sustainability

The progression from our Gen 1 to Gen 2 devices highlighted a pivotal lesson: addressing device malfunctions due to field damage or component failures on the PCB required specialized technical proficiency for intricate repair work. Such procedures are not only time-intensive but also impose significant costs if entire PCBs are to be replaced. Modular designs not only facilitate expedited repairs and reduce costs but also bolster the device's dependability and consistency in production. The capacity for swift component testing and validation is crucial for the authenticity of data in free-living studies. Our system's modularity, supported by I2C communication and Qwiic connectors, also promotes expandability and seamless sensor augmentation, thereby extending its utility to monitor various health-risk behaviors. For instance, the platform could be reconfigured to assess UV exposure or outdoor time for melanoma survivors, creating a versatile tool for various health applications [81], [82], [83], [84]. Furthermore, in light of the recent supply chain constraints on electronic components post-COVID, the capacity to interchange parts highlights the critical importance of a modular system design. This adaptability not only ensures uninterrupted research continuity but also contributes to the sustainability and recyclability of the device, as components can be replaced or upgraded without the need for discarding the entire unit, reducing electronic waste and fostering a more

sustainable life cycle for research tools.

2.5 Methods: Design and Development of the HabitSense System

This section explores the concrete application of our design implications in developing the novel HabitSense system. It includes the rationale behind component selection and the iterative enclosure design process, resulting in a comprehensive end-to-end system. Further, we detail our methodology for the free-living study and describe our proposed SECURE algorithm to address PoLP, as well as our proposed real-time on-device obfuscation algorithms to protect privacy. The section concludes with details about our proposed gesture-detection framework.

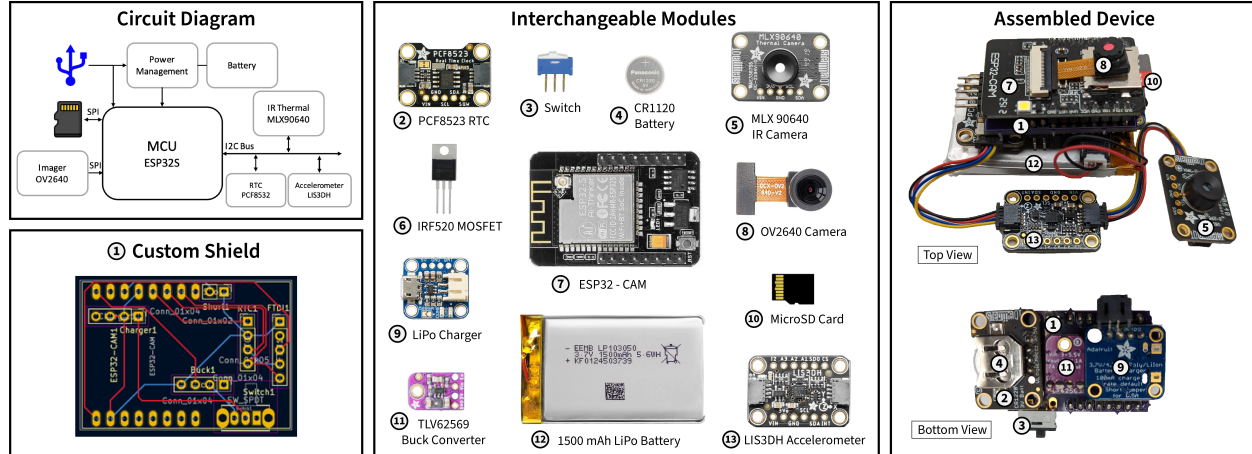


Figure 2.4: Internal design schematic and electronic components used in HabitSense device.

Table 2.4: HabitSense hardware design space. Components used in HabitSense are marked by *.

Modules	Alternatives	Features
Camera (RGB)	OV7670	VGA (640 x 480), 60mW/15fps(YUV)/maxRes
	OV2640*	2 megapixel (1600 x 1200), 140mW/15fps(JPEG)/maxRes
	OV5640	5 megapixel (2592 x 1944), 294mW/15fps(JPEG)/maxRes
Thermal Camera	Adafruit AMG8833	8x8 IR Array with 60° FOV, 0°C to 80°C range, 4.5 mA
	Grove MLX90641	16x12 IR Array with 110° FOV, -40°C to 300°C range, 12 mA
	Adafruit MLX90640*	24x32 IR Array with 110° FOV, -40°C to 300°C range, 23 mA
IMU	Adafruit LIS3DH*	3-axis IMU, inexpensive, 2 uA
	Adafruit LSM6DSOX	6-axis IMU, built-in gesture recognition, 550 uA
	Adafruit BNO055	9-axis IMU, ARM Cortex-M0, 13.7 mA
RTC	Adafruit PCF8523*	Dedicated battery, inexpensive, moderately precise
	Adafruit DS3231	Dedicated battery, high-precision
	SparkFun RV-1805	Rechargeable supercapacitor, high-precision
Voltage Regulator	Adafruit LM3671	600 mA, 90-95% efficiency, 2 MHz frequency
	Adafruit TLV62569*	Inexpensive, 1.2 A, 90-95% efficiency, 1.5 MHz frequency
	Adafruit TPS62827	2 A, 90-95% efficiency, 2.2 MHz frequency

2.5.1 Hardware Design Details

In our endeavor to develop HabitSense, a device capable of running obfuscation algorithms and real-time detection with a focus on reliability and reproducibility, we embraced a modular approach in its design. The development of Habitsense involved the following critical components: a microcontroller unit (MCU), custom shield, sensors, and power management modules and support components such as MOSFETs and external memory (see Figure 2.4). Our meticulous review of the hardware design landscape, as

outlined in Table 2.4, guided our choice of the ESP32-cam development board to serve as the central platform for the system. This choice was motivated by its affordability (\$5), compatibility with a 24-pin camera, built-in SD card port, WiFi and BLE capabilities (to trigger timely smartwatch or smartphone notifications to obtain further user context), a dual-core processor (240MHz) with substantial RAM (520 KB SRAM and an external 4MB PSRAM) for on-device machine learning through TensorFlow Lite, and nine input/output ports that support various communication protocols (e.g., UART, I2C, SPI, ADC, and DAC). To seamlessly integrate additional components such as the IMU, thermal camera, and RTC, we engineered a custom shield facilitating easy connection to the ESP32-CAM board. This shield utilized the board's pins, channeling them to meet our data transfer, sensor integration, and power management needs. Addressing the focus groups' demand for high resolution and all-day battery life, we selected the OV2640 camera module with a 160-degree fisheye lens. This module not only offers a 1600x1200 resolution with a power draw of 140 mW, as detailed in Table 2.4, but is also equipped with an on-chip JPEG encoder to reduce the processing load on the MCU and RAM usage. The camera's standby mode draws less than 1 mA current. Should a future need arise for higher resolution without power constraints, we can effortlessly switch to the OV5640 camera module, which offers a 2592x1944 resolution but at a higher power consumption rate. To overcome the limitations of the 8x8 IR array in Gen 2, we selected the MLX90640 thermal camera, comprising a 24x32 IR sensor array and a wide field of view. For motion detection, we selected the LIS3DH triple-axis accelerometer for its low power consumption and proficiency in discerning device orientation and user engagement. This empowers the system to actively discern optimal periods for data collection or to execute complex real-time analyses. The accelerometer module can be upgraded to more advanced IMUs like the BNO055 for sophisticated motion detection tasks. Furthermore, in response to the drawbacks encountered with RTC integration in Gen 1 and 2, we opted for the PCF8523 RTC

due to its standalone power source, guaranteeing independent time keeping from the system's main power supply while also allowing for potential upgrades to high-precision RTCs like the DS3231 for users necessitating superior time-keeping accuracy. To optimize the size, weight, and unobtrusiveness of HabitSense, we equipped the system with a 3.7 V, 150 0mAh LiPo battery. We ensured the provision of stable voltage and current to all components by integrating a high-efficiency (90-95%) voltage regulator (Adafruit TLV62569), which converts the input voltage to a consistent 3.3 V in conjunction with a micro-USB LiPo charger. While this configuration met our initial requirements, we designed the system with the flexibility to accommodate alternative voltage regulators, such as those with a broader input range (4.5-21 V), for adaptation to higher-voltage power sources as necessary. The charging circuit of our micro-USB LiPo charger, capable of delivering up to 500 mA, enabled a charge cycle of approximately three hours and included LED indicators to signify full charge status. In contrast to directly soldering the battery to the PCB in Gen 1 and 2, we adopted a JST connector for the battery in HabitSense, simplifying the battery replacement process and allowing for adjustments in battery capacity according to user needs.

The 3D-printed enclosure for the device underwent numerous iterations to refine its size, shape, mechanism, materials, and printing technology (FDM to SLA, SLS), as seen in Figure 2.5. These modifications ensured secure hardware containment, unobstructed camera views, easy micro-USB access, and compatibility with a neck lanyard attachment. To address issues of detachment and instability observed in earlier designs, a new triangular magnetic back plate was introduced (Figure 2.3). The final design also featured a hinge mechanism that orients the camera lenses toward the wearer's face, allowing for an activity-oriented, privacy-preserving design (see Section 9.3 and Figure 2.5 for details regarding the iterative case design).

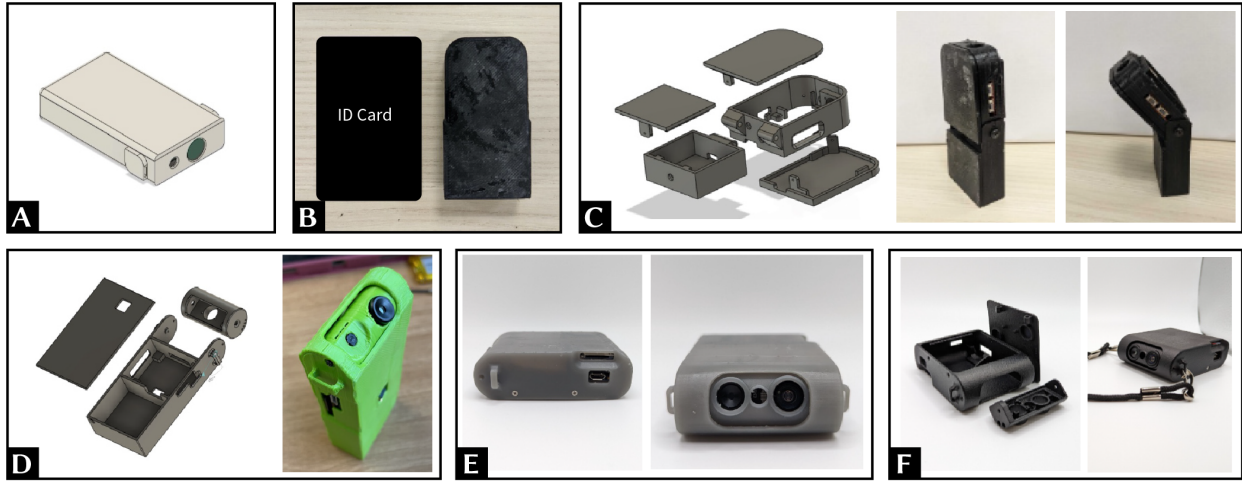


Figure 2.5: Iterative Design of HabitSense Enclosure. (A) Slim initial 3D design, (B) size comparison with ID card, (C) two-piece hinge-connected enclosure, (D) rotatable internal mechanism for sensor adjustment, (E) revised enclosure with optimized dimensions, (F) final product-quality design.

2.5.2 Iterative Case Design

2.5.2.1 Iterations through different 3D printing technologies

The initial prototypes of the device’s enclosure, depicted in Figure 2.5, were fabricated using fused deposition modeling (FDM) with polylactic acid due to its rapid and accessible printing capabilities despite producing lower-fidelity prints. Progressing to later stages, we adopted Stereolithography (SLA) with a FormLabs printer, capitalizing on its ability to handle resins with varied mechanical properties and create complex geometries beyond the scope of FDM. The higher resolution of the FormLabs printer was crucial, as it permitted the precise replication of small design features and components, achieving an optimal fit with minimal tolerances. Employing high-resolution SLA printing with a FormLabs printer allowed for the precise placement of RGB and thermal sensors, facilitating the standardization of calibration parameters across all devices, a process detailed in section 5.4 – on-device obfuscation. For the final design iteration, we utilized selective

laser sintering (SLS) technology with nylon to surpass the resolution offered by SLA. We applied vapor smoothing to the SLS prints to alleviate their inherent rough texture.

2.5.2.2 Iterations in design: hinge mechanism

We systematically enhanced the device design through empirical prototype testing using various 3D printing techniques. Generations 1 and 2 utilized L-shaped components set at diverse angles, allowing the device to be positioned by leveraging the L-shape for variable inclinations. We made a notable improvement in Gen3 – HabitSense by adding a hinging mechanism into the enclosure itself to adjust the field of view for wearers, as documented in Figure 2.5. Initial designs for this mechanism included separate compartments for the battery and circuitry, resulting in a cumbersome design with exposed wiring [Fig 2.5C]. Later designs [Fig 2.5 D,E,F] compacted the structure by limiting the hinge’s movement and relocating only the thermal and RGB cameras to a rotatable cylindrical section, which could be fixed at specific angles with screws for stability. These modifications not only reduced the bulkiness compared to Generations 1 and 2 but also improved the adaptability for users with different body types. We also refined the case design from an initial compact, rectangular shape with sharp corners to a form with rounded edges, improving ergonomics and comfort. Adopting SLA and SLS printing technologies enabled the creation of these curves, which were not attainable with FDM. Despite a slight increase in size, this trade-off was considered beneficial for its ergonomic and aesthetic gains.

2.5.2.3 Magnetic backplate design

Ensuring the device remained stationary and did not sway when worn was critical for collecting high-quality data. To address the detachment issues and instability stemming from an inadequately strong magnetic plate in the Gen 1 and Gen 2 designs, which led to the device swaying or detaching during moderately active movements, we developed

a new magnetic back plate. This redesigned plate featured a triangular shape with three points of contact, utilizing three magnets instead of the previous two, providing a more robust and secure device attachment through the fabric. The advancements in the system's volume and weight, as delineated in Table 2.3 (Gen 3: HabitSense), along with the adoption of stronger magnets and a triangular design for augmented contact points, effectively anchored the device in place, eliminating any sway. Consequently, users could wear the device and engage in their daily routines without the distraction or concern of its stability, assured by the magnetic backplate's reliable hold.

2.5.2.4 Privacy-preserving case design

We designed the device's case to position the RGB and thermal cameras towards the user's face, concentrating data collection on the target activity—hand-to-mouth gestures associated with eating and smoking. This orientation minimized the recording of bystanders, maintaining privacy for those who had not given consent.

2.5.3 Study Design and Data Collection

Utilizing the final iteration of our embedded RGB-T camera system, contained within the final iteration of its corresponding 3D-designed enclosure, we assembled multiple HabitSense devices and recruited 15 participants (7 smokers, 5 with obesity, 3 without obesity) from the Chicago area to participate in a free-living observational study. Recruitment was conducted through online advertisements, seeking adults that were over 18 years of age, fluent in English, who own a smartphone, with a body mass index (BMI) over 30 and under 30 kg/m^2 for people with obesity and without obesity, respectively; to obtain enough data, smokers needed to average at least ten cigarettes daily. We screened and subsequently enrolled eligible participants, equipped them with a HabitSense wearable

device, and provided instructions for its use. Participants installed SNaPsHOT (Smoking, Nutrition, and Personal Habit Observation Tool), an app developed at HABits Lab, and were instructed to log eating or smoking events by opening the app and tapping the appropriate "I'm Eating" or "I'm Smoking" button to record a timestamp for each activity. Over a week, participants wore the device during waking hours and used the app to track their behaviors before returning the device to the lab and filling out surveys about their experience. Notably, participants were not restricted to using the device exclusively during daylight hours or at night, with the only directive being not to wear the device while showering. This precaution was necessary for their safety, as the current version of the device is not waterproof.

2.5.4 SECURE: Sensor-Enabled Control for Ubiquitous Recording and Evaluation

We propose SECURE, a sensor-driven algorithm designed to optimize data collection for devices such as HabitSense, ensuring the acquisition of only essential data useful for researchers, dietitians, or smoking cessation counselors and the users themselves. The core idea of SECURE is to capture obfuscated RGB data that preserves privacy, initiating recording solely during instances with a high probability of a person engaging in eating or smoking activities. Unnecessary data acquisition results in (1) excess power consumption, (2) suboptimal utilization of computational resources, and (3) potential privacy concerns due to over-collection of data. To tackle these challenges, our algorithm employs a three-tiered hierarchical strategy tailored for real-time operations. We leveraged the microcontroller's deep sleep functionality, intermittently awakening it to assess the user's device wearing status using accelerometer data (see Figure 2.6). Subsequently, we analyzed a batch of thermal frames for hand-to-mouth gestures, activating the recording of

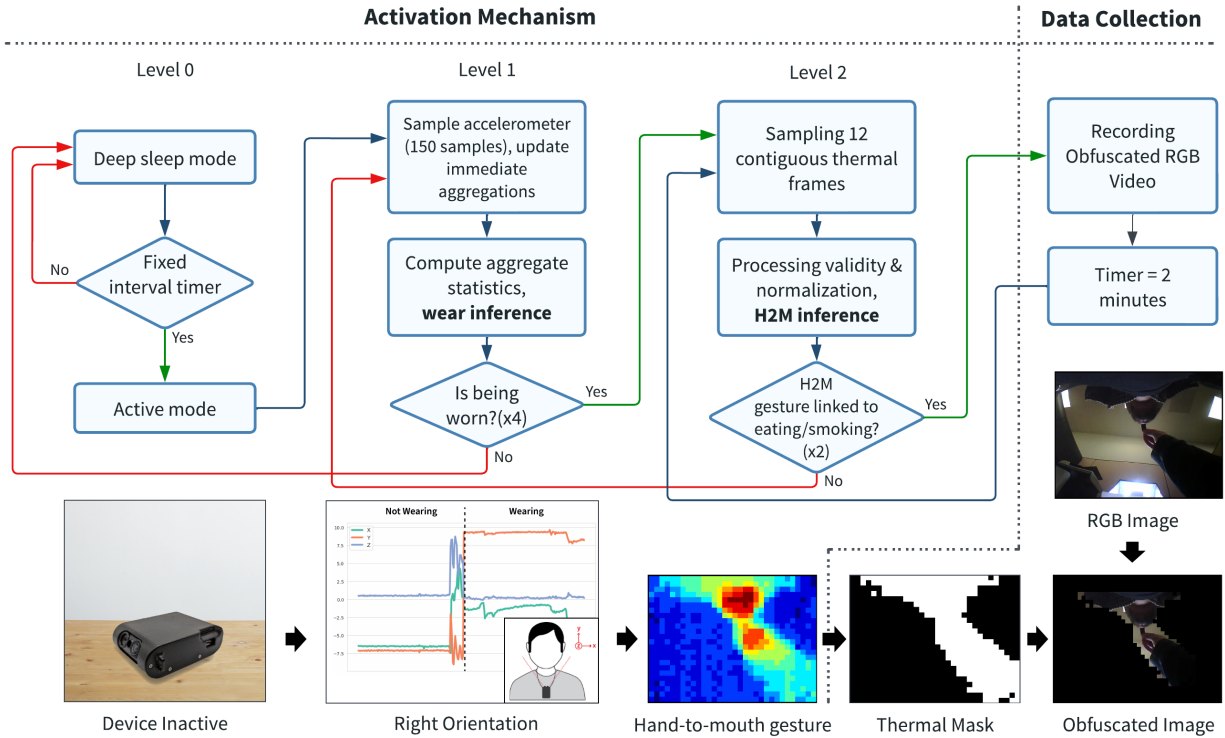


Figure 2.6: Sensor-Enabled Activation Algorithm. Outlines the process from power-on to data recording, starting with deep sleep mode, followed by accelerometer checks for the device being worn (level 1), thermal frame sampling for hand-to-mouth gesture detection (level 2), and concluding with obfuscated RGB data recording, designed to record only essential data to optimize battery life and memory use.

obfuscated RGB frames upon gesture detection before setting the microcontroller back to its sleep state.

2.5.4.1 Pipeline

Level 0

Our approach involves managing the clock frequency of the ESP32 MCU for power efficiency (see Figure 2.6). By default, the MCU operates at 240 MHz, but we set it to enter deep sleep mode when not in active use by using the ESP32's ultra-low power (ULP) co-processor, reducing the clock frequency to 150 KHz. After a predetermined time, the primary processor periodically awakens and operates at its minimum clock frequency of 80

MHz. To preserve temporary variables during sleep mode, we store them in the system's RTC memory, effectively conserving energy during device inactivity.

Level 1: Training a wearing-detection model

Upon reactivation, the primary processor performs accelerometer polling to determine the device's wear status, as shown in Figure 2.6. We detect wear over a 150-sample window (30 seconds) of three-axis accelerometer data by computing a set of ten running metrics: mean and variance of the x, y, and z-axis; mean, standard deviation, and variance of the L2-normalization of the three axes (removing gravity [energy]); and max change in energy differential. At the end of each 30-second segment, we compute a wear/not wear prediction using a decision tree. The decision tree was trained with a maximum depth of 10 to run on the device efficiently. The metrics were chosen based on individual thresholding optimization using f-beta scores. To stabilize and reduce false predictions, we only commit a change in wear-state when we observe five sequential stable predictions (2.5 minutes; e.g., five repeats of "wearing" or "not wearing"). The state of "not wearing" is rare since the device is usually switched off when doffed. Acceleration alone cannot detect all edge cases of "not wearing" precisely when the device is in motion but not in the correct positioning; therefore, so as to not miss data, the model was optimized to reduce false-negative predictions of wear time (0.60% false-negative rate), allowing other levels to filter out false positives (20.26% false-positive rate).

Level 2:

Should a person be detected wearing the device at level 1, we activate our thermal camera through a MOSFET and initiate sampling thermal frames. We then use a window of 12 contiguous thermal frames to estimate the presence of potential hand-to-mouth gestures associated with eating or smoking gestures. To identify these hand-to-mouth ges-

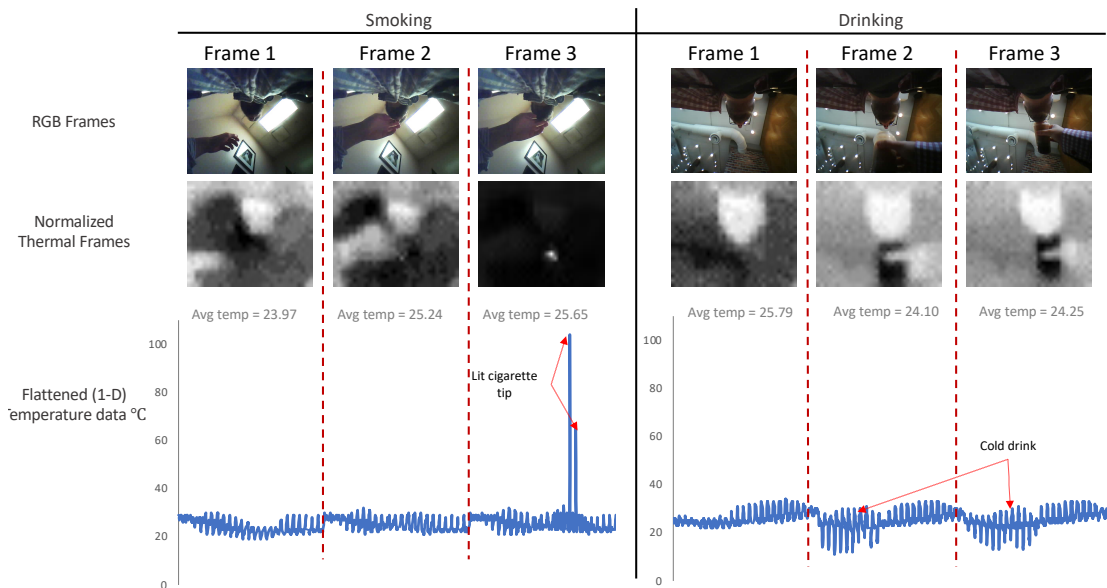


Figure 2.7: HabitSense gesture characterization. Illustrates smoking and drinking gestures in RGB and thermal data, with RGB frames confirming gestures, normalized thermal frames showing temperature distribution, and flattened thermal histograms highlighting the heat signatures during eating and smoking.

tures, we developed a lightweight machine learning model capable of real-time detection on-device. We trained a two-layer neural network with dense layers on a sequence of contiguous thermal frames to incorporate the temporal aspect of thermal patterns associated with hand-to-mouth gestures. This is supported by the distinct temperature signatures observed during eating or smoking gestures (see Figure 2.7). A potential challenge when working with thermal frames lies in the fluctuations of thermal signatures caused by changes in the surrounding environment, which we addressed by applying min-max normalization to the entire window of thermal frames before their utilization in model training.

Data collection level:

When a hand-to-mouth gesture is detected, the system starts capturing obfuscated RGB video (obfuscation algorithm described in the following section) for a user-defined period

of time. The on-device obfuscation mechanism provides privacy-preserving data collection while providing visual confirmation (ground truth) of the activities of interest.

2.5.4.2 Energy profiling

To determine the achievable hours of daily usage with our SECURE algorithm, we calculated the device's average current usage, considering its dynamic operation across level 0, level 1, and level 2 and the data-collection level. We estimated the average daily wear time (t^w) and non-wear time (t^{nw}) based on our wearing detection algorithm, as well as the average obfuscated recording time (t^{obr}). We simulated 'Z' minutes of data recording following a detected hand-to-mouth gesture. Denoting the current consumption of each level as i_{l1} for level 1, i_{l2} for level 2, and i_{dc} for data collection, we calculated the average daily current consumption:

$$\text{Average current} = \frac{(i_{l1} \cdot \sum t_j^{nw}) + i_{l2} \cdot (\sum t_j^w - \sum t_j^{obr}) + (i_{dc} \cdot \sum t_j^{obr})}{\sum t_j^{nw} + \sum t_j^w}$$

Careful consideration was given to the distinction between the system's operational states. Specifically, when the system operated in level 2, the device was being worn but not actively recording; this contrasts with periods where obfuscated data recording was in progress, necessitating a higher current draw as seen in Figure 2.8. Therefore, in the equation presented, we ensured the accuracy of our average current estimation by subtracting the obfuscated recording time (t^{obr}) from the total wear time (t^w) to reflect the differentiated current levels during device wear (level 2).

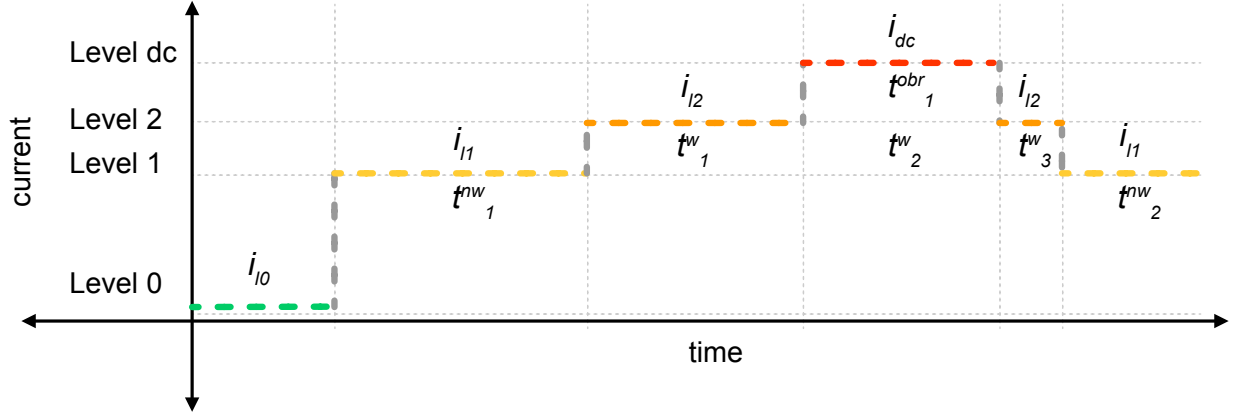


Figure 2.8: SECURE pipeline energy profiling current consumption during different levels.

2.5.5 On-Device Obfuscation Algorithm

On-device obfuscation strategically masks background details to maintain focus on foreground activities, particularly hand-to-mouth gestures relevant to eating and smoking detection. We utilize the thermal sensor’s capabilities, leveraging the proximity of the wearer’s head and hands to the camera, which manifests as a distinct heat signature. This thermal information is overlaid onto the RGB data to create a human thermal mask that isolates and preserves the wearer’s head and hand imagery while rendering the surrounding region opaque.

Our on-device obfuscation algorithm utilizes an adaptive masking operation to overlay thermal camera data onto RGB frames while accommodating disparities in resolution (640x480 for the RGB sensor and 32x24 for the thermal sensor), as well as differences in field of view. The superimposition process requires a one-time initial calibration to synchronize the two data modalities. Following this calibration, the derived parameters apply to all HabitSense devices, attributable to our 3D design that maintains a fixed positional relationship between the RGB and thermal sensors. For calibration, we built a

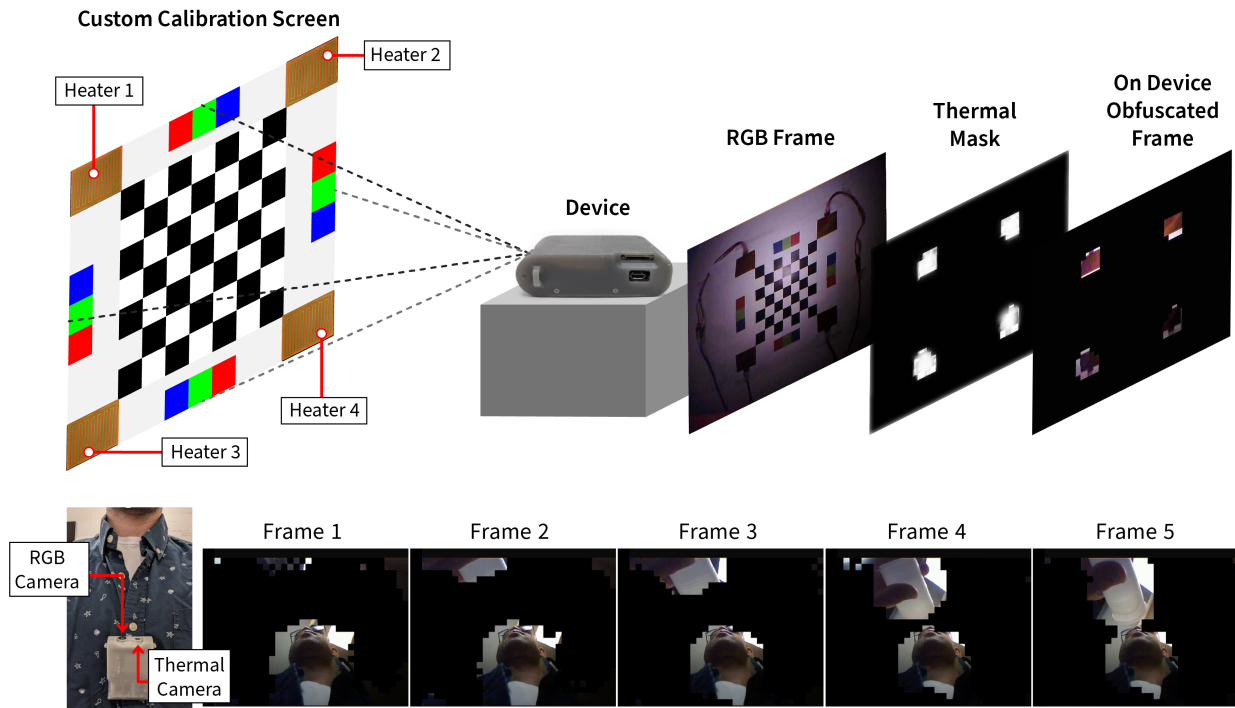


Figure 2.9: Sensor Calibration and RGB Obfuscation Process. Depicts calibration via a screen with heated and colored sections for sensor alignment and shows how calibrated thermal data enables obfuscation in RGB sequences, keeping heads, hands, and objects in hands clear while obfuscating the background.

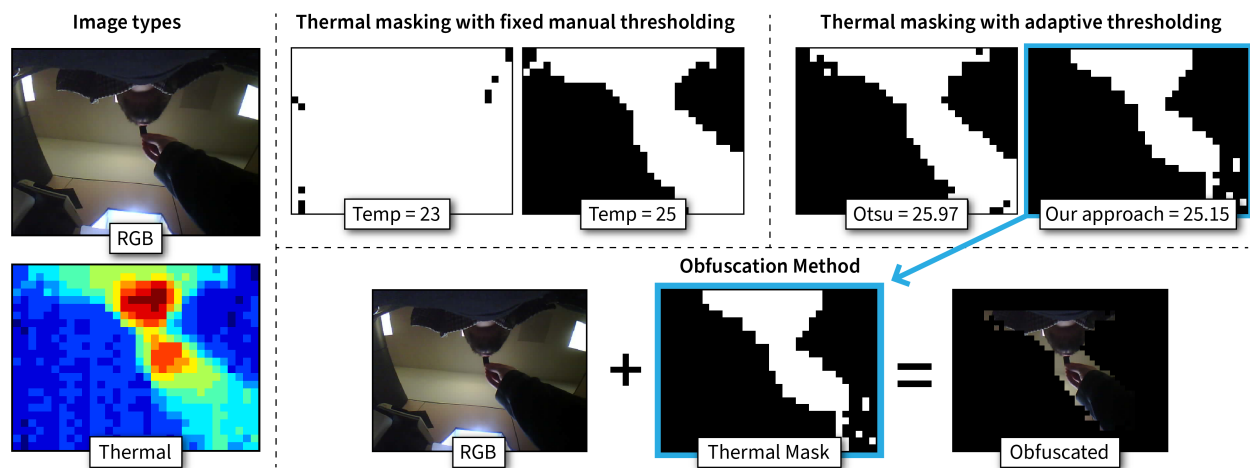


Figure 2.10: Obfuscation threshold determination.

Algorithm 1 On device thermal mask-based obfuscation

```

1: function MASKCALIBRATION(buffer, height, width, mlxFrame, threshold)
2:   offsetLeft, offsetRight, offsetTop, offsetBottom  $\leftarrow$  calibratedOffsetValues
3:   scaleX  $\leftarrow$  (width - (offsetLeft + offsetRight))/32.0  $\triangleright$  Calculate X scaling factor
4:   scaleY  $\leftarrow$  (height - (offsetTop + offsetBottom))/24.0  $\triangleright$  Calculate Y scaling factor
5:   for y  $\leftarrow$  0 to height - 1 do  $\triangleright$  Iterate over height
6:     for x  $\leftarrow$  0 to width - 1 do  $\triangleright$  Iterate over width
7:       bufferIndex  $\leftarrow$  (y · width + x) · 2  $\triangleright$  Calculate buffer index
8:       isEdge  $\leftarrow$  y  $\leq$  offsetTop or y  $\geq$  height - offsetBottom
9:       isEdge  $\leftarrow$  isEdge or x  $\leq$  offsetLeft or x  $\geq$  width - offsetRight
10:      if isEdge then  $\triangleright$  If it's the offset, blackout pixel
11:        buffer[bufferIndex], buffer[bufferIndex + 1]  $\leftarrow$  0  $\triangleright$  Zero out offset
12:      else
13:        if mlxFrame[mlxBufferIndex] < threshold then  $\triangleright$  Check threshold
14:          buffer[bufferIndex], buffer[bufferIndex + 1]  $\leftarrow$  0  $\triangleright$  Apply
15:        end if
16:      end if
17:    end for
18:  end for
19:  return buffer  $\triangleright$  Return the masked buffer
20: end function

```

custom-designed calibration screen with four heaters at the corners and a checkerboard pattern in the center, as shown in Figure 2.9. We positioned the HabitSense device in front of this screen and recorded frames while actively running a calibration algorithm on the device. This algorithm superimposed thermal masks onto the RGB frames while iterating through varying positional and scaling parameters in real time. Upon manual inspection of the recorded data, we identified and utilized the parameters that yielded the most accurate superimposition of the thermal mask on the RGB frames for our on-device obfuscation algorithm. Our algorithm [ALG 1] initializes by constructing a margin ribbon for the RGB image, zeroing pixels within this margin, using the positioning and scaling parameters to account for the smaller field of view of the thermal sensor (compared with the RGB). We rescaled the thermal frame to match the dimensions of the active area of the RGB image, excluding the margin. Subsequently, the algorithm evaluates the thermal value of each pixel against a predetermined threshold, masking those in the visual buffer that fall below this criterion. Our experiments, illustrated in Figure 2.10, explored several strategies for determining this threshold. Initially, we considered fixed temperatures based on the human temperature range, but these proved suboptimal due to the influence of the device’s distance from the subject’s face and ambient temperature variations. Consequently, we adopted an adaptive approach. Although the Otsu method (used in prior work [12]) for image binarization was tested, we found that computing the average temperature of the pixels in each thermal frame yielded results comparable to those of Otsu, with the added benefit of reduced computational complexity for on-device execution. Finally, the algorithm generates a composite image that accentuates thermally notable regions, focusing on pertinent activities such as the clear depiction of the wearer’s head, hands, and held objects—food, a glass of water, or a cigarette—while the surrounding area is obfuscated to safeguard the privacy of both the wearer and any non-consenting bystanders.

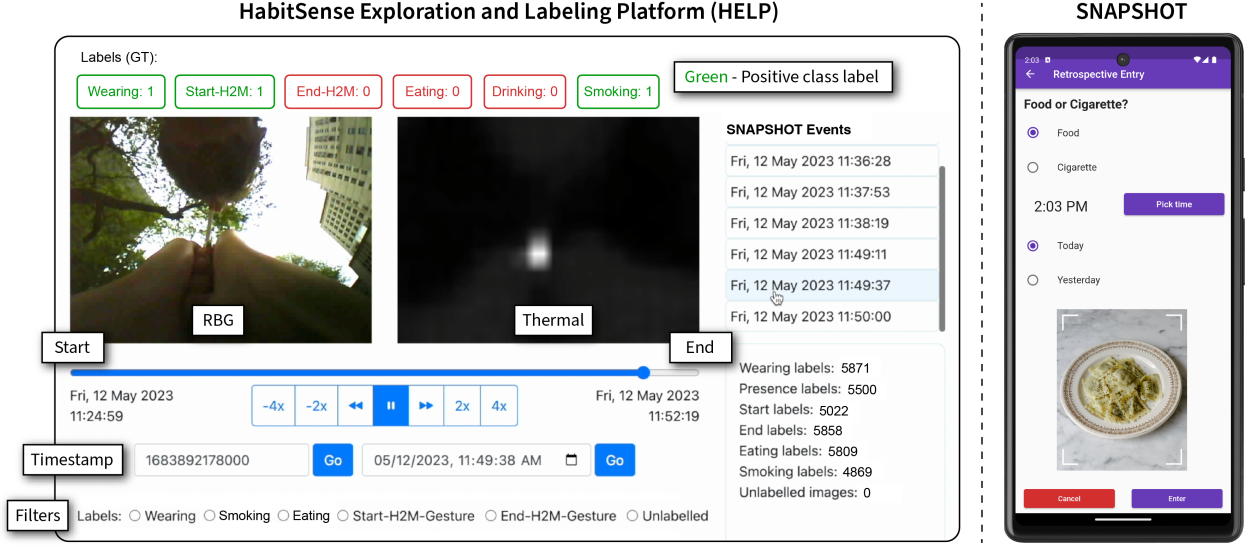


Figure 2.11: HabitSense Exploration and Labeling Platform (HELP) web application (left) and the Smoking, Nutrition, and Personal Observation Tool (SNAPSHOT) mobile application (right).

2.5.6 Gesture Recognition Framework

Having established an efficient data acquisition strategy, we aim to investigate the utility of HabitSense data in identifying instances of eating and smoking behaviors. To ascertain the effectiveness of our collected data in identifying eating and smoking behaviors, we annotated the raw data acquired during the free-living study, thereby enabling the training of machine learning models for behavior detection.

2.5.6.1 Data annotation: HELP and SNAPSHOT

Data annotation represents a significant commitment of time and effort in the data processing workflow. We explored existing annotation tools, such as labelImg and label studio [85], to annotate the start and end of gestures while simultaneously displaying multi-modal data (RGB/thermal/IMU) during the annotation process. However, we found that these platforms are predominantly optimized for object-based annotation and do not provide, by default, the functionality for annotating gesture-specific segments or visualizing

multimodal data in a manner that aligns with our project requirements. To address these challenges, we designed and developed the HabitSense Exploration and Labeling Platform (HELP) [Fig. 2.11], a dedicated system for streamlined data annotation of raw data acquired from video-based behavior monitoring tools.

During our free-living study, we deployed SNAPSHOT, a multi-platform mobile app developed in Flutter, allowing participants to document their eating and smoking activities while wearing the Habitsense device. We engineered the HELP platform such that annotators could reference events logged via the SNAPSHOT application to more efficiently find eating and smoking activities, given the prevalence of inactivity in the dataset. We systematically annotated the start and end of hand-to-mouth gestures associated with eating and smoking activities, in addition to distinguishing periods of device wear from non-wear. The onset of a gesture was defined as the hand approaching the mouth for food intake or puffing from a cigarette, and its termination was marked by the hand retracting from the mouth.

2.5.6.2 Gesture recognition using state-of-the-art spatiotemporal architecture:

Timesformer

In the context of consumption behaviors, the primary gesture preceding eating or smoking is a hand-to-mouth gesture. While 3D convolutional networks have successfully identified spatiotemporal gestures, the emergence of state-of-the-art models, notably transformers [86] built upon attention mechanisms, introduces more efficient approaches for training video data. TimeSformer [87] introduces a significant shift in video analysis using self-attention mechanisms instead of traditional convolutional methods. This change leads to better classification accuracy, quicker training times, and improved efficiency in testing, allowing the model to process long video sequences with spatial and temporal

context effectively. Utilizing this architecture, we could also generate self-attention maps, which illuminate the model’s focal areas, thereby rendering the model not only effective but also interpretable [88], [89].

We trained binary (eating vs. none, smoking vs. none, hand-to-mouth vs. none) and multi-class TimeSformer (eating vs. smoking vs. none) classifiers using annotated RGB and thermal data to detect eating and smoking activities. For each of the four classifiers, we train and test models on (1) RGB-only data; (2) Thermal-only sensor data; (3) RGB Obfuscated-blur (lower-privacy version that blurs the background, gaussian blur with radius 15); and (4) RGB Obfuscated-mask (a high-privacy version masking all background pixels). Through surveys, these two obfuscation methods effectively reduced privacy while maintaining contextual information surrounding hand-to-mouth activity [41]. To inform the sensors and data type used in the online model (Thermal-only, Obfuscated-blur, and Obfuscated-mask), we perform significance tests using an analysis of variance (ANOVA), followed by post hoc pairwise t-tests. Our data were segmented into video clips, each defined by the commencement and conclusion of the specific gesture pertinent to the model—eating, smoking, or hand-to-mouth. For the "none" class clips, a distribution was established mirroring the gesture lengths of the respective positive classes. Using this distribution as a benchmark, we randomly selected sequences of frames for the "none" class, ensuring they adhered to the established length distribution of the positive classes. These clips were then resized to a short edge of 256 pixels following the procedures specified in [90]. We maintained a balanced dataset with an equal representation of 50% for both positive and negative classes. For validation, we implemented a leave-one-participant-out (LOPO) cross-validation strategy, with an 80:20 split for training and validation.

2.5.6.3 Gesture recognition using quantized neural networks on-device

Training state-of-the-art offline models demonstrated the potential of the data collected by HabitSense, but for real-world applications, models that can detect hand-to-mouth gestures associated with eating and smoking in real time are essential. Based on the winning data type from the significance test, we train and test the data type that provides the greatest privacy preservation without impacting the accuracy of activity detection. For the real-time detection capability and to capture the temporal nature of the gestures, we selected a window size of 12 frames. This selection was grounded by the distribution of the length of eating and smoking gestures (Supplementary Figure 2.21). We also mirrored the approach used in the offline models to select the negative class, ensuring class balance during training. We used a training-to-validation ratio of 80:20 and performed leave-one-out cross-validation to evaluate the models. The on-device models, based on a two-layer neural network, were quantized to enhance the real-time detection performance. The inference time was estimated by calculating the number of multiply-accumulate operations (MACs) from the model weights/parameters. Then, by knowing our MCU's number of operations per second, dictated by its clock frequency, we estimate the inference time.

2.5.7 Calculations for Day/Night, Variable Lighting Conditions, and Head Movement

The data collection for HabitSense included a free-living study where participants used the device continuously for 7 days, with unrestricted usage during both day and night. This resulted in data from varied lighting conditions—darker and brighter environments—and included variations in head movement. To ensure a rigorous evaluation, it is crucial to test the robustness of our models under these diverse conditions.

Using an API [91], we obtained the sunrise and sunset times for the participants' locations on the days they wore the device. Subsequently, we calculated the number of eating, smoking, and drinking gestures recorded during daytime versus nighttime.

Addressing potential concerns about the low-light performance of HabitSense, particularly as participants might wear the device at night in well-lit indoor environments, we analyzed the perceived luminosity of each of the 13 million RGB frames, totaling approximately 768 hours of data. Luminosity was calculated for each pixel; thus, for the entire frame, luminosity was determined by averaging the values for all the pixels in frames. We defined a luminosity threshold to distinguish between brighter and darker frames, guided by the perceptual evaluation of light intensity in RGB images and the statistical distribution of luminosity values across all frames (refer to Figure 2.12). The median of

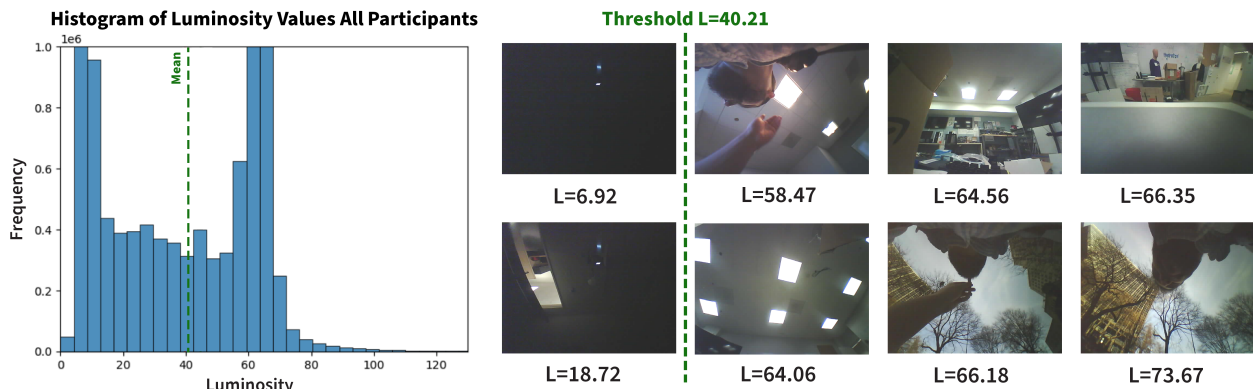


Figure 2.12: Luminosity threshold determination

this distribution served as the threshold, satisfying our criteria for differentiating light conditions. Accordingly, frames exceeding this threshold were classified as bright and those below as dark.

HabitSense was meticulously designed to ensure that the field of view (FOV) of our camera system lenses comprehensively captures the full rotational and translational mo-

tion of the wearer’s neck and head (figure 2.13). To address the challenges associated with accurately detecting eating and smoking activities amid frequent head movements, it was essential first to quantify these movements precisely. We annotated the wearer’s head with a bounding box and the nose with a keypoint across the 13 million frames collected from our free-living study. We analyzed head movements using three key metrics (Figure 2.13):

- The Euclidean distance between the centroids of the bounding boxes across adjacent frames to measure the movement of the head across frames.
- The amount of neck rotation was measured using the nose key points and the centroid of the bounding box across frames.
- The change in the area of the bounding box across adjacent frames to measure the movement of the head towards and away from the camera.

These metrics provided a comprehensive measure of head movements, encompassing both rotational and translational head motions captured throughout our study. We calculated these metrics for all eating and smoking segments in our dataset and identified instances of high versus low head motion.

We assessed the robustness of our models across various sub-categories of data and incorporated attention maps in the supplementary materials as an interpretability metric. This allows us to determine whether the models focus on relevant features for accurate gesture detection.

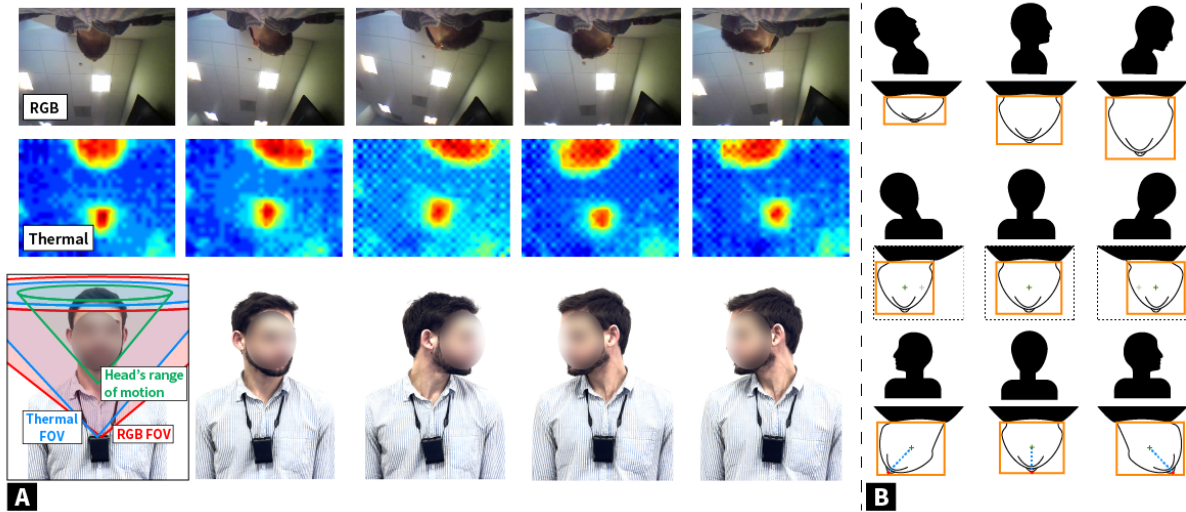


Figure 2.13: (A) Head movement range and fields of view for RGB and thermal cameras, (B) Head movement characterization, including extension and flexion (top), lateral flexion (middle), and rotation (bottom)

2.6 Results

In this section, we present (1) the statistical data from our study; (2) an evaluation of our device conducted with 15 participants (comprising 8 people with and without obesity and 7 smokers), highlighting the advantages of capturing smoking and eating episodes through obfuscated RGB video, triggered by the SECURE algorithm; (3) the evaluation of machine learning algorithms for both the offline and real-time detection of eating and smoking behaviors; and (4) the positive feedback received for HabitSense following its third iteration.

2.6.1 Data Statistics

In total, we collected 46,060 minutes (768 hours) of data with a sampling rate of 5 Hz. The recording duration for each participant ranged from 1,644 to 4,901 minutes (27.4-81.7 hours). Table 5.2 presents the demographic characteristics of the participants. From

these recordings, we captured 10,495 hand-to-mouth gestures across all participants (see Table 2.13). On average, each participant had 807 hand-to-mouth gestures (min=35, max=1726, std=531). Each hand-to-mouth gesture is represented by a start frame and an end frame. By analyzing the frame lengths of all the hand-to-mouth gesture clips captured across all participants, we found that, on average, each hand-to-mouth gesture clip consisted of 12.2 frames (min=9.39, max=20.10, std=2.98). Supplementary Figure 2.21 provides a visual representation of the distribution of hand-to-mouth gesture clip counts per frame length. The combined duration of all the hand-to-mouth intakes amounted to 420.91 minutes (7.02 hours) of smoking, eating, or drinking gestures. When examining individual participants, the time spent smoking, eating, and drinking varied from 1.7 to 97.2 minutes (min=1.7, max=97.2, std=25.32). Two participants were omitted from our data and analysis because of device malfunctions, including water damage, during the free-living study.

Table 2.5: Demographic information of participants.

Participant	Age	Sex	Race	Hispanic/Latino	Smoker	Hours recorded	Ave. hrs per Day
P1	36	male	white	no	no	27.4	5.5
P2	22	male	white	yes	no	60.1	7.5
P3	26	female	black	no	no	59.3	4.2
P4	28	female	white	yes	no	62.9	9.0
P5	47	female	black	no	no	61.9	6.9
P6	40	female	white	no	yes	62.8	7.0
P7	43	female	black	no	yes	69.1	9.9
P8	27	male	black	no	yes	81.7	7.4
P9	27	female	white	no	yes	47.0	5.9
P10	69	male	white	no	yes	74.3	9.3
P11	31	male	asian	no	yes	36.5	7.3
P12	30	female	white	no	yes	70.5	8.8
P13	25	female	white	no	no	54.4	5.4
Mean						59.1	7.2
Total						768.0	

2.6.2 Data Statistics for Day/Night, Variable Lighting Conditions and Head Movement

To assess the robustness of our models, we calculated statistics for different conditions, including day/night, brightness levels, and head motion intensity. Table 2.6 shows that 77.07% of the eating and smoking gestures were recorded during the daytime, with the remaining 22.93% occurring at night. Furthermore, as indicated in Table 2.7, 66.37% of the gestures were recorded under bright conditions, while 33.63% were in darker settings. According to Table 2.8, 80.85% of the gestures involved high-intensity head motion, and 19.15% featured low-intensity movements. These data points highlight the variability in our dataset and highlight the importance of testing model performance across these diverse conditions.

Table 2.6: Daytime-Nighttime table

	no. of gestures captured during day time				no. of gestures captured during night time			
Participant	Smoking	Eating	Drinking	Total	Smoking	Eating	Drinking	Total
P1		128	48	176		0	10	10
P2		386	350	736		152	131	283
P3		95	68	163		115	109	224
P4		1184	266	1450		49	33	82
P5		8	13	21		10	4	14
P6	644	455	194	1293	313	31	89	433
P7	136	539	70	745	21	7	6	34
P8	7	406	48	461	0	28	53	81
P9	164	297	55	516	130	334	85	549
P10	573	93	184	850	467	16	118	601
P11	507	66	43	616	24	0	0	24
P12	72	177	48	297	7	22	6	35
P13		566	199	765		16	20	36
Mean	300	338	122	622	137	60	51	185
Total	2103	4400	1586	8089	962	780	664	2406
% of gestures	77%				23%			

Table 2.7: Luminosity table

	no. of gestures in bright environments				no. gestures in dark environments			
Participant	Smoking	Eating	Drinking	Total	Smoking	Eating	Drinking	Total
P1		128	52	180		0	6	6
P2		391	438	829		147	43	190
P3		78	17	95		132	160	292
P4		902	203	1105		331	96	427
P5		18	13	31		0	4	4
P6	317	331	167	815	640	155	116	911
P7	46	142	34	222	111	404	42	557
P8	7	336	88	431	0	98	13	111
P9	271	601	135	1007	23	30	5	58
P10	749	66	208	1023	291	43	94	428
P11	327	32	17	376	204	34	26	264
P12	47	117	30	194	32	82	24	138
P13		477	181	658		105	38	143
Mean	252	278	122	536	186	120	51	271
Total	1764	3619	1583	6966	1301	1561	667	3529
% of gestures	66%				34%			

2.6.3 SECURE Algorithm Evaluation

2.6.3.1 Assessment of volume of data captured through the SECURE algorithm

Using our pipeline, we evaluated the device wear-detection and hand-to-mouth gesture-detection models and, by assessing different parameter settings, determined the minimum amount of data required to capture the maximum number of hand-to-mouth gestures associated with eating and smoking (see Figure 2.14).

Wearing detection (Level 1)

A decision tree was trained with an 80/20 train/test split on 150-second aggregated samples, resulting in an F1-score of 0.86 on "no wearing" predictions and 0.98 on "wearing" predictions. On the ESP32 microcontroller, the time to compute intermediate values for

Table 2.8: Head Motion Table

	no. of gestures in high motion				no. gestures in low motion			
Participant	Smoking	Eating	Drinking	Total	Smoking	Eating	Drinking	Total
P1		105	43	148		22	15	37
P2		456	398	854		74	74	148
P3		156	113	269		49	44	93
P4		956	229	1185		188	46	234
P5		16	14	30		1	3	4
P6	643	424	221	1288	171	44	34	249
P7	81	405	36	522	24	120	23	167
P8	5	359	69	433	0	64	17	81
P9	243	551	102	896	45	64	38	147
P10	665	65	184	914	353	41	117	511
P11	426	58	32	516	40	0	3	43
P12	58	171	40	269	4	20	5	29
P13		451	131	582		58	71	129
Mean	303	321	124	608	91	57	38	144
Total	2121	4173	1612	7906	637	745	490	1872
% of gestures	81%				19%			

aggregate metrics at each pull of the accelerometer was negligible ($<1 \mu s$), whereas the time to compute metrics for each window was measured at $24 \mu s$ and the decision tree inference time at $5 \mu s$. Overall, this time was insignificant compared to other steps in the pipeline.

Hand-to-mouth gesture detection (Level 2)

Our two-layer network-based hand-to-mouth gesture detection model achieved an F1-score of 0.71. As detailed in Table 2.9, the model occupied 134 KB of memory and necessitated 48 ms for on-device inference time. Additionally, due to the model's training on a 12-frame window, frame normalization on the device required an extra 24 ms, resulting in a total inference time of 72 ms.

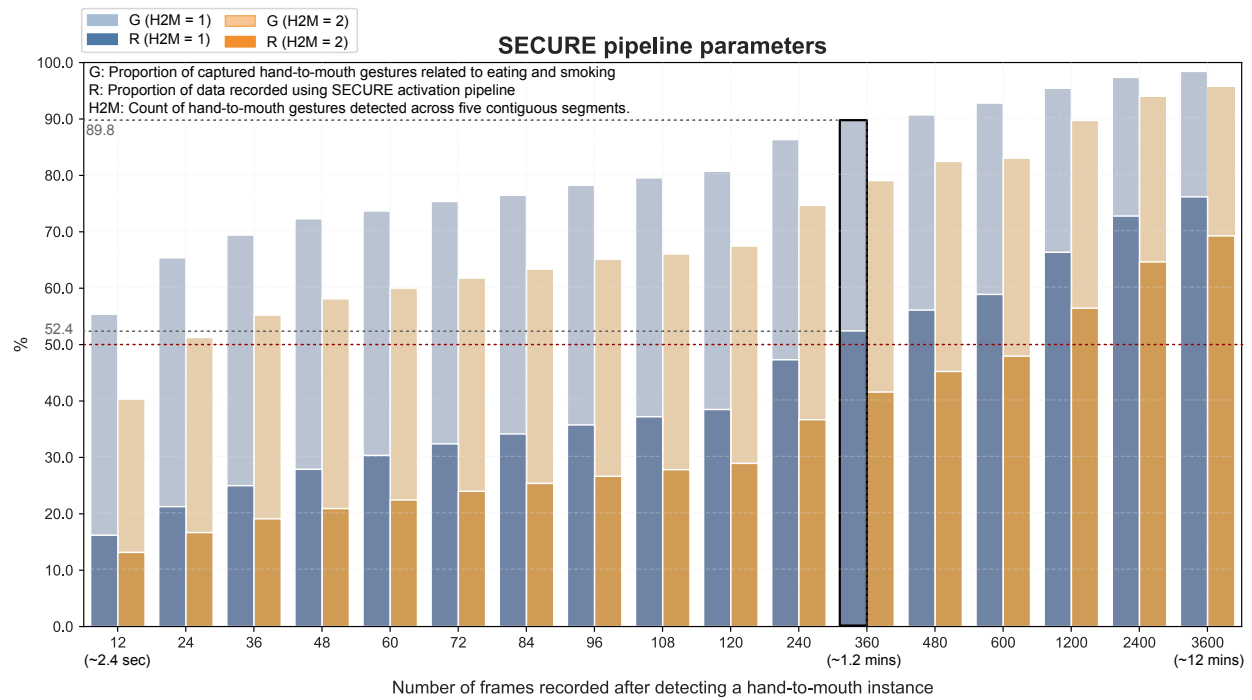


Figure 2.14: SECURE pipeline parameter determination. This figure shows the effects of varying parameter settings on data collection. The graph displays the effectiveness of choosing a threshold of one hand-to-mouth gesture across five segments and a recording duration of 360 frames (1.2 minutes), achieving 89.8% capture of eating and smoking gestures with only 52.4% data recorded.

Table 2.9: Performance of each SECURE pipeline level of the system based on the leave-one-participant-out evaluation.

Level	Precision	Recall	Acc	F1	Model Size	Inference Time
Wearing detection (level 1)	0.97	0.99	0.97	0.98	15.4 KB	$<50 \mu s$
Hand-to-mouth detection (level 2)	0.79	0.67	0.74	0.71	134 KB	48 + 24 ms

Table 2.10: Percent of data recorded versus percent of captured gestures and resulting storage and energy savings using the SECURE pipeline. Utilizing the SECURE pipeline, we captured approximately 89.8% of hand-to-mouth gestures while recording only 52.4% of the data, thereby conserving about 47.6% of storage space and achieving 30.0% energy savings.

Participant	Recorded (%)	Hand-to-Mouth Captured (%)	Storage Saving (%)	Energy Saving (%)
P1	26.35	77.37	73.65	47.12
P2	72.25	96.43	27.75	16.54
P3	60.52	90.27	39.48	20.38
P4	60.82	93.96	39.18	22.25
P5	50.28	82.90	49.72	30.92
P6	55.15	86.66	44.85	23.81
P7	49.61	94.48	50.39	27.59
P8	41.89	93.64	58.11	43.57
P9	44.50	72.99	55.50	33.09
P10	71.94	96.63	28.06	16.63
P11	33.34	96.76	66.66	60.53
P12	64.38	96.85	35.62	18.51
P13	50.36	88.23	49.64	29.33
Mean	52.41	89.78	47.59	30.02

2.6.3.2 Energy profile

We evaluated the energy efficiency of our SECURE pipeline by measuring the HabitSense system’s current consumption through the INA219 current sensor at each SECURE pipeline level, as depicted in Table 2.11. The SECURE pipeline’s design inherently minimizes power usage. It maintains the system in its lowest power state—level 0—when idle, using only the ULP co-processor and drawing approximately 2.8 mA. Based on a 1500 mAh lithium polymer battery and assuming a constant current draw, this translates to a standby time of approximately 2.5 weeks per charge. During active periods, the system periodically polls the accelerometer to determine the device wear status, consuming 57.44 mA. If worn, the

thermal camera is activated via a MOSFET to scan for hand-to-mouth gestures, increasing consumption to 79.44 mA. Should no gestures be detected, the system would operate for approximately 18 hours; however, upon detecting a gesture, it would begin recording obfuscated data using the RGB camera and storing data on the SD card, necessitating a current draw of 139.60 mA. Using the SECURE pipeline, we captured approximately 89.8% of hand-to-mouth gestures while recording only 52.4% of the data, thereby conserving about 47.6% of storage space and achieving 30.0% energy savings (Table 2.10).

Table 2.11: Energy requirements for each level of the SECURE pipeline.

Level	Active Components	Power (mW)	Current (mA)
Level 0	ULP co-processor	10.36	2.80
Level 1	ESP32 (80 MHz), accelerometer	212.52	57.44
Level 2	ESP32, accelerometer, thermal camera	295.03	79.74
Data collection	ESP32, accelerometer, microSD thermal and RGB cameras	471.42	139.60

2.6.4 Evaluation: Usability of Data Collected

In this section, we evaluate the overall performance of the method by assessing the F1-score for all eating and smoking activities of all the participants.

2.6.4.1 Eating and smoking recognition offline analysis

We trained a series of offline models for each activity and input data, reporting the F1 score; Figure 2.15 highlights these results. The TimeSformer architecture-based classification models excelled in hand-to-mouth gesture recognition with an F1-score of 92.7%, closely followed by the obfuscated-blur model and eating detection using RGB data, each achieving 89.7%. However, for smoking detection, the results of RGB versus thermal were much closer, with thermal trailing closely behind RGB video by 85.4%. In addition to the models trained on individual activities, we explored multiclass models, which yielded less effective results, with an F1 score of 82.6% in RGB, thermal performance slightly below

TimeSformer models (offline)



Figure 2.15: Performance comparison of TimeSformer models (offline). H2M, hand-to-mouth.

2-Layer Neural Network (on-device), trained on thermal data

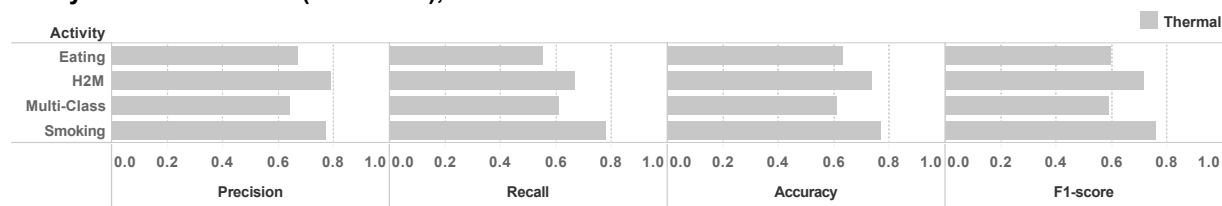


Figure 2.16: Performance comparison of 2-Layer Neural Network, on-device models. H2M, hand-to-mouth.

RGB at 80%, and obfuscated-blur around 71%. We observed a trend where the RGB-trained model generally outperformed models using only thermal data; however, unaltered RGB images pose privacy concerns and are computationally problematic on embedded systems. As a result, we aimed to process either the thermal, obfuscated-blur, or obfuscated-mask on-device.

2.6.4.2 Model Robustness Performance Metrics

As seen in figure 2.17, our eating detection model demonstrates robust performance, achieving F1-scores exceeding 89% in both day and night conditions, as well as during high-intensity head motions. In darker settings, the model achieves an F1-score of 85.58%. Similarly, our smoking detection model shows high efficacy, with F1-scores above 85% during daytime and high-intensity head motions. However, during nighttime and low-light conditions, the model's F1-scores are slightly lower, registering at 79.07% and 83.84%, respectively.

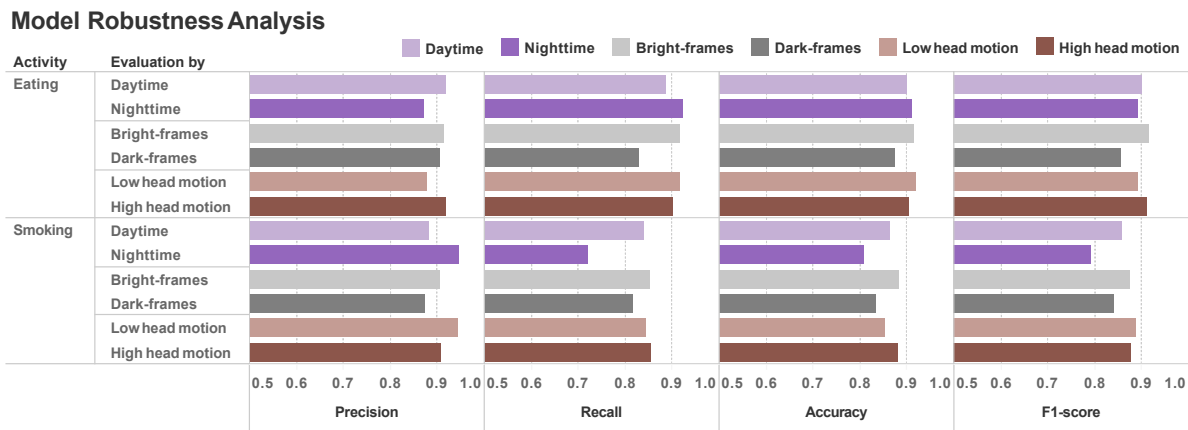


Figure 2.17: Performance of eating and smoking detection TimeSformer models trained on RGB data in different circumstances to highlight robustness

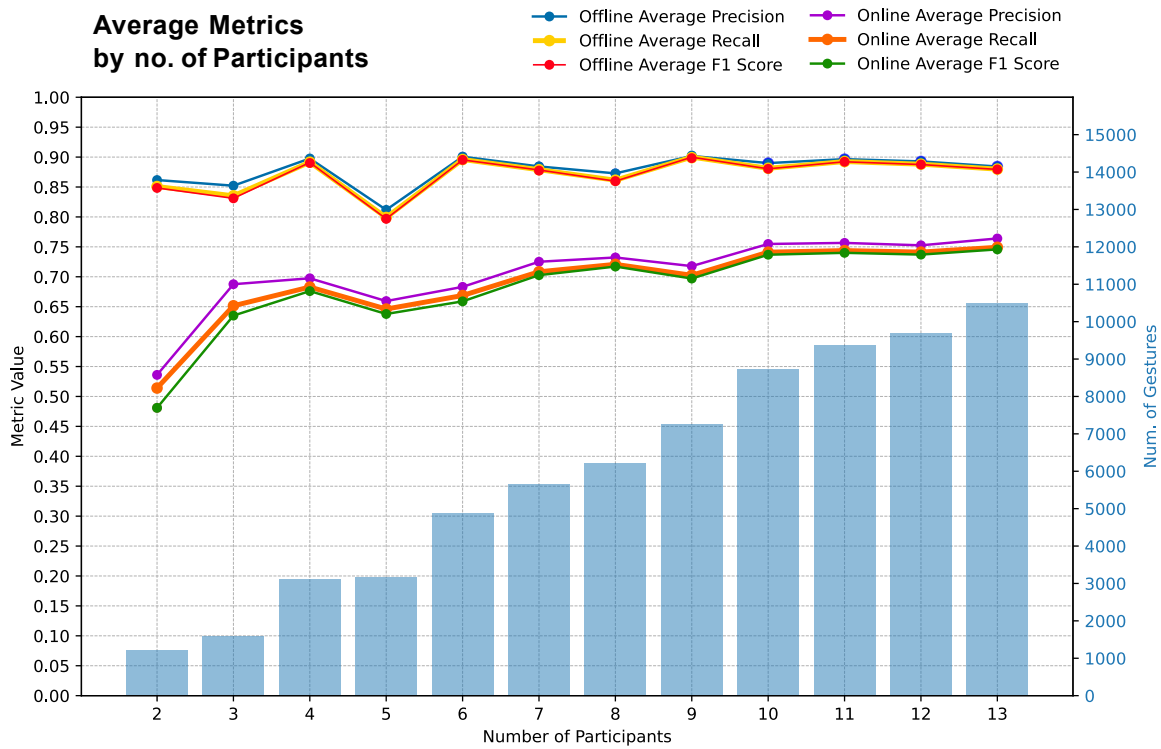


Figure 2.18: This figure illustrates the results of our incremental data addition experiment designed to evaluate the impact of incorporating additional participant data into our model training.

2.6.4.3 Individual Performance metrics

The following list of tables provides individual performance metrics for each of the machine learning models we trained across different settings and data modalities:

- Eating Detection, Off-device RGB: Table 2.16
- Eating Detection, Off-device Thermal: Table 2.17
- Eating Detection, On-device Thermal: Table 2.18
- Eating Detection, Off-device RGB Obfuscated (Mask): Table 2.14
- Eating Detection, Off-device RGB Obfuscated (Blur): Table 2.15
- Smoking Detection, Off-device RGB: Table 2.25
- Smoking Detection, Off-device Thermal: Table 2.26
- Smoking Detection, On-device Thermal: Table 2.27
- Smoking Detection, Off-device RGB Obfuscated (Mask): Table 2.28
- Smoking Detection, Off-device RGB Obfuscated (Blur): Table 2.24
- H2M Detection, Off-device RGB: Table 2.21
- H2M Detection, Off-device Thermal: Table 2.22
- H2M Detection, On-device Thermal: Table 2.23
- H2M Detection, Off-device RGB Obfuscated (Mask): Table 2.19

- H2M Detection, Off-device RGB Obfuscated (Blur): Table 2.20
- Multiclass Detection, Off-device RGB (Weighted Scores): Table 2.31
- Multiclass Detection, Off-device Thermal (Weighted Scores): Table 2.32
- Multiclass Detection, On-device Thermal (Weighted Scores): Table 2.33
- Multiclass Detection, Off-device RGB Obfuscated (Mask) (Weighted Scores): Table 2.29
- Multiclass Detection, Off-device RGB Obfuscated (Blur) (Weighted Scores): Table 2.30

2.6.4.4 Assessing the potential of on-device detection for real-time application

A one-way ANOVA revealed a statistically significant difference in F1-score between at least two groups ($p < .05$) for all eating, smoking, hand-to-mouth, and multi-class activities. Post hoc pairwise analyses indicate no significant difference between thermal and obfuscated-blur ($p < .05$ for all activity types; eating, smoking, hand-to-mouth, and multi-class). However, post hoc pairwise analyses indicated a significant difference between obfuscated-blur and obfuscated-mask ($p > .05$ for all activity types; eating, smoking, hand-to-mouth, and multi-class). In light of these findings, we considered thermal to be the optimal target for the input of the on-device model due to the drastic reduction in input data (pixel count) and given no significance difference between thermal and obfuscated-blur.

Using the internal input data, the on-device models were quantized two-layer fully connected neural networks (reduced size and inference time). These models were expected to have reduced performance due to these optimizations. The best-performing model, individual smoking detection, yielded an F1-score of 75.7%, whereas hand-to-mouth

gesture recognition yielded an F1-score of 71.1%. Although the precision and recall of the smoking model were closely matched, with an approximately 1% difference, the hand-to-mouth gesture model displayed a much higher precision than recall. The model for eating detection produced an F1-score of only 59.5%, which warrants future investigation. As with the offline model, the multi-class model performed below the individual models with an F1-score of 58.8%.

Table 2.12: On-device smoking detection model (thermal data), participant-level analysis.

	Precision	Recall	F1-score	Accuracy
P6	0.61	0.65	0.63	0.62
P7	0.79	0.94	0.86	0.84
P8	0.75	1.00	0.86	0.83
P9	0.89	0.39	0.54	0.67
P10	0.84	0.79	0.81	0.82
P11	0.80	0.82	0.81	0.81
P12	0.70	0.88	0.78	0.76
mean	0.77	0.78	0.76	0.77

2.7 Discussion and Limitations

2.7.1 User Burden and Privacy Evaluation

In the study of participants' perceptions regarding the device, several key insights emerged (Figure 2.19).

When inquiring about participant concerns regarding the information shared by the device (panel A), 87% (13/15) of participants had limited to negligible worries, implying minimal privacy concerns related to the data. For comparison, 77.7% (14/18) of participants who completed the Gen 2 study and 77.6% (38/49) participants who completed the Gen 1 study reported being "a little bit" or "not at all" concerned about what information

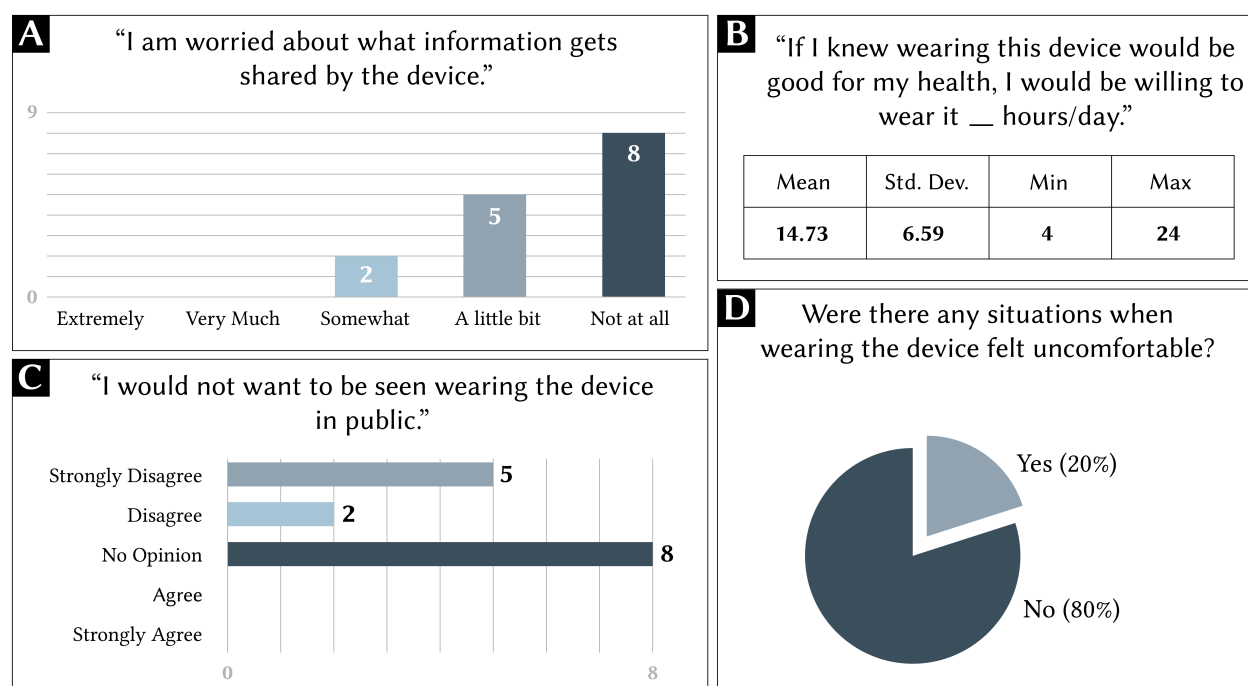


Figure 2.19: Participant responses to burden metrics concerning (A) privacy, (B) obtrusiveness, (C) willingness-to-wear, and (D) physical comfort.

gets shared by the device. When asked how long participants would be willing to wear the device for its health-related benefits, the average intended wear time was approximately 14.73 hours daily (panel B). It is worth noting, however, that a standard deviation of 6.59 hours was observed, and the wear time spanned a range from a minimum of 4 hours to a maximum of 24 hours daily. The actual mean daily wear time of HabitSense study participants was 7.2 hours, compared with 4.3 hours observed in the Gen 2 study. Together, these results indicate a 3 hours-per-day improvement in wear time between Gen 2 and HabitSense. Within the framework of social acceptability, approximately 53.33% of the participants assumed a neutral stance, indicating no distinct preference concerning their visibility while wearing the device in public (panel C). Meanwhile, about 47% of the participants were comfortable with being seen wearing the device in public spaces. One participant commented on the social comfort of the system, saying, "I did not really no-

tice it until I saw people looking at me. They didn't ask, but they 'asked with their eyes.' It didn't bother me, but I was maybe a little conscious of it." Another participant said, "Nobody was even looking at it. They probably thought it was a radio or something." These responses and quotations suggest that the HabitSense wearable constitutes a meaningful step forward in the mitigation of social burden for wearable camera systems. Lastly, 80% of participants found the device comfortable in all situations they encountered, whereas only 20% reported instances of discomfort while wearing the device (panel D).

2.7.2 Privacy Algorithm: SECURE

At level 1 of the SECURE pipeline, acceleration data were used to determine whether the device was worn. However, this method faced challenges in detecting edge cases where the device, while active, was oriented similarly to when worn (while seated or stationary) but placed on the hook of a door hanger. However, the model was optimized to minimize false negatives (0.6%), i.e., instances where wearing was falsely detected as non-wearing, as this is more critical to the system's functionality. Future exploration could involve integrating gyroscopic data to enhance the detection of the orientation of the device. This addition, coupled with a revised case design that limits similar wear orientations when the device is not worn, may improve performance. Another potential avenue is to leverage the thermal camera in wear detection, provided that it is aligned with the system's energy budget constraints.

An important observation from Table 2.10 is the suboptimal performance of the SECURE system on participant P9, who was a smoker. Further investigation into the data, as seen in Table 2.12, showed that the smoking gesture of P9 often involved covering the cigarette tip with their hand, which likely contributed to these detection errors. Moreover, most of the participant's eating and smoking activities occurred inside a car, an environ-

ment characterized by higher indoor temperatures, and less contrast in heat signature. Future work should address the performance of eating and smoking in such scenarios. The final step of the SECURE pipeline involves recording obfuscated data using thermal information to mask the wearer and obfuscate background pixels. We can achieve a frame rate of about one frame per second with an image resolution of 640x480. In comparison, on-device obfuscation techniques have been implemented in various camera systems such as privacyCAM [73], which encrypts the facial region; trustCAM [74], which applies a Canny edge filter to the face; and trustEye [75], which renders the entire frame with a cartoon effect. These systems can operate at higher frame rates, with privacyCAM and trustEye testing at lower resolutions of 320x240. TrustCAM has achieved on-device obfuscation at 640x480 resolution and a higher frame rate of 5 frames per second. However, these systems are not designed as wearable technologies with all-day battery life, nor do they utilize thermal data for masking. Future work should focus on enhancing the frame rate at higher resolutions.

2.7.3 Comparative Analysis of Model Performances for Gesture Classification

In our study, we trained the on-device models using thermal data to achieve feasible inference times and to comply with the device’s resource limitations while enhancing privacy for real-time applications. Our benchmarks using TimeSformer models with thermal data demonstrate performance comparable to obfuscated blur models across all scenarios. However, the on-device models demonstrated superior performance in detecting smoking gestures, with an F1-score of 75.7%, compared with eating gestures, which had an F1-score of 60%. This disparity likely stems from the distinct thermal signatures of smoking gestures, which are more discernible in thermal imagery.

Previous studies, including work by Bi et al. [61], have investigated eating detection using RGB data from head-mounted cameras, achieving F1-scores of 73.8% and 78.7% with 3D CNNs and slow-fast networks, respectively. In contrast, our use of TimeSformers on thermal data alone yielded F1-scores of 85.4% for smoking and 85.6% for eating, highlighting thermal data’s potential to learn eating gestures. Despite these promising results, deploying transformer-based architectures on-device remains challenging. However, recent advancements in transformer technology, such as one-bit based mixtures of experts [92], and ongoing research into on-device integration [93], [94], suggest potential pathways for implementation. To overcome the limitations of resource-constrained devices, combining thermal and RGB data could enable smaller models to extract more comprehensive features, thereby improving on-device eating gesture detection.

2.7.4 Model Robustness

As depicted in Figure 2.17, our eating and smoking detection models demonstrated robustness not only during the night or dimly lit conditions but also during scenarios of intense head movement. The training dataset included participants with a diverse representation of age, sex, race, and ethnicity, as detailed in Table 5.2. To further evaluate robustness, we implemented an additional experiment to assess the impact of incorporating additional participant data into our model training. The results, illustrated in Figure 2.18, indicated that the F1-score of our hand-to-mouth detection model initially increased and then reached a plateau, showing no significant improvement as more participant data were added.

2.7.5 Cigarette vs. E-Cigarette: Gesture Detection Challenges

HabitSense aims to detect hand-to-mouth gestures associated with eating and cigarette smoking. Despite the increasing prevalence of electronic cigarettes (e-cigs), smoking remains the predominant preventable cause of death in the U.S., responsible for approximately 480,000 deaths annually [24], or approximately one in every five individuals. Therefore, in our free-living study, we recruited participants who primarily smoked cigarettes, averaging at least ten daily (recruitment details in section 2.5.3). However, we also found instances of e-cig use in our study data. Such instances were labeled as non-cigarette-smoking gestures, resulting in high precision observed in our LOPO analysis for P8. This is detailed in Table 2.12 of the on-device smoking detection model (thermal data), participant-level analysis. It is important to note that because the SECURE pipeline does not rely on specific eating or smoking detection models for data recording, relying instead on hand-to-mouth gestures, it will still record e-cig use instances.

The growing prevalence of e-cig use, which poses confounding factors for both conventional smoking and the recognition of eating gestures, warrants dedicated investigation [95]. To evaluate the effectiveness of HabitSense on different e-cigs, we performed preliminary tests on three popular devices on the market: Juul [96], Posh Max 5200 [97], and IQOS [98]. Figure 2.20 illustrates the heat signatures of a conventional cigarette versus the e-cigs tested. Although none of the e-cigs tested produced pronounced heat signatures, the IQOS signature is more promising than the Juul and Posh Max signatures, suggesting the potential for future iterations of HabitSense to detect e-cigs that operate based on thermal-passing heating mechanisms. The detection of e-cigs that do not have thermal-passing mechanisms may require the integration of more sophisticated techniques and the fusion of data from both thermal and RGB cameras. This could be made

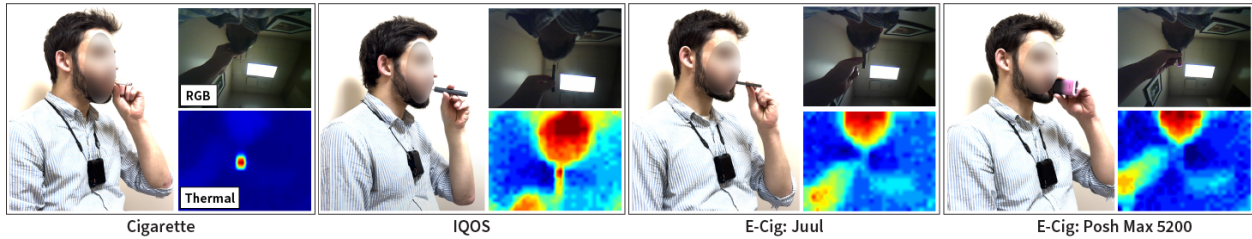


Figure 2.20: Heat signature, cigarette vs e-cigarettes

feasible through HabitSense, which supports both modalities, unlike other wearable camera devices such as Smokemon [13], which only includes a thermal camera. Future work can also test the benefits of a gas sensor or enhanced data collection to discern subtle differences in hand-to-mouth gestures between conventional smoking and e-cig use.

2.7.6 Thermal Sensor Selection

In the transition from the first to the second generation of devices, we replaced the thermal camera with the MLX90640 thermal IR array, which comprises a higher resolution (32x24) and an expanded field of view (110 degrees). This enhancement improved the recognition of hand-to-mouth gestures, increasing the efficacy of the SECURE pipeline. However, the MLX90640 has a notable current draw of approximately 23 mA, as reported in the sensor specifications. During the operation of HabitSense, predominantly at level 3 of the SECURE pipeline, the thermal camera was responsible for 28% of total energy consumption. Another option exists with the MLX90641 sensor, which has a 110-degree field of view but a lower pixel count (16x12), better noise suppression, and lower current draw (12 mA). However, future work is needed to explore the trade-off between improved temperature noise and pixel resolution. If functionality is feasible, this substitution would decrease the current draw at level 2 of the SECURE pipeline from 79.74 mA to 68.74 mA, thus reducing the total energy consumption at level 3 by up to 14%. Another sensor, which would require significant hardware rework due to the higher supply voltage, is the Omron

D6T-32L-01A MEMS thermal sensor, which has 1024 pixels (32x32) and a current draw of 12 mA. Future efforts to scale HabitSense should include efforts to optimize the selection of sensors based on the results of this work.

2.8 Summary

This research has successfully demonstrated the practicality and efficacy of the HabitSense platform, a novel, open-source RGB-T wearable camera system designed for the continuous monitoring and automated detection of health-risk behaviors such as eating and smoking. Informed by thematic analysis of focus groups with 36 experts, including dietitians and smoking treatment specialists, and bolstered by our iterative design, HabitSense balances user privacy considerations with the need for automated detection of hand-to-mouth activity. One of the system's most notable attributes is its SECURE algorithm. SECURE efficiently triggers obfuscated RGB video recording, significantly improving storage savings, reducing battery consumption, and minimizing overall data accumulation, while still maintaining robust event capture rates. Our comprehensive evaluations, in real-world scenarios with 15 participants, have validated the high user acceptability and willingness to use HabitSense. Our study yielded extensive RGB-T data (768 hours), enabling the continued development and testing of advanced machine learning models for offline and real-time on-device analysis, which are robust to free-living environments with varying lighting conditions and head posture. HabitSense offers a low-cost, modular, and scalable solution for reliable all-day data collection while also addressing critical privacy concerns through on-device obfuscation. Our evaluation shows no significant difference between thermal-sensing and obfuscated RGB data in detecting eating and cigarette smoking, paving the way for two privacy-conscious approaches to the detection of health risk behaviors. Finally, the ability of HabitSense to objectively monitor health-risk behav-

iors in real-world scenarios, reducing reliance on self-reporting, elevates the practicality and appeal of wearable cameras in healthcare contexts, setting a new benchmark for the design, development, deployment, and translation of such technologies.

These findings answer (RQ1) by showing how wearable sensing systems can unobtrusively capture real-world behavioral data related to eating and smoking while ensuring user comfort and data quality, thereby reducing reliance on self-report and setting a new benchmark for the design, deployment, and translation of such technologies.

2.9 Supplementary

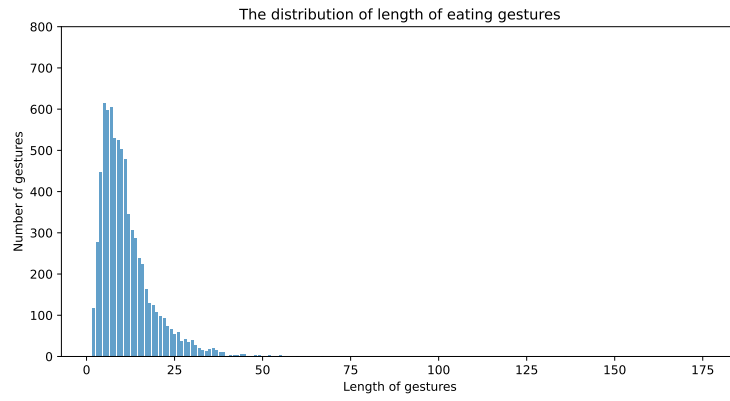
2.9.1 Prior Deployment Protocols

2.9.1.1 Gen 1 Study

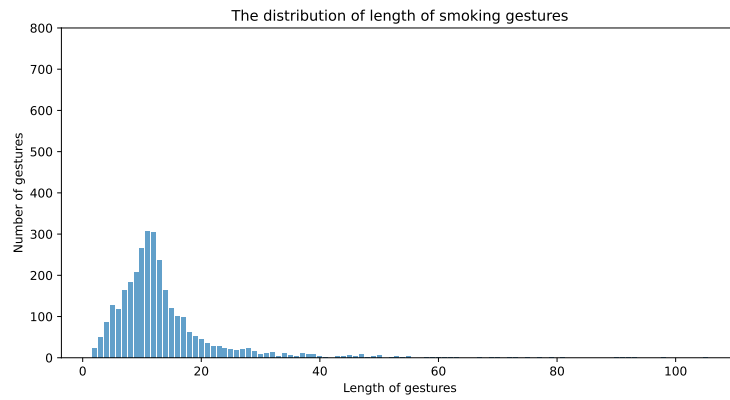
With the Gen 1 device, we collected 14 waking days of data with 60 participants in a free-living setting. Participants attended a laboratory visit to introduce them to the Gen 1 device and calibrate the neckband for optimal body placement. Participants were then instructed to take the device home, equip and turn on the device each morning, perform a synchronization event by aiming the device at a digital clock smartphone application, wear the device throughout the day, remove and turn off the device at the end of the day, and charge the device overnight each night. Participants were informed that recording status was indicated by a blue LED light inside the device's enclosure (blinking = on, solid/absent = off).

2.9.1.2 Gen 2 Study

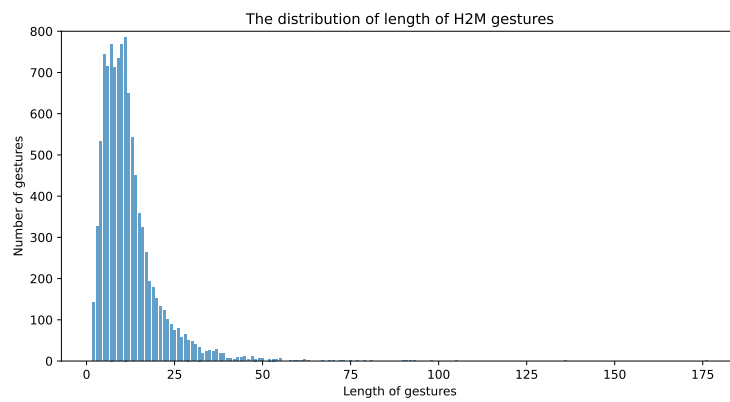
We deployed the Gen 2 device in a free-living study to collect 14 waking days of data from 30 participants, systematically varying which obfuscation technique (blurring, edge obfuscation, cartoon/avatar obfuscation, or raw data/no obfuscation) each participant's device used to test the impact of obfuscation method on wear time. After a laboratory visit to introduce participants to the Gen 2 device and calibrate the neckband for optimal body placement, participants were instructed to take the device home, equip and turn on the device each morning, wear the device throughout the day, remove and turn off the device at the end of the day, and charge the device overnight each night. Participants were informed that recording status was indicated by a blue LED light inside the device's enclosure (blinking = on, solid/absent = off). Participants used a provided laptop to view brief clips of their data collected throughout the day, allowing them to view the effect of the obfuscation technique on their collected data, thus allowing the obfuscation method to affect their wear time.



(a) Eating gesture (mean=10)



(b) Smoking Gesture (mean=12)



(c) Hand-to-mouth gestures (mean=10)

Figure 2.21: Distribution of number of frames comprising one eating gesture (mean=10), smoking gesture (mean=12), and hand-to-mouth gestures (mean=10)

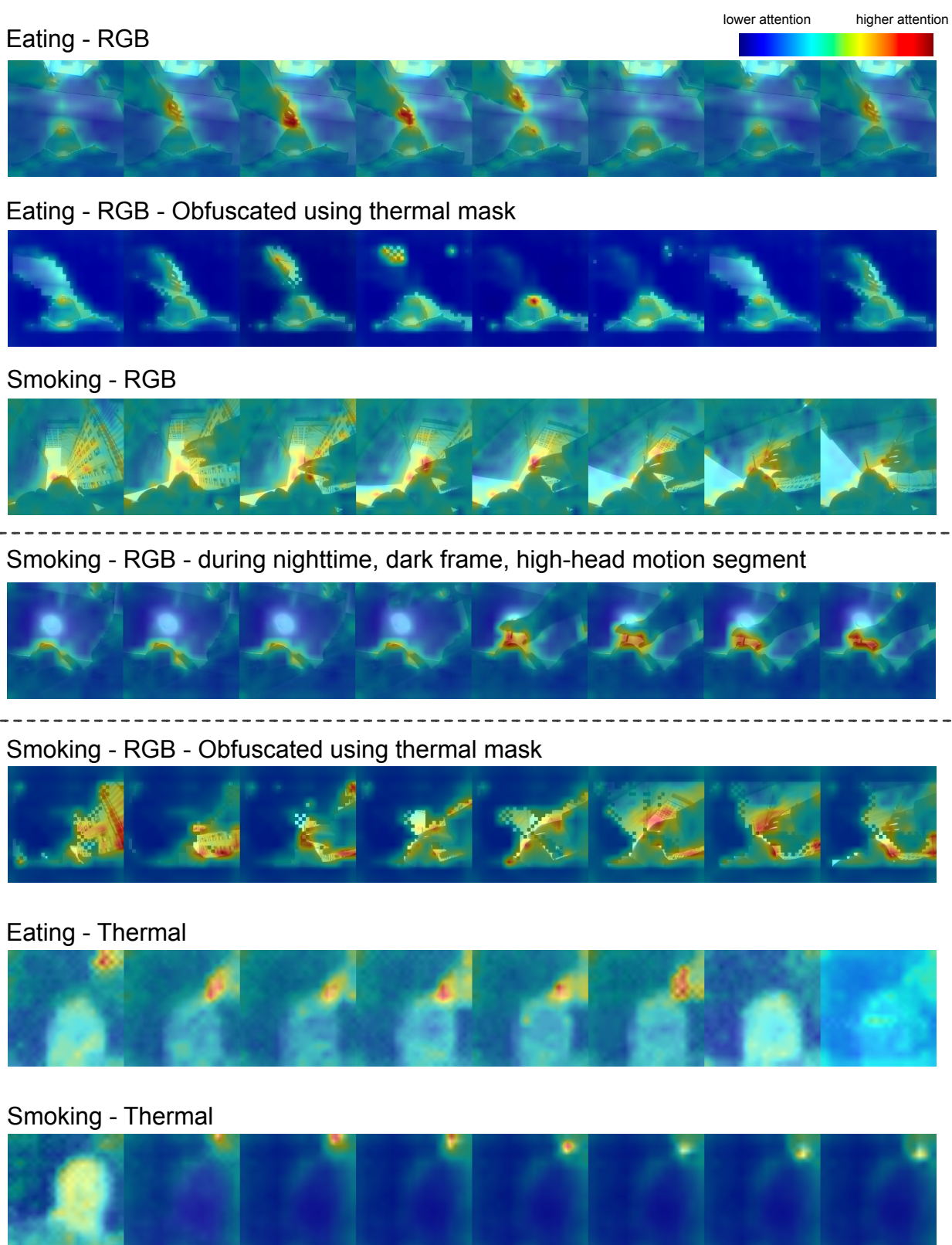


Figure 2.22: Attention maps

Table 2.13: Hand-to-mouth (H2M) statistics across participants.

Participant	Num. of H2M Gestures				H2M Gestures	
	Smoking	Eating	Drinking	Total	Total Frames	Total Duration (mins)
P1	non-smoker	128	58	186	2130	7.1
P2	non-smoker	538	481	1019	9574	31.9
P3	non-smoker	210	177	387	5547	18.5
P4	non-smoker	1233	299	1532	15750	52.5
P5	non-smoker	18	17	35	503	1.7
P6	957	486	283	1726	17892	59.6
P7	157	546	76	779	9610	32.0
P8	7	434	101	542	6257	20.9
P9	294	631	140	1065	10636	35.5
P10	1040	109	302	1451	29161	97.2
P11	531	66	43	640	6441	21.5
P12	79	199	54	332	4766	15.9
P13	non-smoker	582	219	801	8005	26.7
Mean	438	398	173	807	9713	32.4
Total	3065	5180	2250	10495	126272	420.9

Table 2.14: eating offdevice rgb obfuscated mask

pid	precision	recall	f1_score	accuracy
P1	0.80	0.76	0.78	0.78
P2	0.84	0.89	0.87	0.86
P3	0.85	0.77	0.81	0.82
P4	0.91	0.74	0.82	0.83
P5	0.86	0.86	0.86	0.86
P6	0.79	0.82	0.80	0.80
P7	0.80	0.67	0.73	0.75
P8	0.95	0.53	0.68	0.75
P9	0.85	0.79	0.82	0.82
P10	0.63	0.82	0.71	0.67
P11	0.65	0.83	0.73	0.69
P12	0.88	0.79	0.83	0.84
P13	0.93	0.79	0.86	0.87
Mean	0.82	0.77	0.79	0.80

Table 2.15: eating offdevice rgb obfuscated blur

pid	precision	recall	f1_score	accuracy
P1	0.88	0.92	0.90	0.90
P2	0.86	0.93	0.89	0.89
P3	0.92	0.88	0.90	0.90
P4	0.95	0.79	0.86	0.87
P5	0.82	0.80	0.81	0.81
P6	0.87	0.91	0.89	0.89
P7	0.88	0.65	0.75	0.78
P8	0.96	0.65	0.78	0.81
P9	0.88	0.84	0.86	0.86
P10	0.72	0.85	0.78	0.76
P11	0.71	0.81	0.76	0.74
P12	0.86	0.94	0.90	0.89
P13	0.92	0.88	0.90	0.90
Mean	0.86	0.83	0.84	0.85

Table 2.16: eating-offdevice-rgb

pid	precision	recall	f1_score	accuracy
P1	0.92	0.91	0.92	0.92
P2	0.92	0.93	0.92	0.92
P3	0.96	0.84	0.89	0.90
P4	0.96	0.80	0.87	0.88
P5	0.94	0.91	0.93	0.93
P6	0.86	0.98	0.92	0.91
P7	0.90	0.83	0.87	0.87
P8	0.95	0.83	0.89	0.89
P9	0.93	0.94	0.93	0.93
P10	0.77	0.84	0.80	0.79
P11	0.88	0.85	0.87	0.87
P12	0.90	0.95	0.93	0.92
P13	0.95	0.91	0.93	0.93
Mean	0.91	0.89	0.90	0.90

Table 2.17: eating-offdevice-thermal

pid	precision	recall	f1_score	accuracy
P1	0.84	0.97	0.90	0.89
P2	0.96	0.63	0.76	0.80
P3	0.89	0.94	0.91	0.91
P4	0.87	0.90	0.88	0.88
P5	0.76	0.80	0.78	0.77
P6	0.79	0.89	0.84	0.83
P7	0.92	0.90	0.91	0.91
P8	0.95	0.81	0.87	0.88
P9	0.90	0.87	0.88	0.88
P10	0.92	0.85	0.88	0.89
P11	0.81	0.89	0.85	0.84
P12	0.87	0.87	0.87	0.87
P13	0.86	0.82	0.84	0.84
Mean	0.87	0.86	0.86	0.86

Table 2.18: eating-ondevice-thermal

pid	precision	recall	f1_score	accuracy
P1	0.73	0.49	0.58	0.65
P2	0.81	0.60	0.69	0.73
P3	0.72	0.53	0.61	0.66
P4	0.71	0.60	0.65	0.68
P5	0.60	0.48	0.53	0.58
P6	0.58	0.47	0.52	0.57
P7	0.67	0.63	0.65	0.66
P8	0.78	0.46	0.58	0.66
P9	0.45	0.24	0.31	0.47
P10	0.57	0.71	0.63	0.59
P11	0.55	0.78	0.64	0.57
P12	0.65	0.63	0.64	0.65
P13	0.96	0.54	0.69	0.76
Mean	0.67	0.55	0.60	0.63

Table 2.19: h2m-offdevice-rgb-obfuscated-mask

pid	precision	recall	f1_score	accuracy
P1	0.83	0.82	0.82	0.82
P2	0.86	0.94	0.90	0.89
P3	0.91	0.90	0.91	0.91
P4	0.91	0.87	0.89	0.89
P5	0.85	0.80	0.82	0.83
P6	0.92	0.89	0.91	0.91
P7	0.91	0.77	0.83	0.85
P8	0.92	0.85	0.88	0.89
P9	0.93	0.89	0.91	0.91
P10	0.87	0.88	0.87	0.87
P11	0.78	0.87	0.82	0.81
P12	0.89	0.86	0.88	0.88
P13	0.92	0.79	0.85	0.86
Mean	0.88	0.86	0.87	0.87

Table 2.20: h2m-offdevice-rgb-obfuscated-blur

pid	precision	recall	f1_score	accuracy
P1	0.88	0.89	0.89	0.89
P2	0.86	0.98	0.92	0.91
P3	0.94	0.95	0.94	0.94
P4	0.95	0.88	0.91	0.92
P5	0.89	0.91	0.90	0.90
P6	0.96	0.90	0.93	0.93
P7	0.91	0.84	0.87	0.88
P8	0.96	0.89	0.92	0.92
P9	0.96	0.89	0.92	0.93
P10	0.91	0.84	0.87	0.88
P11	0.80	0.83	0.82	0.81
P12	0.89	0.84	0.86	0.87
P13	0.94	0.87	0.90	0.91
Mean	0.91	0.88	0.90	0.90

Table 2.21: h2m-offdevice-rgb

pid	precision	recall	f1_score	accuracy
P1	0.91	0.86	0.89	0.89
P2	0.91	0.99	0.95	0.94
P3	0.95	0.91	0.93	0.93
P4	0.96	0.93	0.94	0.94
P5	0.87	0.94	0.90	0.90
P6	0.95	0.96	0.96	0.96
P7	0.93	0.92	0.92	0.92
P8	0.95	0.95	0.95	0.95
P9	0.95	0.98	0.96	0.96
P10	0.94	0.90	0.92	0.92
P11	0.90	0.89	0.90	0.90
P12	0.92	0.89	0.90	0.90
P13	0.95	0.90	0.92	0.92
Mean	0.93	0.92	0.93	0.93

Table 2.22: h2m-offdevice-thermal

pid	precision	recall	f1_score	accuracy
P1	0.87	0.90	0.89	0.88
P2	0.96	0.66	0.78	0.82
P3	0.88	0.93	0.90	0.90
P4	0.91	0.91	0.91	0.91
P5	0.73	0.77	0.75	0.74
P6	0.91	0.89	0.90	0.90
P7	0.93	0.84	0.88	0.89
P8	0.94	0.87	0.90	0.91
P9	0.92	0.89	0.91	0.91
P10	0.87	0.83	0.85	0.85
P11	0.82	0.86	0.84	0.84
P12	0.89	0.93	0.91	0.90
P13	0.89	0.79	0.84	0.85
Mean	0.89	0.85	0.87	0.87

Table 2.23: h2m-ondevice-thermal

pid	precision	recall	f1_score	accuracy
P1	0.88	0.42	0.57	0.68
P2	0.75	0.72	0.74	0.74
P3	0.79	0.68	0.73	0.75
P4	0.77	0.69	0.73	0.74
P5	0.68	0.48	0.56	0.62
P6	0.80	0.71	0.76	0.77
P7	0.85	0.69	0.76	0.78
P8	0.68	0.73	0.71	0.69
P9	0.86	0.50	0.63	0.71
P10	0.72	0.79	0.75	0.74
P11	0.72	0.86	0.79	0.77
P12	0.78	0.82	0.80	0.80
P13	0.97	0.58	0.73	0.78
Mean	0.79	0.67	0.71	0.74

Table 2.24: smoking-offdevice-rgb-obfuscated-blur

pid	precision	recall	f1 score	accuracy
P6	0.88	0.85	0.87	0.87
P7	0.81	0.92	0.86	0.85
P8	0.50	0.29	0.36	0.50
P9	0.97	0.86	0.91	0.92
P10	0.88	0.56	0.69	0.74
P11	0.76	0.69	0.73	0.74
P12	0.82	0.85	0.83	0.83
Mean	0.80	0.72	0.75	0.78

Table 2.25: smoking-offdevice-rgb-no-obfuscation

pid	precision	recall	f1 score	accuracy
P6	0.94	0.90	0.92	0.92
P7	0.93	0.89	0.91	0.91
P8	0.75	0.86	0.80	0.79
P9	0.96	0.95	0.96	0.96
P10	0.91	0.67	0.77	0.80
P11	0.82	0.78	0.80	0.80
P12	0.92	0.76	0.83	0.85
Mean	0.89	0.83	0.86	0.86

Table 2.26: smoking-offdevice-thermal

pid	precision	recall	f1 score	accuracy
P6	0.90	0.72	0.80	0.82
P7	0.97	0.93	0.95	0.95
P8	1.00	0.86	0.92	0.93
P9	0.87	0.76	0.81	0.82
P10	0.83	0.83	0.83	0.83
P11	0.81	0.77	0.79	0.80
P12	0.82	0.92	0.87	0.86
Mean	0.89	0.83	0.85	0.86

Table 2.27: smoking-ondevice-thermal

pid	precision	recall	f1 score	accuracy
P6	0.61	0.65	0.63	0.62
P7	0.79	0.94	0.86	0.84
P8	0.75	1.00	0.86	0.83
P9	0.89	0.39	0.54	0.67
P10	0.84	0.79	0.81	0.82
P11	0.80	0.82	0.81	0.81
P12	0.70	0.88	0.78	0.76
Mean	0.77	0.78	0.76	0.76

Table 2.28: smoking-offdevice-rgb-obfuscated-mask

pid	precision	recall	f1_score	accuracy
P6	0.80	0.61	0.70	0.73
P7	0.81	0.83	0.82	0.82
P8	0.33	0.14	0.20	0.43
P9	0.94	0.74	0.83	0.84
P10	0.78	0.38	0.51	0.64
P11	0.72	0.54	0.61	0.66
P12	0.72	0.76	0.74	0.73
Mean	0.73	0.57	0.63	0.70

Table 2.29: multiclass-offdevice-rgb-obfuscated-mask (weighted-scores)

pid	precision	recall	f1_score	accuracy
P6	0.69	0.68	0.68	0.68
P7	0.66	0.65	0.62	0.65
P8	0.50	0.52	0.46	0.52
P9	0.75	0.73	0.73	0.73
P10	0.61	0.60	0.60	0.60
P11	0.65	0.63	0.63	0.63
P12	0.70	0.70	0.69	0.70
Mean	0.65	0.65	0.63	0.64

Table 2.30: multiclass-offdevice-rgb-obfuscated-blur (weighted-scores)

pid	precision	recall	f1_score	accuracy
P6	0.84	0.84	0.84	0.84
P7	0.72	0.69	0.67	0.69
P8	0.52	0.52	0.52	0.52
P9	0.82	0.80	0.80	0.80
P10	0.72	0.72	0.72	0.72
P11	0.70	0.68	0.68	0.68
P12	0.79	0.78	0.78	0.78
Mean	0.73	0.72	0.71	0.72

Table 2.31: multiclass-offdevice-rgb (weighted-scores)

pid	precision	recall	f1_score	accuracy
P6	0.90	0.90	0.90	0.90
P7	0.79	0.75	0.75	0.75
P8	0.82	0.76	0.76	0.76
P9	0.91	0.91	0.91	0.91
P10	0.82	0.82	0.82	0.82
P11	0.80	0.80	0.80	0.80
P12	0.84	0.84	0.84	0.84
Mean	0.84	0.83	0.83	0.83

Table 2.32: multiclass-offdevice-thermal (weighted-scores)

pid	precision	recall	f1_score	accuracy
P6	0.79	0.78	0.78	0.78
P7	0.85	0.84	0.84	0.84
P8	0.86	0.86	0.85	0.86
P9	0.80	0.79	0.79	0.79
P10	0.76	0.75	0.75	0.75
P11	0.74	0.72	0.72	0.72
P12	0.86	0.85	0.85	0.85
Mean	0.81	0.80	0.80	0.80

Table 2.33: multiclass-ondevice-thermal (weighted-scores)

pid	precision	recall	f1_score	accuracy
P6	0.46	0.46	0.46	0.46
P7	0.65	0.65	0.61	0.65
P8	0.87	0.78	0.75	0.78
P9	0.62	0.51	0.48	0.51
P10	0.69	0.69	0.69	0.69
P11	0.64	0.63	0.62	0.63
P12	0.56	0.56	0.50	0.56
Mean	0.64	0.61	0.59	0.61

Chapter 3

Modality Fusion

3.1 Introduction and Background

Research Question (RQ2). *How can activity-oriented RGB and thermal data from HabitSense be fused using deep learning to address the multi-class problem?*

Guided by RQ2—and building on RQ1’s demonstration of unobtrusive, privacy-preserving capture with HabitSense—this chapter advances from single-modality hand-to-mouth recognition to robust behavior differentiation via RGB–thermal fusion. Using the HabitSense dataset, we developed early-, mid-, and decision-fusion approaches—culminating in *RT-FVNet*—to achieve reliable multi-class classification of eating, smoking, and non-gesture activities. As shown in Fig. 2.15, single-modality TimeSformer baselines exhibit performance drops when moving from within-modality recognition to the more challenging multi-class setting. This gap motivates fusing complementary cues—fine-grained visual detail from RGB and heat signatures from thermal—to recover accuracy and robustness under real-world variability.

RGB–thermal wearable Activity-Oriented Camera (AOC) systems [10], [14] are used

to monitor health-risk behaviors such as eating and smoking. These systems are typically worn as a necklace, with lenses oriented toward the mouth to capture HTM activities from a unique, non-egocentric perspective (Fig. 3.1), which poses additional challenges for gesture classification and detection. While the HabitSense study focused on gesture recognition using either RGB or thermal data, little work has explored distinguishing multiple gestures (eating vs. smoking) using *both* modalities in activity-oriented footage. RGB excels at detailed visuals but struggles in low-light settings (e.g., dining in dim environments). Conversely, thermal is adept at detecting heat signatures, aiding recognition of eating-related HTM gestures (distinct food temperatures) and smoking-related HTM gestures (the heated cigarette tip). However, thermal sensors are sensitive to ambient-temperature drift and occlusions (e.g., smoking in a vehicle or when the cigarette tip is blocked by one’s hand), which can increase misclassification and reduce reliability. By fusing RGB and thermal, we integrate visual detail with heat cues to improve performance and robustness in multi-class gesture classification—specifically distinguishing eating, smoking, and non-gesture activity.

To answer RQ2, we investigated multimodal fusion for activity-oriented, privacy-aware video by (i) developing and comparing early-, mid-, and decision-fusion models for multi-class HTM recognition, and (ii) introducing *RTFVNet*, a mid-fusion architecture that balances accuracy and efficiency for real-time pipelines. Our contributions are:

1. A free-living RGB–thermal video dataset comprising 441.9 hours of annotated activity-oriented footage.
2. A family of fusion models—early, mid, and decision fusion—for multi-class classification of eating, smoking, and non-gesture activity.

3. *RTFVNet*, a mid-fusion architecture tailored for multi-class HTM gesture recognition in AOC video.

3.2 Methodology

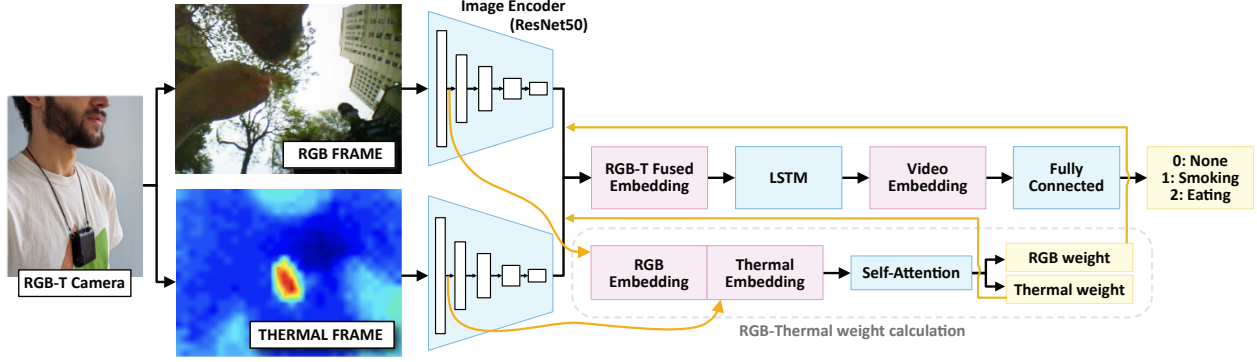


Figure 3.1: RTFVNet Model Architecture

We used the state-of-the-art video classification architecture, TimeSformer [87], to train models on individual RGB and thermal modalities for multi-class classification of eating, smoking, and non-gesture activities. For decision fusion, we combined the outputs from two separate models: the RGB-TimeSformer and the thermal-TimeSformer. In the early fusion approach, RGB and thermal frames were concatenated prior to being inputted into the TimeSformer model. Furthermore, we developed a novel mid-fusion architecture inspired by RTFNet [99], a state-of-the-art architecture for RGB-Thermal-based segmentation. While RTFNet utilizes shortcut connections between encoders to pass feature maps from the thermal encoder to the RGB encoder, which is advantageous for segmentation tasks. To adapt it to our multi-class classification problem, we remove these cross-modal connections. This modification aims to ensure that each encoder specializes in extracting the most relevant features from its respective modality, minimizing cross-modal interference and enhancing the model's ability to leverage distinct modality-specific information prior to fusion [100].

RTFVNet. (Figure 3.1) In our novel RTFVNet model, the initial RGB and thermal frames are fed into an image encoder based on ResNet50, where only the early layers of the ResNet model are used to extract RGB and thermal embeddings separately. These embeddings are concatenated into an RGB-T embedding that captures complementary information from both modalities. A self-attention mechanism then computes the relative importance (weights) of each modality—RGB and thermal. The calculated weights are applied to the respective latent representations at the end of the encoding process, dynamically adjusting their contributions for each RGB-T image pair within the model. The latent vectors are fused and passed through an LSTM layer to capture temporal dependencies across frames. Finally, a fully connected layer outputs a classification into three categories: none, smoking, or eating; leveraging both spatial and temporal features for robust behavior detection. All models were evaluated using leave-one-participant-out cross-validation, with performance metrics in terms of F1-score, precision, recall and complexity measured by TFLOPs.

3.3 Experiments

Dataset. We collected 441.9 hours of RGB-Thermal data from 7 smokers using HabitSense [14], an RGB-Thermal activity-oriented, neck-worn camera system designed with lenses directed toward the wearer’s face. The start and end points of eating and smoking gestures were annotated.

3.3.1 Experiments

We evaluated the performance of our mid-fusion architecture, RTFVNet, using leave-one-participant-out cross-validation and compared it against single-modality and other fusion techniques based on F1-scores and computational costs (TFLOPs) [Table 3.1]. The RGB and

Model	F1-Score	Precision	Recall	TFLOPs
RGB	82.6 ± 2.4	84.0 ± 1.8	82.6 ± 2.3	0.22
Thermal	79.6 ± 2.0	80.7 ± 1.9	79.7 ± 2.0	0.22
Decision Fusion	87.7 ± 1.6	88.7 ± 1.4	87.7 ± 1.5	0.44
Early Fusion	83.8 ± 2.2	84.6 ± 2.2	83.9 ± 2.2	0.22
Mid (RTFVNet)	84.8 ± 2.2	84.8 ± 2.2	85.6 ± 2.1	0.37

Table 3.1: Performance metrics (% , mean $\pm$ standard error) for different models, and computational cost (TFLOPs).

thermal models achieved F1-scores of 82.6% and 79.6%, respectively, with computational costs of 0.22 TFLOPs each. While the decision fusion model achieved the highest performance with an F1-score of 87.7%, it incurred a significantly higher computational cost of 0.44 TFLOPs. Our RTFVNet model balanced performance and efficiency, achieving an F1-score of 84.8% with a reduced computational cost of 0.37 TFLOPs compared to decision fusion.

3.3.2 Model Implementation Details

The following sections outline the specific implementation details of the training and inference procedures for the fusion model experiments.

3.3.2.1 Training.

For early fusion and decision fusion TimeSformer models, we follow the same preprocessing and training process as in TimeSformer [87]. For RTFVNet, we train our custom model for 15 epochs with an initial learning rate of 0.005, which is reduced by a factor of 10 at epochs 11 and 14. This follows the same learning rate schedule as described in TimeSformer. For each video, we sample 9 frames at a rate of 1 frame per second. During training, we resize the images to 224×224 , as we use ResNet-50 weights pretrained on ImageNet, which requires this input size. We train all our models using synchronized SGD across 4 NVIDIA GeForce RTX 2080 Ti GPUs. With Distributed Data Parallel (DDP)

and 4 GPUs, the total effective batch size is 20. The momentum is set to 0.9, and the weight decay is set to 0.0001.

3.3.2.2 Early Fusion.

One approach to combining RGB and thermal inputs is through input-level fusion. Most video classification frameworks accept 3-channel video inputs, so to accommodate the additional 1-channel thermal data, we implemented two early-fusion methods. In the first method, we compute the average of the R, G, and B channels, $(R+G+B)/3$, to form the first channel, use the thermal image as the second channel, and retain the R channel as the third channel. In the second method, we replicate the single-channel thermal image across 3 channels and horizontally stack the RGB and thermal images, row by row. Early fusion achieves F1-scores of 83.8% and 83.5%, respectively, increasing the single RGB modality score of 82.6% by approximately 1%. Early fusion allows for combined feature prediction while keeping the model architecture and computational resources unchanged. However, it does not weigh the importance between RGB and thermal modalities given different inputs.



Figure 3.2: Early fusion inputs for experiment 1 and 2

Table 3.2: Performance metrics of decision fusion experiments.

Fusion	F1-Score	TFLOPs
Experiment 1	83.8 ± 2.2	0.22
Experiment 2	83.5 ± 1.3	0.22
Mid (RTFVNet)	84.8 ± 2.2	0.37

3.3.2.3 Decision Fusion.

Another way of combining RGB and Thermal inputs is at the decision-level. First, we obtain class predictions and probabilities from single-modality Timesformer models. Modality-specific weights are then applied to the RGB and thermal probabilities, and the weighted probabilities for each class are summed across modalities. The final label is assigned based on the highest combined probability. These modality-specific weights are optimized to maximize the overall F1 score in the LOPO (Leave One Participant Out) test. Decision fusion leverages complementary information from both modalities by integrating their independent predictions, with the average score being 87.7%, highest among all fusion methods. However, this approach relies on two separate video encoders, one for each modality, which increases computational costs by 0.07 TFLOPs.

Table 3.3: Performance metrics of decision fusion experiments.

Fusion	F1-Score	TFLOPs
Decision fusion	87.7 ± 1.6	0.44
Mid (RTFVNet)	84.8 ± 2.2	0.37

3.3.2.4 Mid Fusion.

Cross-modal Connections

We constructed two separate encoders to independently extract features from RGB and thermal images, employing ResNet as the feature extractor backbone, with the same setting with RTFNet [99]. The architectures of these encoders are identical except for the number of input channels in their first layers. Since ResNet is originally designed for 3-channel RGB images, we modify the convolutional layer in the initial block of the thermal encoder to accept a single input channel.

We conducted two experiments to transmit thermal information to the RGB encoder. In the first experiment, we fuse the RGB and thermal information in the RGB encoder by element-wise addition of the corresponding RGB and thermal feature maps after the initial block (a convolutional layer, a batch normalization layer, and a ReLU activation layer) and after each residual layer. The addition follows a weighted addition pattern, where the weights for each modality are learned by the attention mechanism mentioned in the next section. The output of the last fusion layer is taken as input for the LSTM module. In the second experiment, we keep the RGB and thermal encoders independent without cross-modal connections. The results show that allowing the model to learn each modality independently is more effective than applying cross-modal connections, regardless of the attention mechanism applied.

Table 3.4: Performance metrics of experiments on cross-modal connections.

	F1-Score	TFLOPs
With cross-modal connections	82.2 ± 3.9	0.37
Without cross-modal connections	82.7 ± 1.5	0.37

Note: Both experiments use a fully connected layer for the attention mechanism and employ early attention.

Attention Mechanism

To assign different weights to the RGB and thermal representations, we utilize attention mechanisms to determine the importance of each modality. We expect it to dynamically change with the input data instead of static weights. We experiment with two attention mechanisms to achieve this objective.

In the first approach, we do not apply any attention mechanism. In the second approach, after obtaining the RGB and thermal outputs from the residual layers, we concatenate these outputs and pass them through a fully connected layer followed by a sigmoid activation function. This process learns two scalar values, representing the weights for the RGB and thermal modalities. In the third approach, we replace the fully connected layer with a self-attention layer—specifically, a multi-head attention mechanism with 8 heads and an attention dimension of 256. Beyond learning to weigh the modalities differently, the self-attention mechanism captures context-dependent relationships between the input modalities. Our results indicate that employing self-attention leads to a 2% increase compared with a fully connected attention mechanism, and a 5% increase compared with no attention mechanism, in the LOPO (Leave-One-Participant-Out) test.

Table 3.5: Performance metrics of experiments on attention mechanism.

Attention Mechanism	F1-Score	TFLOPs
No attention	79.5 ± 4.2	0.37
Fully connected layer	82.7 ± 1.5	0.37
Self-attention	84.7 ± 2.3	0.37

Note: All attention-related experiments employ early attention without cross-modal connections.

Early Weighting vs. Late Weighting

During the image encoding phase, we experimented with calculating weights using embeddings from the first layer compared to those from the fourth residual layer. The results show that the early weighting mechanism outperforms the late one, achieving an F1-score of 82.2% compared to 70.6%. This indicates that the weighting between RGB and thermal modalities relies more on lower-level features, such as texture and illumination than on higher-level, abstract features, such as scenes. Meanwhile, as the depth increases, the number of feature channels increases, and the representation becomes larger. To learn the weighting at these deeper layers, we need to learn more weights to capture the relationships between representations, which results in higher computational cost, with TFLOPs increasing from 0.37 to 0.58.

Table 3.6: Performance metrics with early and late attention experiments.

	F1-Score	TFLOPs
Early weighting	82.2 ± 3.9	0.37
Late weighting	70.6 ± 4.5	0.58

Note: Both experiments use a fully connected layer for the weighting mechanism with cross-modal connections.

3.4 Summary

Our findings indicate that decision fusion is the most effective method for recognizing eating, smoking, and non-gesture activities. However, its computational expense, requir-

ing two separate models, can pose challenges when designing real-time pipelines. In contrast, RTFVNet provides a relatively efficient solution with comparable performance using a single model. In future work, we plan to further optimize the mid-fusion approach for real-time use cases in health-risk behavior detection. We will also expand the dataset and explore advanced multimodal pretraining approaches [101], to improve both performance and robustness.

In addressing RQ2, we demonstrated that multimodal RGB-thermal fusion effectively distinguishes eating, smoking, and non-gesture activities, even under challenging real-world conditions. Our systematic evaluation revealed a critical accuracy-efficiency trade-off: decision fusion achieves the highest recognition performance but requires dual-model computation, while RTFVNet offers a practical single-model alternative with only about a 3% reduction in performance. These findings provide actionable guidance for deploying multimodal wearable systems in clinical settings, where computational constraints often dictate fusion strategy selection.

Chapter 4

Mitigating Privacy Concerns

4.1 Introduction and Background

Research Question (RQ3). *How do we best capture the wearer’s activity of interest while reducing both wearer and bystander privacy concerns?*

Guided by RQ3, this chapter evaluates design choices to reduce user and bystander privacy risk while preserving analytic utility for detecting eating, smoking, and related behaviors.

Embedded wearable cameras are becoming more ubiquitous among researchers interested in studying human behavior in real-world settings, allowing them to visually confirm problematic health behaviors. Wearable cameras provide accurate and impartial data on the normal daily activities of the user (i.e., wearer) and their interactions with the environment [6], [102], [103]. Prominent examples of this in the research community include SenseCam [103] for lifelogging and ActiSight [12] for detection of hand-related activities (e.g., smoking, eating, and drinking). One of the major challenges in the study of human behavior is the capture of authentic or naturally occurring behavior, and given

the known privacy concerns associated with wearable cameras [7], [104], [105], [106], [107], researchers have attempted to design methods that reduce privacy concerns while maintaining their utility for visually confirming human behavior. One type of camera that has generated concerns is the egocentric camera that captures first-person views. Alharbi et al. [7] set out to study cameras on different body locations with different lens orientations to provide device recording affordances that reduce surveillance and social discomfort compared with egocentric cameras. These are referred to as activity-oriented cameras (AOCs) because they are designed to capture a particular activity rather than the field of view of the wearer. To capture hand-related activities of the wearer (e.g., eating, smoking, drinking), Alharbi et al. showed through multiple studies [6], [7], [12] that the use of a wearable video camera in a necklace form factor, with the lens oriented upward toward the mouth, combined with the use of a thermal sensor that aids the obfuscation of RGB video by distinguishing the foreground pixels (i.e., the wearer and activity of interest) and applies obfuscation methods to the background pixels, significantly increases people's willingness to use wearable cameras. AOCs have thus been shown to be highly promising for the capture of hand-related behaviors; however, the obfuscation methods tested are quite primitive, including applying complete masks (the most extreme obfuscation method in which background pixels are set to a single color, such as black), blurring (which retains some information about the colors and shapes of the objects), or canny edge (which maintains the shape of the objects by only displaying an outline). Despite these advances, the partial nature of obfuscation makes the wearer's face identifiable, and participants still report privacy concerns for bystanders that are recorded without consent.

Researchers evaluated blur and edge filters on footage from activity-oriented wearable cameras and found the blur filter to be most preferred by the wearers [6]. However,

several references [108], [109], [110], [111], [112] have shown that using cartoons/avatars for obfuscation is more privacy preserving over image-processing filters such as blurring, canny edge detection, and pixelation. Cartoon-based obfuscation techniques are being used to deidentify faces [109], [111], [113], [114], [115], [116], [117], entire persons [109], [110], and the surrounding environment [108], [111] to preserve user privacy in surveillance videos. Cartooning can be implemented in different ways, from implementing a clever combination of canny edges and box blurs to using advanced generative algorithms [117], [118], [119], [120], [121]. The latest developments in generative deep-learning techniques have led to the creation of high-quality cartoonization models, such as Cartoon-GAN [118] and white-box [119] models that convert real-world footage into anime-style cartoon videos. However, the influence of generative cartooning methods on privacy concerns and user acceptance using AOC systems has not yet been studied.

Videos recorded with an AOC contain three important elements that can pose privacy concerns to the wearer and bystanders. These include (1) the wearer’s face, which roughly occupies one-third of the frame [6], [7]; (2) bystanders captured in the field of view of the camera; and (3) the surrounding environment of the wearer. In order to maximize privacy, we tested a combination of cartooning techniques that address privacy issues across all three elements of the scene. Our approach uses an avatar face to obfuscate the wearer’s and bystanders’ faces; the bystanders’ bodies and surrounding environment are obfuscated using a white-box Cartoon-GAN [119] for an enhanced cartoonized effect. In addition, this method also helps obfuscate information on screens, phones, or other sensitive personal objects in the user’s environment.

In this paper, we aim to show the benefits of cartoonization of video in AOCs in reducing privacy concerns. We set out to answer the following questions: (R1) How does

cartoon-based obfuscation of activity-oriented video footage affect the wearer’s and bystanders’ privacy concerns and visual confirmation utility? (R2) Do wearers prefer cartoonization using an avatar over other obfuscation methods in AOC video footage? The following contributions in this paper enable us to answer the above research questions: (1) a unique cartooning approach to reduce privacy concerns of the wearer and bystanders in the field of view; (2) a Qualtrics survey with 60 participants to measure the privacy concerns of participants deployed using Prolific (a higher standard of online research); and (3) an in-person study with 49 participants for an entire day, where participants use the device and perform normal daily activities, followed by 1:1 semi-structured interviews with the participants to understand their privacy concerns.

Our results show that our cartoon-based obfuscation approach of AOC video significantly reduces privacy concern levels, particularly for bystanders in high-risk scenarios, all while maintaining the ability to visually confirm context. By increasing acceptance of wearable video cameras, our research can drive greater adoption of AOC and ultimately improve understanding of naturally occurring problematic human behavior in real-world settings.

4.2 Related Works

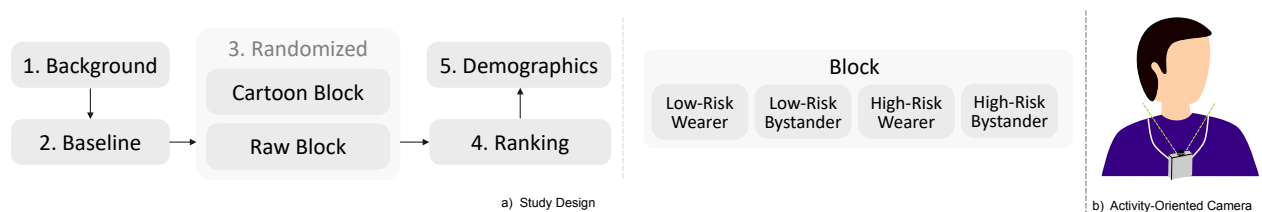


Figure 4.1: (a) Study design for the survey. (b) Illustration of a person wearing a necklace form-factor activity-oriented camera with lens oriented upward. The dashed lines represent the field of view of the camera.

Prior wearable cameras, such as Sensecam, were used as memory recall tools [102]

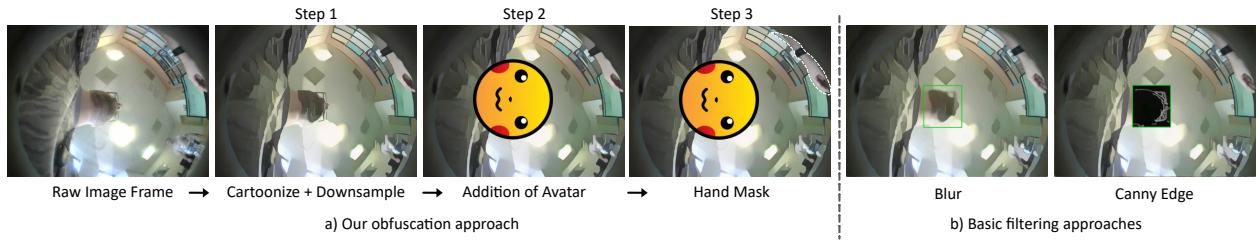


Figure 4.2: Obfuscation Pipeline: The process of generating a cartoonized obfuscated frame from a raw frame.

and as behavioral monitoring tools to capture eating patterns [103]. However, privacy concerns of Sensecam users and bystanders remained unclear. To test the feasibility of using privacy filters for embedded cameras, Erdelyi et al. [111] developed a video cartooning approach combining blurring and edge detection. Their assessment pointed to the balance between privacy and utility compared with other approaches, such as blurring and pixelation. Similarly, Winkler et al. [122] proposed a mean shift filtering method to preserve privacy by globally cartoonizing every frame. However, the impact on user acceptance of the video was unknown, as was the utility of the video footage in maintaining visual confirmation of the wearer’s activity. More recently, researchers have evaluated people’s perceptions of different wearable camera affordances and coined a new, activity-oriented camera (AOC) [7]. To reduce privacy concerns, the findings suggested wearable cameras should orient the lens toward the activity of interest [7] while using thermal data to identify the foreground pixels (i.e., the wearer and objects in close proximity) and obfuscate all other background pixels (partial obfuscation) [6]. Through a subsequent study, the authors concluded that AOCs using blur and edge filters achieved a balance between privacy and visual confirmation of the wearer’s activity [6]. However, filters such as blur and edge allow for identifying features such as the texture and color of clothes and personal belongings of both the wearer and bystander [109]. Avatars and generative methods, on the other hand, have been shown to more successfully obfuscate the

wearer and bystander while maintaining the ability to visually confirm context [110]. This motivates our first research question regarding assessing whether cartoon-based obfuscation of activity-oriented video footage affects wearers' and bystanders' privacy concerns and/or visual confirmation utility.

Based on data collected from an AOC, video can be split and obfuscation applied to each of three segments in a scene: (1) the wearer's face and hands; (2) bystanders; and (3) the environment (everything else in the scene). Researchers have proposed several techniques for face obfuscation, for example, convolution-based filtering [123], geometry-based face morphing [115], and in-painting [123], as well as deep learning methods such as the use of generative adversarial networks (GANs) to distort faces [124]. GAN-based obfuscation can also be applied to the face and body [109]. Moreover, Hasan et al.'s [125] cartoon obfuscation algorithm for deidentifying sensitive objects and personal details in the environment showed that the cartoon approach strikes an acceptable balance between understanding the video's context and protecting privacy. GANs can also be utilized to obscure the wearer's surrounding environment. CartoonGAN [118] is one of the more popular GANs for cartooning real-world videos and has been shown to outperform other cartoonization GANs in quality. Wang et al. [119] improved cartoon quality and overall image quality by using white-box representations (separating out the ability to discriminate between different textures, structures, and surfaces in images to reduce the black-box nature of CartoonGAN). Methods such as blurring and blocking [126], as well as masking, silhouette, avatar, and pixelation, have been tested by Li et al. in their Amazon Mechanical Turk (MTurk) survey [110], where they highlight users' preference to replace people with human avatars to address privacy concerns.

The above-mentioned obfuscation methods were designed to be used on egocentric

video footage, focusing primarily on people present directly in front of the camera, and because the wearer is not in the scene, they fail to assess the effects of obfuscation on the wearer. Moreover, these approaches have not been tested on footage from AOCs, especially in studies involving normal daily activities. This motivates the need for our second research question, namely whether wearers prefer cartoonization using an avatar over other obfuscation methods in AOC video footage. In this paper, we investigate these two research questions and perform thematic analysis to identify themes that optimize the design of cartoon-based obfuscation.

4.3 Methods

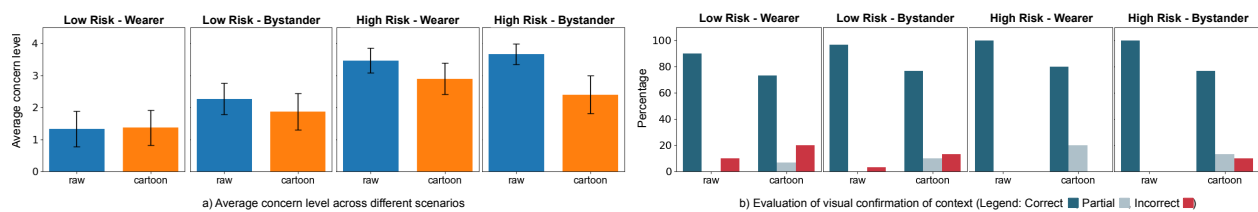


Figure 4.3: (a) The mean (95% confidence interval) concern level of participants in each of the conditions (raw, cartoon) given low-risk or high-risk scenarios with questions focused on either the wearer or bystander. The y-axis represents the Likert scale (0, not concerned at all; 4, extremely concerned) (b) The percentage of participants that got the context (of the video shown to them) correct, partially correct, or incorrect. Each column represents the scenario described.

Video	Statistic (W)	p-value	Raw Median	Cartoon Median
Low-Risk Wearer	453	0.97	1	1
High-Risk Wearer	544	0.11	4	4
Low-Risk By-stander	516.5	0.32	2	2
High-Risk By-stander	646	0.00079	4	3

Table 4.1: U-tests (Wilcoxon rank-sum test). Four hypothesis tests are being conducted, therefore, p-values are set to $\alpha = 0.0125$.

To answer our research questions, through two independent studies, we evaluated

the privacy concerns of wearers and bystanders in AOC obfuscated video footage and the ability to verify their context. Here we describe the first Qualtrics survey with 60 participants and the second in-lab study with 49 participants.

4.3.1 Qualtrics Survey

We conducted a Qualtrics survey [127] to analyze the impact of cartoon-based obfuscation techniques on the wearer’s and bystanders’ privacy concerns and the ability to verify the video context. Participants for the survey were recruited using Prolific’s Platform [128]. The survey was completed by a total of 60 participants, whose ages spanned from 23 to 76 years old (mean [SD], 41.6 [14.0] years). Overall, 49 individuals identified as White, 2 as Black/African American, 6 as Asian, and 3 as Hispanic/Latino; 30 participants identified as female, 29 as male, and 1 as nonbinary. A small number of the participants (8.3%), had worn a camera, such as a Go-Pro, in the past, whereas the majority of the participants had not used any form of wearable camera. At the start of the study, participants were asked about their background and a set of baseline concern-level questions. The participants were randomly assigned to either watch videos processed using our cartoon-based obfuscation approach (cartoon condition) or to watch raw videos (raw condition), as seen in Figure 4.1. Sixty participants were recruited and randomly assigned to either the cartoon or raw video footage condition. If assigned the cartoon block, participants watched four videos: (1) low-risk scenario with the wearer obfuscated; (2) low-risk scenario with the bystander obfuscated; (3) high-risk scenario with the wearer obfuscated; and (4) high-risk scenario with bystander obfuscated. However, if assigned the raw block, participants would see the same four videos without any obfuscation. The order of the four videos within each block was shuffled for all participants to mitigate carry-over effects. After each video, participants were asked to recognize the context of either the wearer or the

bystander and then answered a question asking them how concerned they would be if they were in the wearer's/bystander's position using a Likert scale (see questionnaire in the appendix).

We captured the low-risk and high-risk scenario using two actors wearing a necklace form factor-based AOC (lens oriented upward toward the wearer's face) comprising an RGB sensor equipped with a fisheye lens so that the wearer's face and hand-related activity were visible in the camera's field of view [6]. The low-risk concern scenario comprised a wearer working at a desk using their laptop and a bystander entering the scene to drink water from a water bottle at the wearer's desk. The high-risk concern scenario comprised a wearer using the restroom while a bystander was using an adjacent urinal, after which they proceeded to wash their hands at a wash basin. The wearer's face and hands and the bystanders' faces and bodies were annotated using annotation software.

Our cartoon-based obfuscation method consisted of obfuscating the wearer, the bystander, and the surrounding environment using a three-step process as shown in Figure 4.2. (1) We first globally cartoonized the entire frame using the white-box cartoonization GAN [119] and down-sampled the image. (2) We then superimposed faces in the frame with avatars. Subsequently, (3) we used the labels for the wearer's hand and the raw image to make the hands and objects in hand visible (in the final frame) so that the wearer's activities were not obscured. Our goal was to preserve the wearer's activity context as much as possible while addressing the privacy concerns of both wearers and bystanders. This led to choosing avatars for obfuscating faces to maximize privacy while keeping the wearer's hands unobfuscated to allow for context verification.

In addition, at the end of the study, we showed the participants blurred and canny-edged versions of the videos (Figure 4.2b), where the faces of wearers and bystanders were

blurred/edged instead of blocked using avatars. We then asked them to compare and rank our obfuscation methods with blur and edge filtering based on their willingness to wear the camera, given that the footage was obfuscated using one of these methods.

4.3.2 In-Lab Study

Although effective at providing an initial understanding of privacy concerns, the online Qualtrics survey did not afford participants the opportunity to utilize the device or review their own data. To understand potential differences in participants' privacy concerns when directly interacting with an activity-oriented wearable camera and viewing their own footage, we conducted an in-lab study (n=49) in which participants engaged in their normal daily activities while wearing the AOC in a controlled social environment. Throughout the study, participants were provided with breakfast, lunch, dinner, and snacks; they were also provided with video games and asked to perform the following activities: (1) eat; (2) rest; (3) work on their laptops or phones; (4) use the restroom; (5) play board games; and (6) watch television. Additionally, participants were permitted to leave the laboratory and return between scheduled activities. At the end of the study day, the data from an eating episode (involving hand-related activity) collected during the day was obfuscated using blur, edge, and our cartoon-based approach. Semi-structured interviews were conducted to show participants short video clips of the obfuscated data, along with raw footage for comparison, to solicit their feedback on the effectiveness of each obfuscation method. During the interviews, the questions were aimed at understanding the participants' filter preferences, privacy concerns, and apprehension regarding data sharing in diverse social situations, such as sharing the footage with friends, family, medical professionals, or strangers. Thematic analysis was conducted to assess the participants' obfuscation preferences. The analysis used a predetermined set of codes based on the

researchers' expectations of the participants' responses to the questions posed. In instances where the assigned codes did not adequately capture the context of the response, new codes were created, and themes were generated based on the frequency of different codes.

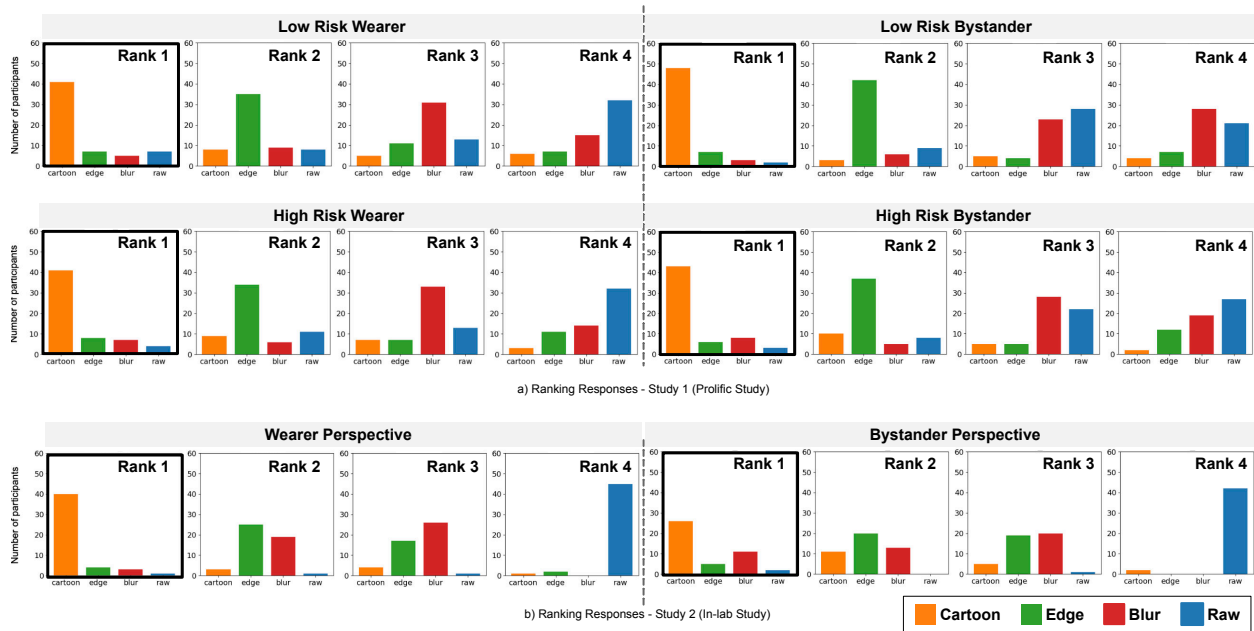


Figure 4.4: Ranking Question for (a) prolific study and (b) in-lab study. The distribution are displayed for participants' rankings for each of the obfuscation techniques in response to the given video scenarios (range, 1–4).

4.4 Analysis and Results

We conduct quantitative analysis on data from our Qualtrics survey, which includes concern levels for raw versus cartoon conditions, context verification, ranking of different obfuscation techniques, and qualitative analysis from themes arising from our interviews with participants in the in-lab study.

4.4.1 Raw Versus Cartoon

We analyzed participants' responses when asked about their concern level if they were in the same situation as either the wearer or the bystander in the video they were watching. Figure 4.3 shows the average concern level of participants for each scenario. The increase in concern from low-risk to high-risk videos is more pronounced for wearers than it is for bystanders. This indicates that if an individual were a wearer, low-risk videos were viewed with low risk, and high-risk videos were viewed with high risk. However, if an individual were a bystander, low-risk videos were viewed with moderate risk, and high-risk videos were viewed with high risk. In the low-risk wearer, high-risk wearer, and low-risk bystander videos, there was no statistically significant difference between the raw and cartoon conditions (Wilcoxon rank-sum test [Mann Whitney U Test]; Table 4.1). However, there was a statistically significant ($p\text{-value} = 0.00079$) difference between raw and cartoon conditions in the high-risk bystander scenario.

4.4.2 Context Verification

We defined key terms for context in the low-risk and high-risk videos to analyze and tally participants' responses when asked about the context of the video they watched. If the participants' responses contained all or some of the defined keywords for the scenario in which they were identifying the context, we classified the responses as correct responses. If the participants missed some of the terms, we classified the responses as partially correct; otherwise, we classified their responses as incorrect. As seen in Figure 4.3b, in the low-risk wearer scenario, 90% of the participants identified context correctly in the raw condition, and about 80% (73.3% correct + 6.7% partially correct) identified the context in the cartoon condition. In the low-risk bystander scenario, 96.7% participants identified the context accurately in the raw condition, and about 86.7% (76.7% + 10%) identified the

context in the cartoon condition. In the high-risk wearer scenario, 100% identified the context in both raw and cartoon conditions. In the high-risk bystander scenario, 100% identified the context accurately in the raw condition and about 90% (76.7% + 13.3%) were able to identify the context in the cartoon condition.

4.4.3 Ranking

When asked to compare the different obfuscation methods at the end of the study, about 72.1% of the participants ranked the cartoon approach with the avatars on faces to be the most preferred over edge, blur, and lastly raw (Figure 4.4a).

Theme	Description
Biometric identifiers	Our obfuscation approach was the most privacy preserving because it curbed recognizable personal traits
Wearer's avatar	The avatar on the wearer's face, although most privacy preserving, obscured certain finer details during certain activities
Bystander avatar	This approach is privacy preserving but can draw unwanted attention at times

Table 4.2: Recurring themes captured during the in-lab semi-structured interviews

4.4.4 In-Lab Results

During the semi-structured interviews, when participants were asked to compare the different obfuscation methods at the end of the in-lab study, a majority (71.7%) of the participants ranked the cartoon approach with the avatars on faces to be the most preferred over edge, blur, and lastly raw (Figure 4.4b). Through thematic analysis of the semi-structured interviews, three main themes were revealed, which are presented in Table 4.2.

4.4.4.1 Biometric identifiers

A total of 53.1% of participants reported that our cartoon-based obfuscation technique obscured their facial features and mitigated identifiable biometric markers and personal

traits. When asked to rank the obfuscation methods based on their preference, participant 38 (P38) stated (referring to our obfuscation method as "cartoon"): *"Cartoon best because none of my features are distinguishable not even the shape of my head. Blur #2 because you can see the shape of my head but can not distinguish any of my features. Edge #3 because you can make out my features, but not recognize me. Raw #4 because you can easily recognize me."* The participant appreciated that our cartoon-based obfuscation approach was successful in occluding several personality traits. P25 stated, *"Cartoon completely occludes any recognizable features; skin color, facial features. Blur only gives very rough contours and blurred colors. Edge gives only contours, but facial angles might be identifiable or informative. Raw shows all information, so it is the least. Maybe if I had a prominent chin, you could recognize it with edge. With blur, I think the only way to get caught is by a birthmark or tattoo or something."*

4.4.4.2 Wearer's avatar

A total of 4.1% of participants reported that our cartoon-based obfuscation approach hindered the identification of finer details during certain activities. P44 stated, *"Cartoon is totally blocked out, you wouldn't know what you are looking at. Edge still has outlines. Those outlines don't show any personal features, and so it's better than a blur for that reason."* Similarly, P47 stated, *"Cartoon is something covering your face completely, so there is nothing to analyze besides your hands or whatever is showing. Edge has no color or anything that could be anyone's face. Blur is still not an accurate showing of the face, so it is still somewhat privacy protecting vs raw."*

4.4.4.3 Bystander avatar

A total of 12.2% of participants noted that cartoon-based obfuscation preserved the privacy of bystanders while being fun; P28 stated, *"[The avatar in cartoon-based obfuscation approach] removes their face, and is funny."* Despite the presence of avatars obfuscating by-

standers, some participants were unable to identify them in the footage (likely due to the low number of pixels representing the bystander). On the contrary, P46 expressed that *"Cartoon draws attention to the person in the background. I wouldn't have even noticed him in the raw, and he blends in in the blur and edge versions."* P45 stated, *"The cartoon draws your attention. Edge is also noticeable compared to blur."* P39 also mentioned that, *"The edge kind of defined the face a little bit. Cartoon draws attention that there is someone there."*

4.5 Discussion

Our results show that our cartoon-based obfuscation approach can alleviate not only wearers' but also bystanders' privacy concerns; for especially high-risk scenarios (e.g., using the restroom), there is a significant improvement in privacy concern level when participants view footage in the cartoon condition compared with raw video footage. During the use of such AOC systems for studying behavior, researchers aim to study the wearer's and bystander's behaviors (e.g., working at a table, eating); therefore, it is essential to maintain the ability to identify the activities even after the obfuscation process. Our results further show that the cartoon-based obfuscation strikes a balance between curbing privacy concerns and maintaining the ability to visually confirm context within the scene. Results from our Qualtrics survey show that despite the use of intense image processing using a GAN for cartoonization and down-sampling the frame thereafter, 90% of the participants assigned to the cartoon block were able to accurately determine the context in the video. In addition, our cartoon-based obfuscation approach was preferred by users not only in the Qualtrics survey but also by participants in the in-lab study. Our findings in the Qualtrics survey also hold true in the in-lab study, where participants donned the device and were able to visually see their own data.

The themes generated from the analysis showed that participants ranked our cartoon-based obfuscation approach the best because they felt it was the most privacy-preserving approach, primarily because identifying features (e.g., facial features, skin color) of the wearer were more difficult to recognize. Although privacy preserving, participants noted that using an avatar in the cartoon-based obfuscation method resulted in complete occlusion of the wearer's face, impeding the ability to identify nuanced behavioral cues, especially in activities such as eating. Based on participants' responses, we conclude that even though using an avatar might be the most privacy preserving for the wearer, determining fine-grained activities might be more difficult. To achieve a balance between visual confirmation and privacy in instances where fine-grained activity recognition such as eating or chewing gestures is important, we could use the edge or blur filter. However, with more advances in artificial intelligence, we could map an avatar's facial features to that of the wearer's face to mimic exact fine-grained facial activities. Participants stated that the addition of the avatar to the bystander's face successfully reduced their wearer and bystander privacy concerns. However, they were drawn to the presence of a bystander without being aware of the bystander's presence had the avatar been absent. Thus, unless this is desired, to alleviate the privacy concerns of bystanders, it may be advisable to refrain from using avatars on individuals positioned that are difficult to identify, as well as those positioned at a considerable distance from the wearer, as this may attract undue attention. Although the GAN and down-sampling process utilized in our obfuscation approach inherently made it challenging to recognize bystanders, the participants' feedback regarding the effectiveness of avatars in obfuscating biometric identifiers warrants consideration. Thus, it may be advantageous to incorporate avatars in situations where bystanders are in closer proximity to the wearer.

4.6 Summary

This paper presents a framework for cartoon-based obfuscation and evaluates its effectiveness in mitigating privacy concerns for both wearers and bystanders in the scene while maintaining the ability to identify video context. A Qualtrics survey and an in-lab study were conducted to gather participants' perceptions of privacy concerns and to evaluate their ability to identify the context in AOC videos. Our findings indicate that our cartoon-based obfuscation approach significantly reduced privacy concerns, especially for bystanders in high-risk scenarios, without impeding the ability to discern the context of the video across scenarios. Although avatars applied to both wearers and bystanders effectively preserved privacy, the use of avatars on bystanders situated at a substantial distance from the wearer could draw unwanted attention. Additionally, using avatars to obfuscate the wearer's face could impede the validation of fine-grained activity. Our future plans include developing techniques to map an avatar's facial features to that of the user to preserve visual confirmation of fine-grained wearer activity. Furthermore, to gauge the effectiveness of these obfuscation methods in real-world settings and to gain a better understanding of naturally occurring behavior, we plan to conduct a free-living longitudinal study wherein participants wear AOCs throughout their day.

These findings directly address RQ3 by demonstrating that effective privacy preservation in wearable sensing requires context-aware obfuscation strategies that balance multiple stakeholder needs. Our cartoon-based approach significantly reduced privacy concerns (particularly for bystanders in high-risk scenarios) while maintaining 90% context verification accuracy, providing a human-centered alternative to traditional blur or edge filters. Most importantly, our thematic analysis revealed that privacy strategies must adapt to situational context—what preserves privacy in one scenario (avatars on distant bystanders)

may inadvertently compromise it in another by drawing unwanted attention. This work establishes that successful privacy-preserving wearable sensing depends not just on technical obfuscation strength, but on understanding the social dynamics of how privacy interventions are perceived in real-world settings.

Chapter 5

Explainability

5.1 Introduction and Background

Research Question (RQ4). *How can AI-driven predictions derived from passively sensed health-risk behaviors be made interpretable, transparent, and actionable for clinicians and end users?*

Guided by RQ4, this chapter translates model outputs into privacy-aware, clinician- and patient-facing explanations (e.g., feature attributions and uncertainty), designs visual analytics aligned with clinical workflows, and evaluates interpretability and actionability with expert feedback.

Obesity-attributable medical costs remain high for healthcare systems and patients and are on the rise [129], [130]. Although there has been an increase in the number of mobile intervention-based weight loss treatment programs [131], [132], [133], effective and economical interventions have not been adequately identified or disseminated. To save costs and improve health outcomes, interventions that can predict nonresponse early during a treatment program enable dynamic modification of intervention components

(ie, stepped care models [134]) to steer the course of treatment toward achieving response or the desired health outcome, analogous to clinical decision-making over time. Clinical decision-making involves making a decision that maximizes what the patient values, and so treatment decisions could be informed by better predictions of what treatments work for whom and when. Machine learning (ML) models help make early predictions; however, building models to predict human behavior early on (eg, if a patient will lose weight in the near future) poses several challenges [135], [136], [137], [138].

A growing body of literature provides evidence for using ML models to understand health behavior [139], [140], [141], [142]; however, 3 critical challenges are associated with using ML models to predict health behavior. The first is that ML models used to represent complex health behavior data are often black boxes that are overly complex; for example, a random forest (RF) model—a classifier known for increasing predictive accuracy even without hyper-parameter tuning, comprising many decision trees where each tree uses a different number of features or variables to determine a classification—is difficult to understand intuitively and, by extension, hard to use in practice. Although RF outperforms several other ML methods in prediction accuracy, it is notoriously hard to interpret [143]. There is an apparent tradeoff between the performance of classifiers (accuracy) and their ability to explain the reasoning behind their results (explainability). Despite this, researchers have been quick to ride the predictability wave of RFs. Several publications have used RF algorithms for critical tasks such as risk prediction [144], [145], estimating energy prediction [146], and early detection of depression [147], yet their clinical utility has yet to be realized, which may be because of lack in using explainability.

Explainability methods developed by computer scientists might help users gain insight into the inner workings of a model. However, it is unknown whether weight man-

agement experts will agree with predictions provided by ML models even when visualization and interactive design components included in these techniques attempt to convey to domain experts the reasoning behind the decisions made by the model [148], [149]. Research has highlighted an over-reliance problem among clinicians in primary care settings, where misuse of a system can occur when placing too much trust in automated systems resulting in user agreement with incorrect system suggestions [150]. However, Jacobs et al recently reported that psychiatrists with higher familiarity with ML were less likely to use an ML recommendation of which antidepressant drug to use compared with clinicians with lower ML familiarity [18]. Over-reliance, however, can be mitigated through explainability metrics. This leads to the second challenge: health behavior is often more difficult to predict, so it is unknown whether there would be a lack of adoption of ML or over-reliance. As a result, it remains not known whether weight management experts are less likely to use ML recommendations even when attempting to increase credibility with explainability metrics.

Traditional ML evaluation metrics such as accuracy and sensitivity provide an overall metric of the model's performance but fall short of providing insights into the reasoning for the model's prediction. Recent developments in the field of explainable artificial intelligence (AI) have seen new explainability metrics or explanations that provide a further understanding of the reasoning behind the decisions made by an ML model. One example of an explainability metric, or explanation, is feature ranking to understand the degree to which features influence a model. Model-agnostic explanations such as SHAP [151], LIME [152], and ANCHORS [153] are currently the standard approaches to explainability. LIME and ANCHORS rely on local explanations that are aimed at explaining the model's reasoning for a given instance (eg, to answer the question: why was this instance classified as "X?"); however, these algorithms are based on approximations by defining

the contribution of a feature to the difference between the actual and mean prediction (ie, not accurate explanations). Although local explanations are specific to understanding a specific instance, when combined with explanations of multiple instances, it could provide an understanding of the model's overall behavior, similar to a global explanation. Global explanations are explanations that describe the overall working of the ML model [154], such as feature importance. For example, a local explanation would be generating explanations for a prediction made by an ML model trained to predict if a patient would lose weight or not. However, a global explanation would be looking at feature importance to understand which features contributed more towards the model's capability of discerning one class from the other. Lundberg et al's SHAP [151] can create global explanations by aggregating the Shapley values (an approach with a solid foundation in game theory) to create feature importance, summary, and dependence plots. Shapley values are feature attributions that act as driving forces, either contributing to the prediction or not. This implies that, unlike LIME, SHAP does not train an interpretable model that can make predictions. The literature has since defined a set of principles for designing explanations.

The use of prime implicants as explanations for weight management experts is supported by current literature describing design principles for human explanations [155]. These principles state that explanations should be “contrastive,” explaining why the model predicted an instance as one class over another. A prime implicate explanation denotes a region that is sufficient to arrive at a given prediction. To arrive at a different prediction, it is necessary (but not sufficient”) to go outside that region. Thus, prime implicants also follow the second principle of being exhaustive.

We designed PRIMO (Prime Implicant Maintenance of Outcome; our interactive explainable software) to translate prime implicants into understandable quantities. The ex-

planation's primality leads it to be parsimonious. The authors additionally emphasize the need to provide guidelines on effectively using interactive explanation tools. Our explanation tool provides users with a step-by-step intuitive approach to generating and evaluating explanations. Our interactive explainability tool, PRIMO, is designed based on these design principles in addition to leveraging methods from SHAP, LIME, and ANCHOR.

The need for exact, accurate explanations that health experts can intuitively visualize and interact with led to our unique approach of extending methods developed by Shih et al and others [156], [157], [158] by compiling an RF classifier into ordered decision graphs, a tractable representation also called ordered binary decision diagrams (OBDDs). We use this tractable representation alongside visualization methods used in prior research to build PRIMO, a tool to generate, visualize, and interact with prime implicant-based explanations. However, the third challenge arises despite prior research among other domain experts; it is unknown whether explainability methods advance trust and understanding of ML models among weight management experts. Our primary objective through this case study is to build a weight loss prediction model, to assess whether providing ML-based predictions and explanations from this model increases weight management experts' agreement with ML model predictions of success or failure, and to conduct evaluation studies with weight management experts to understand the impact of such models on trust and understanding of ML models and ultimately adoption in the real world. This work aims to support the use of explainable ML to support decisions related to changing the course of treatment for someone that has already started treatment based on their initial response.

Explainability Definition Several explainability AI methods have been designed to improve the interpretability of ML models. Kim et al [159] tested an interpretability method

designed for conditional recurrent neural networks to predict weight loss at 16 weeks using features collected across the 16-week study. However, this method does not transfer to RF models and does not focus on early weight loss prediction. Explainability AI involves the communication of ML model results and operations for different audiences and purposes, and our main goal is to study its effectiveness in communicating these explanations to the weight management expert audience. To do this, we define explainability as the ability of weight management experts to trust and understand the explanations presented when related to the problem at hand, predicting weight loss.

5.2 Methods

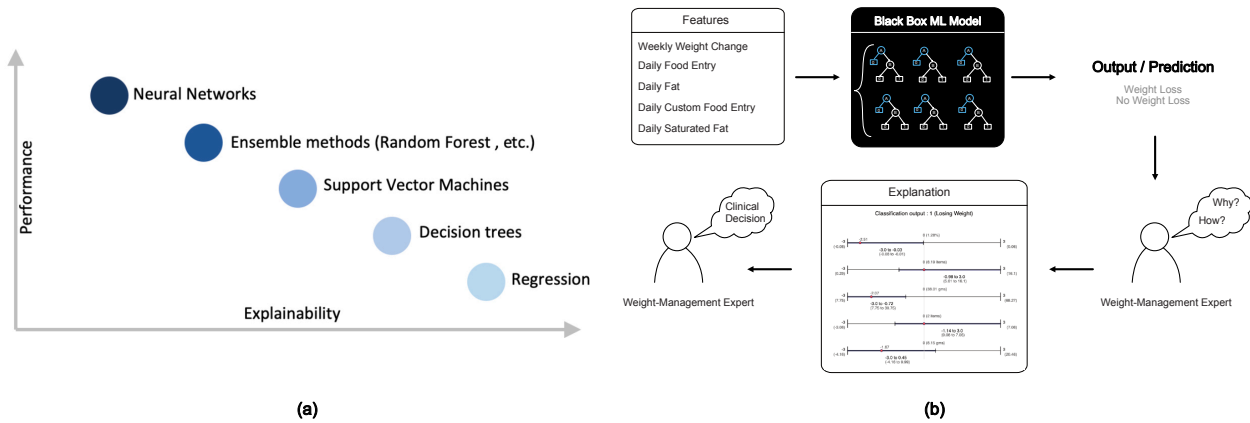


Figure 5.1: (a) Tradeoff between the performance of a model (accuracy) and the explainability. (b) Black box machine learning model.

Figure 5.1 presents our proposed framework in 3 steps: (1) selecting an optimal early prediction time point; (2) generating a tractable decision diagram representation of the RF model using the RF classifier; and (3) building a software tool to enable visualization and interaction with the explanations. We describe each of these steps below, followed by details of our 1-on-1 interviews with domain experts using our software tool, the statistical analysis to assess the agreement of weight management experts with our ML model output, and the questions used to assess trust and understanding of ML models.

5.2.1 Weight Loss Study

The Opt-IN study was a 6-month, theory-guided, technology-supported weight loss intervention to explore factors contributing to significant weight loss. Details of the study have been previously described [160], [161]. The study enrolled adults with body mass index between 25.1 and $39.9\text{kg}/\text{m}^2$. Eligible participants had stable weight (no loss or gain >25 lbs for the past 6 months), were not enrolled in any formal weight loss program or taking weight-reducing medications and were interested in losing weight. Participants obtained their personal physician’s approval to participate, and the physician agreed to receive study reports. An institutional review board approved all study procedures, and all participants provided written informed consent before enrolment. We used demographic information and data collected from participants’ smartphones to determine what factors early in the intervention predicted weight loss at 6 months. We define clinically significant weight loss as $\geq 7\%$ weight reduction from baseline. In the Diabetes Prevention Program, 7% weight loss over 6 months led to a 58% reduction in the development of diabetes [162].

5.2.2 Step 1: Optimal Early Prediction Time Point and Building an RF Model for Early Prediction

To identify the critical early time point for building a machine-learned model, we built several learned models at different time points to select the model–time point pair with the highest predictability. We combined both evidence-based and data-driven practices to guide the process. Evidence-based practice guided our initial selection of features and the subset of time points for early prediction to select from [163], [164], [165], [166], [167], [168], [169]. Data-driven approaches guided the dimensionality reduction through feature

selection and developing a machine-learned model at each time point. We built multiple models using the RF classifier to predict a binary weight loss outcome at the end of 26 weeks. We observed a local optimum in the predictability of models at the end of weeks 2 and 3; we decided to select an earlier time point and therefore selected the end of week 2 as the optimum time point. We used potential predictors to build an RF classifier observing the change over time in data collected from baseline until the end of week 2 to predict a weight loss outcome at the 6-month time point [164], [165].

Table 5.1: Feature Details:

Variable	Description
Age (years)	Baseline age
BMI (kg/m ²)	Baseline BMI
Average weekly food entry	Average number of days per week any food item was entered
Average weekly weight change (%)	Average weight change per week entered by participant
Average daily custom food entry	Average daily number of custom food items created by participant
Average daily food items	Average daily number of food items entered by participant in the app
Average daily fat estimated (grams)	Average daily fat consumed by participant indicated by participant
Average daily saturated fat estimated (grams)	Average daily saturated fat consumed by participant indicated by participant
Average daily physical activity*	Average daily number of physical activity items entered by participant

* this refers to the total number of times participants engaged in moderate-to-vigorous physical activity, as defined by the 2008 Physical Activity Guidelines for Americans.

Based on RF feature importance analysis and prior literature supporting the potential for initial weight loss [163], [164], [165], fat intake [166], [167], and patterns of engagement with mobile health tools to predict health behavior trends [168], [169], we identified a minimal number of highly predictive features (Table 5.1) combining self-reported dietary (initial weight loss, fat intake, saturated fat intake) and engagement variables (entry of food items, entry of custom food items) to build the RF model. This enabled reducing the number of features from 10 features (Table S2) to 5 features, resulting in reduced com-

plexity and further aiding the understanding of generated explanations. After rejecting participants with missing datapoints, we trained the RF classifier (N=419) using these 5 features in the scikit-learn python library and optimized the classifier to have 13 trees, a maximum depth of 3 for each tree, and an accuracy of 81%.

5.2.3 Step 2: Generating a Tractable Representation to Facilitate Computation of Prime Implicant Explanations

The RF model underwent 3 steps (Figure 5.2):

1. Reduction to propositional logic: A RF classifier generally takes continuous variables as input, but each continuous feature becomes a proposition when it appears in any given decision tree. Hence, we can view any RF classifier's input/output behavior as a propositional function. An RF classifier consists of an ensemble of decision trees and classifies an instance by evaluating each decision tree. Note that an RF classifier (and each decision tree in its ensemble) evaluates instances using various tests of the form $x_i \geq t$ or $x_i < t$, both propositions. When we take all the variable tests for a particular feature X_i , they induce a partitioning of the space of X_i into mutually exclusive and exhaustive intervals. In our reduction, we represent each interval as a binary variable in our propositional formula, which can be viewed as a discretization of the original continuous variable.

We define a propositional formula for the RF by defining a propositional variable for each interval appearing in the RF. Given a leaf node in a decision tree, the propositional formula for the leaf node's class label is the conjunction of the decisions found on the path from the root to the leaf. Given a decision tree, the propositional formula for a class label is the disjunction of the paths' formulas to each class label

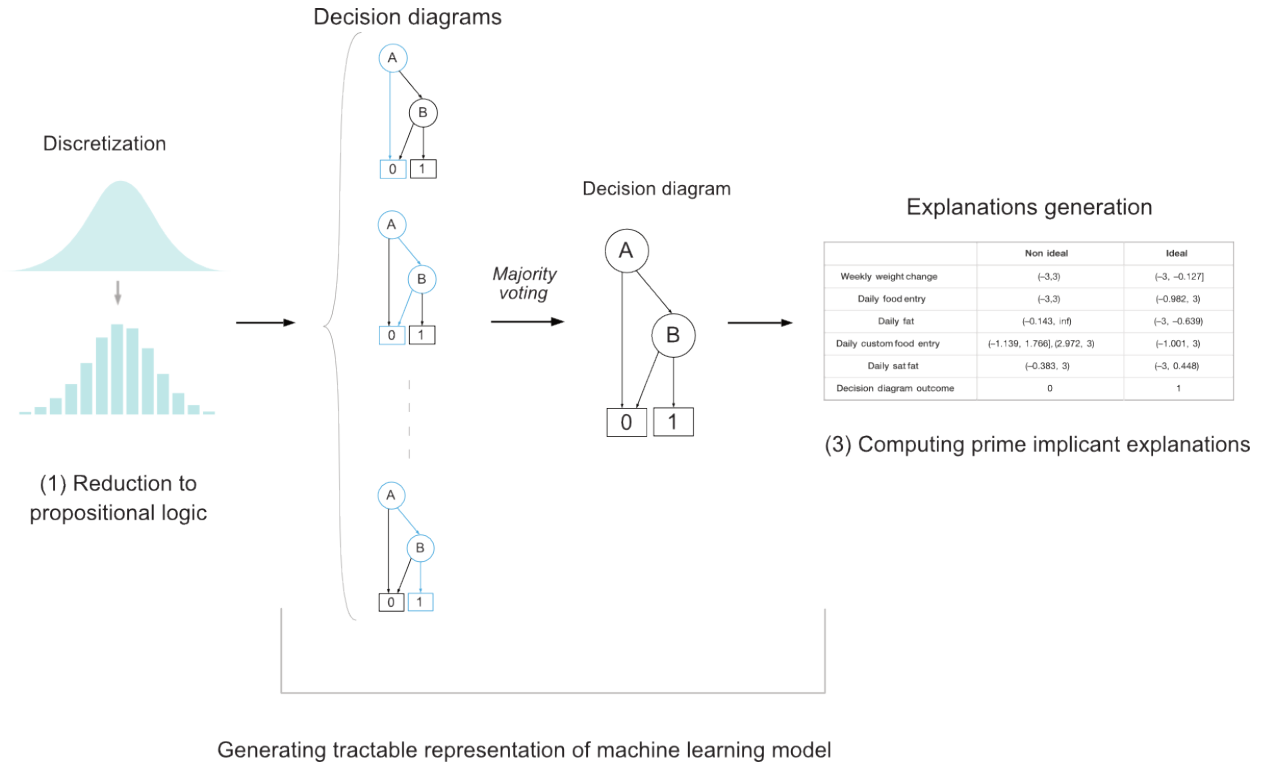


Figure 5.2: Generating tractable representation of machine learning model to compute prime implicant explanations.

leaf. The propositional formula of a random for a class label is found by aggregating the formulas for each decision tree using a majority gate. For a given assignment of features to values, the resulting propositional formula is true if the RF labels the feature vector with the corresponding label.

2. Generating tractable representation (conversion to OBDD): Simply obtaining the propositional formula underlying an RF is not helpful since reasoning with the formula will not be tractable. For example, testing whether there exists a satisfying assignment (ie, testing whether there exists some feature vector that obtains a given label) is an NP-hard (non-deterministic polynomial-time hard) problem. Hence, we appeal to the field of knowledge compilation to obtain a tractable representation of the formula [170]. Knowledge compilation is a subfield of AI that studies, in part,

tractable Boolean circuits and the tradeoffs between succinctness and tractability. That is, by enforcing different properties on the structure of a Boolean circuit, one can obtain greater tractability (the ability to perform specific queries and transformations in polytime) at the possible expense of succinctness (the size of the resulting circuit). We follow Choi et al [158] to compile the propositional formula of an RF into an ordered decision graph, or equivalently, an OBDD. We adapted an RF compiler [171] to compile a propositional formula into an OBDD; many queries of interest become tractable, typically requiring time that is only linear in the size of the resulting OBDD.

3. Computing prime-implicant explanations: To generate explanations, we compiled the discrete RF classifier into an OBDD, which is a tractable function representation and can be used to answer queries and facilitate efficient explanations of classifiers efficiently. Once we have an OBDD representation of our RF classifier, we reason about and generate explanations for the behavior of the classifier [172], [173]. One type of explanation is called a sufficient explanation [174], which corresponds to computing the prime implicants of an RF classifier’s propositional function. A prime implicant of an RF’s propositional formula can be considered a minimal assignment of features to values that will fix the output of the RF classifier. It is a partial feature vector sufficient to fix the classifier’s decision. The behavior of an RF can be wholly described as a disjunction of all its prime implicants (also called its prime cover). Given a propositional formula represented as an OBDD, there are efficient algorithms for computing its prime implicants [175], [176].

For each instance that is input into the classifier, we generated a shortest prime implicant that is compatible with it (ie, the shortest subinstance that is also a prime implicant). As explained by Shih et al [157], one way of verifying the behavior of a

classifier is to verify whether the classifier is compatible with the expectations of a domain expert. A domain expert may define expectations of input-output pairs for a classifier to be reliable. For example, a domain expert may say that anyone who has lost $\geq 1\%$ of their weight and is maintaining low intakes of saturated fat early in treatment is on the trajectory toward clinically significant distal weight loss. To facilitate this understanding of expected behavior, we provide a visual of intervals/ranges for each variable, such that if the value of each variable falls within these ranges, the classifier output will remain the same. Figure 5.3 shows 2 sample visuals of a basic weight loss archetype and a complex weight loss archetype along with their corresponding visuals, which we call PRIMO.



Figure 5.3: In the 2 PRIMO visuals, the red point indicates the patient profile for basic and complex weight loss archetypes. The complex weight loss archetype tends to be difficult because the red points fall at the edges of the PRIMO-generated prime implicant ranges.

5.2.4 Step 3: Designing a Tool for Weight Management Experts

An interactive software tool (Figure 5.4) was designed to enable researchers to query the ML model and generate explanations for hypothetical archetypes of individuals. The user creates custom instances by assigning values to each feature and selecting a “Generate Explanation” button. The software queries the interpretable model to provide a prediction and PRIMO’s visual. PRIMO was designed to be interactive and intuitive by enabling users

to enter values for model features by adjusting sliders that represent standardized values for each feature. The slides show the z-score and corresponding actual value for each feature to create an instance.

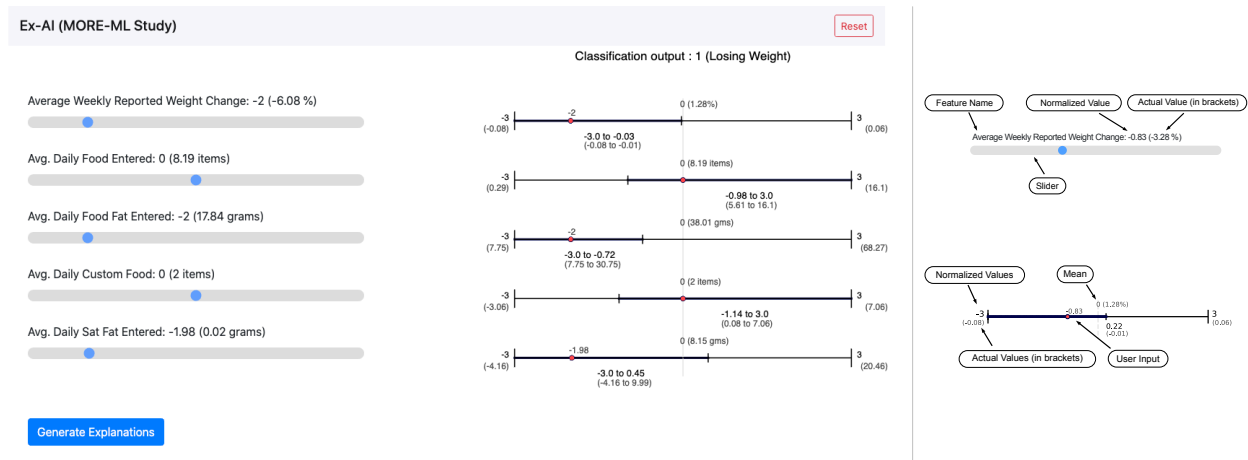


Figure 5.4: PRIMO software interface.

The explanation provided by the model is displayed in the form of highlighted ranges on the interface. If one were to tweak the input values of an instance within the respective ranges specified by the explanation, the model's prediction and explanation are guaranteed not to change. However, on tweaking the input value outside the displayed ranges, the prediction and the explanation may or may not change. This allows users to understand the model's limits that may or may not produce a different output, providing better insight into the model's functionality. In some cases, there is a possibility that, on selecting specific inputs, there is no explanation range for one or more features. This indicates that given the values for all the features with ranges defined, the value of the feature without a range does not affect the output of the predictor.

Table 5.2: Participant Demographics (Interviewees)

Participant	Background	Gender	Familiarity with ML
P01	Clinical Psychology	Female	2 years
P02	Clinical Psychology	Female	-
P03	Clinical Psychology	Female	7 years
P04	Statistics	Male	>15 years
P05	Clinical Psychology	Female	5-6 years
P06	Clinical Psychology	Female	2-3 years
P07	Clinical Psychology (Doctoral Student)	Female	2.5 years
P08	Clinical Psychology	Female	Superficially familiar
P09	Data Science	Male	3-4 years
P10	Clinical Psychology	Male	1-2 years
P11	Clinical Psychology	Female	7 years (collaborated with data scientists)
P12	Clinical Psychology	Female	7 years (started to read about ML, and collaborate)
P13	Nutrition	Female	4 years
P14	Epidemiology	Female	4 years

5.2.5 Evaluation Study With Weight Management Experts

A total of 14 participants (Table 5.2) were recruited (mean [SD] age, 45 [5] years) for the 1-on-1 interviews conducted over an interactive video conference. The length of each interview was approximately between 1 and 1.5 hr. The participants included mobile health researchers in weight loss. Participants were recruited from an obesity research special interest group mailing list and listservs focusing on weight loss research. Participants had backgrounds in 1 or more of the following areas: psychology, nutrition, epidemiology and clinical experience, statistics, and data science. The study's primary goal was to evaluate among end users, the weight management experts, if the explanations are understandable and if they would trust the tool for use in a real-world scenario. Figure 5.5 shows the overall study design.

Participants were asked about their background and experience with analysis metrics and ML. They were presented with details of the weight loss intervention for which the RF model was based. They were then required to provide feedback on 2 other expla-

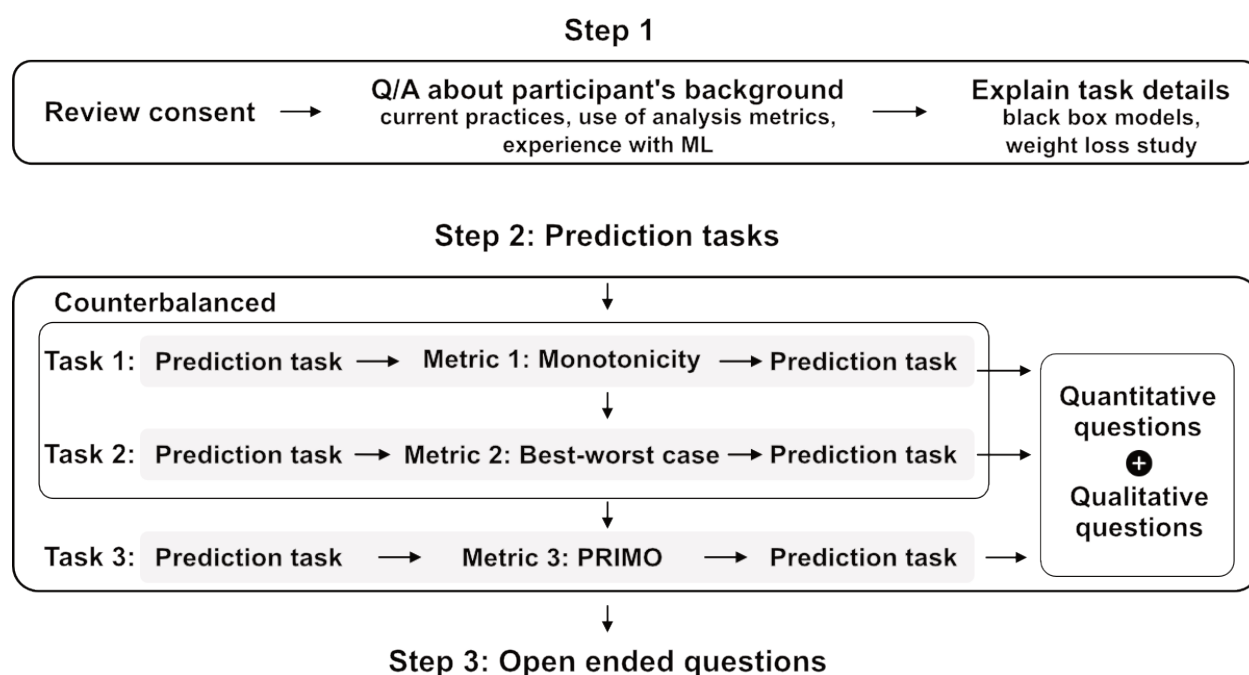


Figure 5.5: Study Design

Monotonicity

if feature increases from -3 to 3	and if prediction goes from	Monotonicity (outcome expected)	Results (How many of the cases were monotonic)
Average Weekly Weight Change	(1 to 0) or constant	decreasing	77.68%
Average Daily Food Items	(0 to 1) or constant	increasing	85.07%
Average Daily Fat Estimated	(1 to 0) or constant	decreasing	74.80%
Average Daily Custom Food Items Generated	(0 to 1) or constant	increasing	82.73%
Average Daily Saturated Fat Estimated	(1 to 0) or constant	decreasing	82.80%

Figure 5.6: Monotonicity visual

nation types besides PRIMO (Figure 5.8). These 2 explanation types were used to help participants get acquainted with the idea of explainability and use explainability tools to understand predictions from an ML classifier. Our explainability framework enabled the

Best/Worst Case Analysis

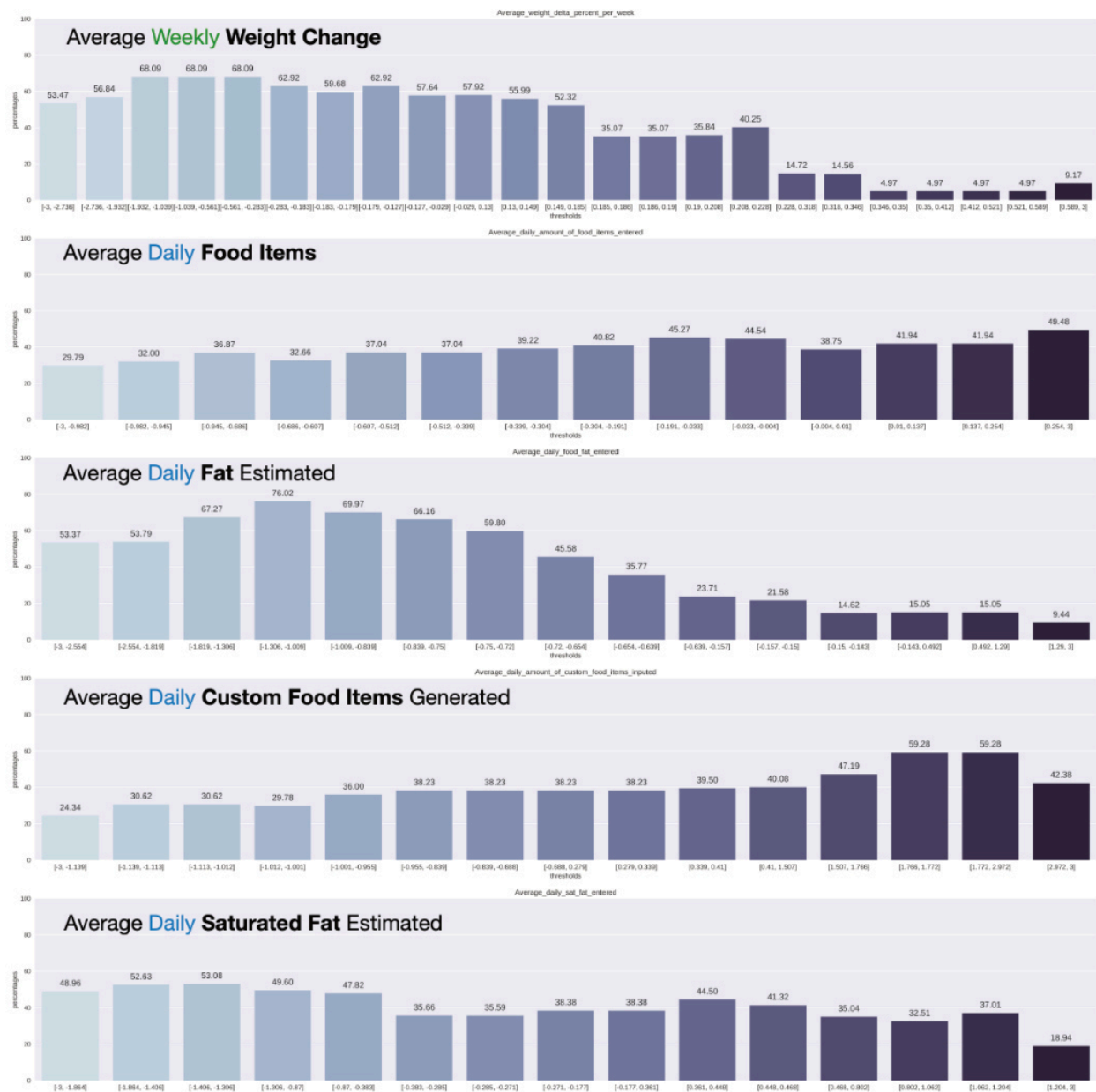


Figure 5.7: Best-worst case visual

generation of these other 2 explanation types: Monotonicity (baseline metric 1) (Figure 5.6), which closely resembles feature ranking, was where we calculated for each feature a percentage that denotes the proportion of instances for which the classifier was either

PRIMO Software

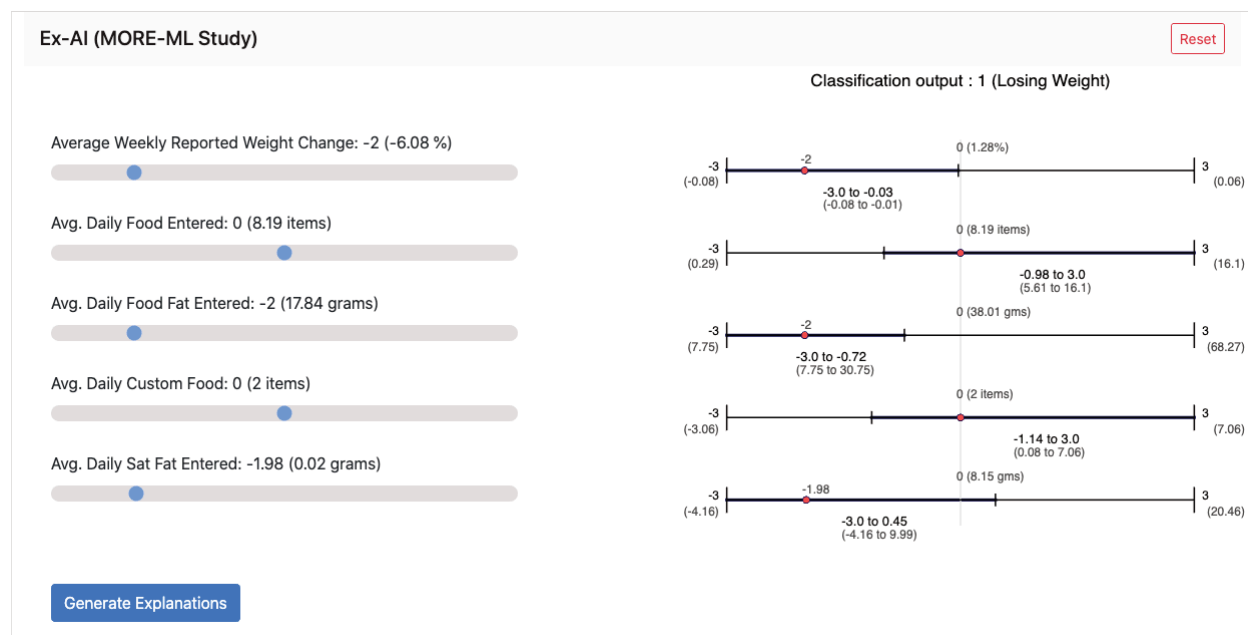


Figure 5.8: PRIMO software UI

monotonically increasing or decreasing. For example, we would expect that the greater the reported weight reduction, the more likely the participant was to lose weight; if this were always the case, it would be 100%. Best-worst case (baseline metric 2) (Figure 5.7), which closely resembled a conditional probability approach, was designed to convey the information gained about the class label of an instance based on observing a particular feature/value pair. In the best-worst case, we compute for a feature variable X , with value x , the proportion of feature vectors classified as either the positive or negative class. The goal of adding these metrics was to get participants to think beyond the existing explainability metrics provided. For each metric, we created 4 hypothetical archetypes of individuals (Figure 5.9 reports z-scores and actual values), 2 basic and 2 complex cases for prediction including: (1) a participant expected to lose sufficient weight at the end of the study (basic weight loss); (2) a participant expected not to lose weight (basic failed weight loss); (3) a participant that was borderline but expected to lose sufficient weight

(complex weight loss); and (4) a participant that was borderline and not expected to lose weight (complex failed weight loss).

Hypothetical Archetypes of Individuals				
	Basic Weight-Loss	Basic Failed Weight-Loss	Complex Weight-Loss	Complex Failed Weight-Loss
Weekly Weight Change	-2.51 (-7.3%)	0.5 (-0.11%)	-0.1 (-1.54%)	-2.05 (-6.2%)
Daily Food Entry	0 (8.19 items)	0 (8.19 items)	1.44 (12 items)	-1.97 (3 items)
Daily Fat	-2.07 (17.1 grams)	0.76 (45.7 grams)	-0.75 (30.45 grams)	-0.11 (36.9 items)
Daily Custom Food Entry	0 (2 items)	0 (2 items)	1.78 (5 items)	-0.12 (1.8 items)
Daily Sat Fat	-1.67 (1.291 grams)	1.79 (15.5 grams)	0.4 (9.79 grams)	1.2 (13.07 grams)

Figure 5.9: Hypothetical archetypes shown to participant types.

We asked participants to predict the outcome of each hypothetical archetype before and after viewing each explanation type.

5.2.6 Ethics Approval

The study was exempt by Northwestern University's institutional review board (approval number: STU00213879) and all weight management experts provided written informed consent before participating in the study.

5.2.7 Statistical Analysis to Assess Likelihood of Agreement with ML Predictions

To evaluate participants' agreement with ML predictions, we used generalized linear mixed-effects models (GLMM). Our outcomes were binary (agreement yes/no), so our models used logit link functions and random participant effects to account for correlation between outcomes within the same participant. The primary analysis evaluated the likeli-

hood of final agreement across predictions/explainability methods. We further examined relationships between explainability methods and 2 participant behaviors: (1) switching initial incorrect predictions to agree with the ML models and (2) switching initially correct predictions to disagree with the ML models. For 1, analyses considered only the initially incorrect predictions by participants, whereas models for 2 included only the initially correct predictions by participants. Models were fit using the lme4 library in the R programming language (version 4.1.2). We fit a model with fixed explainability method effects for each of these analyses and report estimated effects, standard errors (SEs), and 95% confidence intervals (CIs). To examine the overall relationship between explainability methods and outcomes, we also fit a model omitting the explainability method effects and conducted likelihood ratio tests (LRT) for nested models ($\alpha=0.05$).

5.2.8 Advancing Explainability Through Trust and Understanding

To further determine whether PRIMO advanced trust and understanding compared with the other baseline methods, we administered an intelligibility questionnaire and then requested the experts rank each of the 3 explanation types and then probed them to discuss why they preferred 1 method over the other. Quantitative responses were provided using REDCap, and themes were identified based on participants' responses to open-ended questions. The intelligibility questions (Table 1) were evaluated on a 5-point Likert response scale ranging from 0 (not at all), 1 (a little), 2 (somewhat), 3 (quite a lot), and 4 (extremely).

We then calculated and compared the mean (SE) of the Likert response scale values answered by the participants for the 5 quantitative questions shown after each task. At the end of the study, the participants were asked to rank the 3 explanation types and to describe their suggestions regarding improvements and which types enhanced their un-

Table 5.3: Intelligibility questionnaire. The 5 quantitative questions were presented as adapted from Cahour et al [177].

Trust	Trust: How much do you trust the model?
	Reliable: How reliable do you think the explainability metric is?
Understand	Predictable: According to you, how predictable are the outputs of this model?
	Efficient: According to you, how efficient is the explainability metric in describing why the model generated the outputs/predictions?
	Confident: How confident are you in your answers?

derstanding of ML models. We also report on themes regarding participant preference for a specific metric type over another by qualitatively analyzing their responses. Qualitative questions were designed to engage the user in an open-ended response to capture how and why the metric facilitates the prediction of outcomes. The three open-ended questions were presented as follows (with the first 2 adapted from Ribeiro et al [152]):

1. Trust: Would you trust this model guided by this evaluation metric to work well in the real world?
2. Understanding: How do you think the model is able to distinguish between the classes?
3. Understanding: Do you have a better understanding of the model? If yes, why? If not, how would you improve the explainability metric?

5.3 Results

We trained an RF model using 5 engagement and diet-related features, captured during the first 2 weeks of treatment, to predict early on at week 2 if a participant would lose

weight or not at the end of the 6-month weight loss study. The model was trained with 81% accuracy (specificity, 86%; sensitivity, 69%). The 5 features obtained using a smartphone app were: (1) average daily food entry, (2) average daily custom food entry, (3) average weekly weight change, (4) average daily fat, and (5) average daily saturated fat.

5.3.1 Statistical Analysis

GLMM results are shown in Table 2; coefficients and 95% CIs are reported on the odds ratio scale. Across explainability methods, participants' initial predictions were often correct (PRIMO software, 67.9%; monotonicity, 69.6%; best-worst case, 66.1%). Explainability methods were statistically significantly related to the final agreement (LRT $X^2=7.99$, $P=0.018$), and the PRIMO software method appeared to have the strongest final agreement at 89.3% relative to the monotonicity (76.8%) and best-worst case (67.9%) methods. Explainability methods were also statistically significantly related to participants switching initially correct predictions to disagree with the ML models (LRT $X^2=6.85$, $P=0.033$). Participants responding to the PRIMO software method were considerably less likely to change initially correct responses (2.6%) relative to monotonicity (5.1%) or best-worst case (18.9%). Finally, explainability methods were related to participants switching initially incorrect predictions to agree with ML models (LRT $X^2=6.27$, $P=0.043$). The PRIMO software method saw a greater proportion of participants correcting incorrect predictions (72.2%) relative to monotonicity (35.3%) or best-worst case (42.1%). Notably, the between-subject variability was minimal (intraclass correlation coefficient - ICC<0.01) for a final agreement and switching to disagreement but was substantial for switching to incorrect predictions to agree with ML models (estimated between-subject SD: 0.855 on the logit scale; ICC=0.18).

Table 3 displays the results of GLMM fit to only responses on complex items, which

Table 5.4: GLMM Results.

Term	Dependent Variable		
	Final Agreement (n=168 responses)	Switch to Disagree (n=114 responses)	Switch to Agree (n=54 responses)
Intercept	2.11	0.23	0.70
Best, worst case	(1.22, 3.90)	(0.071, 0.501)	(0.23, 2.10)
<i>P</i> value	0.01	0.001	.53
Monotonicity	1.57	0.23	0.75
Best, worst case	(0.68, 3.68)	(0.03, 1.04)	(0.17, 3.29)
<i>P</i> value	0.29	0.08	0.7
Software	3.95	0.12	4.54
Best, worst case	(1.45, 11.8)	(0.006, 0.70)	(0.93, 22.13)
<i>P</i> value	0.009	0.05	0.06

were deemed likely to be difficult to classify. The data exhibited greater prior agreement on complex items in the PRIMO software group (16/28 total trials, 57.1%) than for the Monotonicity (11/28, 39.3%) and Best-Worst Case (11/28, 39.3%) groups. However, Table 3 shows that despite the greater opportunity for participants to switch to disagree, for these items, the odds of participants in the PRIMO software group switching to disagree were 0.05 [95% CI: 0.00, 0.31] relative to the Best-Worst Case group. Similarly, among complex items for which there was prior disagreement, the odds of participants in the PRIMO software group switching incorrect prior responses to agree with ML models were over 4 times that of either of the other groups (OR: relative to the Best-Worst group, 4.1; relative to Monotonicity group, 5.1).

Figure 5.10 shows 5 plots, 1 for each question in the intelligibility questionnaire. Although there was no significant difference in intelligibility between the 3 explainability metrics, the result trended toward using the PRIMO software. Participants were more confident about their answers to the quantitative questions for PRIMO software followed by best-worst case and monotonicity. On average, participants found the PRIMO software

Table 5.5: Comparison between overall agree, switch to disagree and switch to agree groups.

Term	Overall Agree (Complex)	Switch to Disagree (Complex)	Switch to Agreement (Complex)
Intercept	0.714	1.232	0.646
Best, worst case	(0.26, 1.85)	(0.0001, 56,695.11)	(0.14, 2.71)
<i>P</i> value	0.477	0.756	0.518
Monotonicity	1.658	0.168	0.814
Best, worst case	(0.53, 5.42)	(0.00, 1.00)	(0.15, 4.31)
<i>P</i> value	0.389	0.159	0.803
Software	8.056	0.05	4.140
Best, worst case	(2.223, 36.277)	(0.00, 0.31)	(0.705, 35.64)
<i>P</i> value	0.003	0.046	0.143

more trustworthy, predictable, reliable, and efficient.

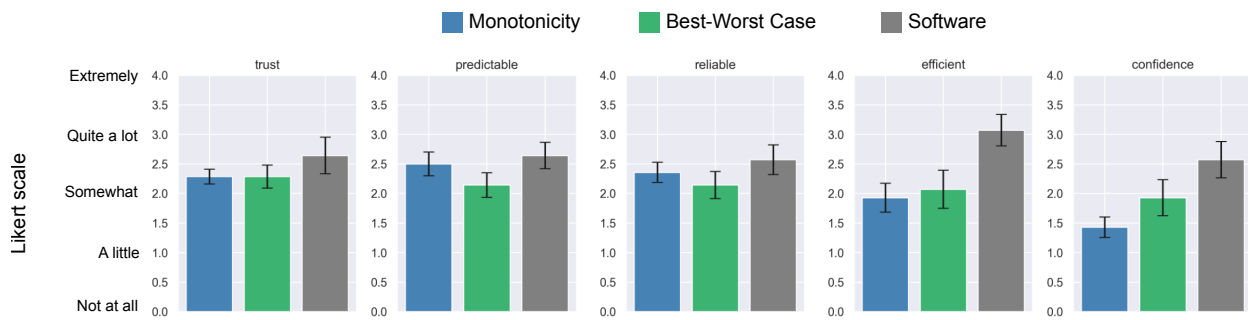


Figure 5.10: Likert scale analysis for intelligibility questionnaire (mean value and standard error)

5.3.2 Qualitative Analysis

Preference for interactive explanations: Participants found the PRIMO software to be most understandable due to its ability to provide an explanation at an instance level. A few participants also hoped to see explanations at the instance level while going over a global understanding of the model. Participant 2 (P2) stated: “Software was the most understandable and allowed for considering multiple individuals, multiple circumstances and really was user friendly even down to the level of a participant.” The ability of the participant to query and generate explanations that are of interest to them also played a

role in their understanding of the model. A similar case was tested by Lakkaraju et al [178] when the authors compared user interest-based subspace explanations with decision lists and decision sets.

Local explanation remains insufficient for answering the question “Why?”: Many participants found the interactive nature of the PRIMO software and its ability to provide explanations to increase their understanding. However, a few participants were looking to reason from a behavioral perspective “why” the interpretable model generates ranges for the explanations the way it does. P2 stated concerning the PRIMO software, “I really like the model. I would just say somewhat predictable. I thought I was setting threshold at 3 weeks that would be likely associated with weight loss and was surprised that the model put the person in the not losing weight category. I think it’s really good, really efficient, allows you to see the impact of your various inputs are.” Having gone through the details of the PRIMO software and using it a bit, P11 stated: “So, if I wanted someone to understand it, I think you know your explanation was really interesting of this one because you said it tells you why, but it does not tell you why. It tells you what ranges are causing you to make that decision, but it’s not telling you why those ranges matter, for like the behavior.” It seems explainability metrics alone are not sufficient in informing the user regarding the precise reasoning behind the outcome.

Missing model performance-related information: Previous literature demonstrates that showing a model’s accuracy on the held-out set can affect people’s perception of the model’s performance and their trust [179]. Therefore, throughout the study, domain experts were not shown any information about the performance of the RF model or the interpretable OBDD-based model, as our goal was to see how the explanation independently affects people’s trust in the ML model. However, participants were curious to know model

performance metrics, especially while interacting with the PRIMO software. P4 felt that “[I] need to see metrics on how the model really performs in some automated fashion rather than me making changes to see. Human/human plus this software makes better predictions—would be the real test of this. In the sense, I can’t answer this because it’s not been evaluated objective.” P5 also agreed stating, “I will still say quite a bit trustworthy. Because I’ll still like to see other overall model performance metrics about, I don’t know if accuracy would be the right one to use, but something like overall model performance would be nice to see as well, but I feel like I understand the model.” These model performance metrics provide an understanding of the overall model performance, along with a global explanation. It seems that local/global explanations alone cannot fulfill domain experts’ need to create mental models that describe the overall behavior of the model. As a result, a combined global and local explanation approach is necessary, reinforcing prior findings in the explainability literature [151], [178].

Features related: Participants were shown results based on the 5 features that were used to train the RF model. These features were also selected based on permutation feature selection methods and guided by experts. However, a few participants, through their clinical/research practice, hoped to see features such as gender or calories. Having gone through the first local explanation type, P1 would have liked to have seen a lot more variables, “I would like a lot of variables; I wouldn’t stick just to these four. I don’t know how much I’d trust it just based on these five features.” P7, based on their practice, felt that “The single most important one is calorie intake.” P6 also hoped to see a demographic feature stating, “What’s interesting is that I wish that the participant’s sex was included as well because usually when I look at weight loss percentages I, evaluate them a little bit different for men vs women just because, in my clinical experience and from a research perspective, men tend to lose more weight than women.” P13 thought it might be helpful

to combine features, stating, “you’re saying that they’re [custom food item] separate from average food items, but what if you just add them together as one total variable.”

Misinterpreting the term model: The term “model” refers to the ML model. In our case, it refers to the RF model that was trained and used to generate the interpretable model. A few participants confused the term model for the explanation type itself. P3 was curious to know how the model weighted factors but intended to understand how the explanation was weighting different features, “I think somewhat [predictable] because I think when it comes to predict the individual participants, it was hard when not everything followed the principle of monotonicity so I could not decide how the model would end up weighting certain factors, so when things were inconsistent I was not sure how that would turn out.” Similarly, when asked if they trust the model guided by the explanation (they referred to trusting the model), P2 while going over the second explanation type, wanted to convey that they trusted the explanation, “Somewhat trust the model, because to a certain extent it predicts and is consistent with what I know from literature. The fact that it seems to discount the completeness of the dietary record that makes me question it a little bit, but I do know it’s important to log fat, so I can see that being potentially important.”

The usefulness of explanations to patients: After having used the PRIMO software and tested it out, P2 felt that it could be beneficial for patients to observe the thresholds determined by the explanations to understand their goal progress, “I can see even participants [patients] messing around with it and seeing what their goal thresholds might need to be. The same limitations in terms of binary categories and what I think is ideal.” Although an exciting hypothesis, we do not know how many patients would find such helpful information. Future research should explore such questions to determine if pro-

viding PRIMO explanations would benefit patients.

Weight loss definition: A few participants were curious to know about the definition of weight loss itself. Specifically, what percentage of weight loss is clinically significant? P13 expressed their views about a fixed threshold for defining weight loss and stated, “For example, if we were to categorize, like what weight change is right, is it, like what’s the difference between someone who has 0.00 vs someone who has -0.01? But then one’s losing weight, and one’s not, and then that kind of like muddies the water in a way, so is it like that those who are in the middle of transitioning to something.” Several definitions of clinically significant weight loss vary based on the clinical outcome of interest. For example, weight loss as low as 2% has shown promise for preventing diabetes, whereas weight loss of 7% to 10% may be warranted to see changes in physical functioning.

5.4 Discussion

We have demonstrated the feasibility of building a model that predicts weight loss as early as 2 weeks into an intervention using 5 engagement and diet-related features obtained from a smartphone app: (1) average daily food entry, (2) average daily custom food entry, (3) average weekly weight change, (4) average daily fat, and (5) average daily saturated fat. The model used yielded an 81% accuracy (specificity, 86%; sensitivity, 69%). Recent papers have demonstrated the predictive value of initial self-reported weight loss on long-term weight loss, which our model validates; however, most of these models achieve low specificity (eg, 50.5%–53.6% [180]), which demonstrates a challenge in predicting nonresponders. Predicting nonresponse early is essential to be able to change the course of treatment. We suggest future predictive models assess engagement (eg, number of times participants log food and create custom food items) and self-reported fat intake (total fat

and saturated fat) to advance their prediction of nonresponders. Although recent findings of engagement with smartphone apps (eg, interaction length and density of usage [181]) have been shown to be predictive of positive health behavior trends, our findings support the use of specific features, such as daily entries of food logs and frequency of creation of new food items, as important predictors of weight loss in the context of weight loss treatment programs. Prior literature has shown support for fat intake positively predicting change in body fat [182], and our findings show that self-report based on food logging may be an important feature in early prediction of nonresponse. This model is a first step toward saving intervention costs and/or improving health outcomes using an effective stepped-care model. This saves intervention costs by predicting responders and potentially stepping down care and improves health outcomes by predicting nonresponders and stepping up care. Current studies, such as the SMART study [183], are beginning to test stepped-care approaches; however, they lack validated weight loss prediction models. Future work should further validate this model in a different study to assess its generalizability. We plan to analyze data from a completed study and assess if RF models help predict weight loss. We also plan to conduct a clinical trial in the future to study the predictive power of the 5 predictors we selected in addition to day-to-day variability [184], macronutrients, and the feasibility of using RF models and PRIMO system as a decision support tool for stepped-care treatment.

We have also addressed one of the critical challenges of representing complex models using an explainability tool, PRIMO, an interactive tool that uses prime implicants as an explainability metric. Out of 14 weight management experts, 10 ranked PRIMO software first in advancing their understanding of ML models, and 3 ranked it second. Some participants felt the PRIMO software was more reliable and user-friendly and allowed for greater engagement to consider different types of people. However, there was agreement

that the PRIMO software system could still be improved by combining global explanations such as the accuracy of the model. Designing a visualization tool that combines interactive instance-based explanations with interactive global explanations seems to be desired by weight management experts. Participants can then query the system to show overall model performance (similar to the other 2 explainability metrics). Another key challenge is whether weight management experts are more or less likely to use ML recommendations when provided with explainability metrics. In basic predictive scenarios, ML with explainability may be more powerful than other tools in encouraging weight management experts to switch to agree (when they make simple errors [185], [186]). Although still powerful in switching to an agreement, in complex scenarios, ML with explainability is not more powerful than other simpler methods when switching a weight management expert's mind when they disagree with the algorithm. However, it is more powerful than other more basic tools when they agree with the algorithm to prevent them from changing their mind (ie, switching to disagree), acting as a great confirmatory tool. When looking at switching to agree, the PRIMO software tool had high numbers in both the basic and complex scenarios, and it showed in complex scenarios to be significantly more likely to agree after viewing the PRIMO software. Future work should further investigate the effect of ML with explainability tools when the predictions are considered less controversial (basic) compared with complex scenarios.

Despite the increased potential for the adoption of ML models with explainability metrics, it is unclear whether an interactive PRIMO software that uses ML models with explainability would significantly advance trust compared to other methods. We do notice that PRIMO software, on average, yields the greatest trust (mean, 2.64 vs 2.29 for monotonicity and 2.29 for best-worst case); however, it is not significantly different from other methods. Most participants liked the PRIMO software owing to its interactive nature and

felt it was more transparent and “trustworthy” than the other 2 methods. However, many still felt that, although they could test different hypothetical individuals and see the impact of these changes on model output, they still did not fully understand how the PRIMO software was generating these predictions and windows of uncertainty. They liked that they could see the combination of the different features and what they were likely to predict; however, it seems that with any local explanation, people still wanted to see overall metrics (eg, accuracy, specificity/sensitivity).

Despite following the principles of designing human explanations, there remain challenges in understanding these models. Our findings support that the understanding of ML models is improved through explainability metrics. However, our interactive PRIMO software method did not show significant improvement over other explainability methods, especially in efficiency, predictability, and confidence. In explainability metrics, we must recognize that we are still only learning a functional relationship between a set of features (ie, consumption of fat or entry of a food item into the app) and a class label (yes/no to weight loss after 6 months of intervention), uncovering the patterns the RF finds in the data. However, as 1 participant mentioned, to truly uncover understanding, we must answer the “Why?” which requires building models of understanding the causal mechanisms (including the latent ones) as discussed by Pearl and Mackenzie [187] and Darwiche [188]. Although researchers often refer to the output of a classifier trained on data as a learned model, in reality, it is a learned function. It has become increasingly apparent that the ability to answer the “Why?” question depends on whether one is learning a function or a (causal) model. For example, we learn a functional relationship between a feature vector and a class label by learning a classifier using RF from data. Hence, a local explanation may only provide insight into what sort of pattern an RF found in the data. In contrast, with a model, we try to represent the underlying causal mech-

anisms of a system (including latent ones), so we may be able to answer questions such as “Why does consuming more fat increase the odds of gaining weight?” In other words, although local explanations may help increase the trust that one has in a classifier that was learned from data, the ability to answer “Why?” questions may require that we go beyond functions and classifiers and seek to learn and develop causal models instead [187], [188]. Our findings show promise in PRIMO gaining the trust of weight management experts; however, the lack of significance compared with other methods may be attributed to our relatively small sample size ($N=14$) and to the fact that defining trust in ML models or functions remains elusive.

5.5 Summary

We performed a qualitative analysis of the one-on-one guided interviews with weight management experts to study if an interactive explainability tool can enhance trust and understanding of the behavior of the original model, thereby increasing adoption. Our findings show that domain experts showed a significant increase in agreement using our interactive explainability tool compared with other methods, especially when it comes to explaining basic weight loss archetypes (people that are easier to predict). This shows the potential for interactive explainability tools to catch errors or misjudgments in weight management experts’ initial predictions. Moreover, we show that participants are significantly less likely to change their response to disagree when their initial predictions are aligned with the interactive prediction tool [189]. In general, when shown local explanations of specific hypothetical archetypes, participants felt the need to understand the model’s behavior at a global level. Alternatively, when shown global explanations, participants felt the need to understand how specific instances would be classified and explained by the model. Therefore, augmenting our interactive explainability tool to in-

clude a visual representation combining local and global explanations could further enhance agreement with ML models. Additionally, model performance metrics such as accuracy, sensitivity, and specificity are also needed to enable domain experts to gauge their reliability on the ML model's decisions to avoid overreliance. This implies that explanations are required at every step in the framework, from visualizing and understanding trends in raw data to using the model for making predictions. Through our findings, we also learn that domain experts often differ, when compared with data scientists, in the use of the term “model” where data scientists often conflate learned models with learned functions. Aside from a more refined use of terminology, expectations and limitations need to be set in terms of what can be gained from a learned function compared with a learned model. Future studies could include testing a more comprehensive explanation interface with aspects of local and global explanations. Weight management interventions may benefit from such explanations by testing the delivery of just-in-time adaptive interventions based on the explanation ranges generated by the PRIMO software to change the course of treatment. The just-in-time adaptive intervention could explain to the patients in the treatment program why the system thinks they need to stay or change a course of treatment, potentially increasing users' trust in the system. This work provides a foundation to aid in translating models in communicating with other disciplines such that use in research contexts is more plausible. To the extent that researchers can trust, use, and explain models to guide decisions, these could be used in research protocols to guide testable treatment decisions based on ML rather than less-complex or less-accurate correlational or clinical intuition methods. Future work should also test the effectiveness of these explanations among patients and clinicians. We plan to provide PRIMO as an open-source software tool for the community, so that not only weight management researchers but also computer science researchers can build on this framework to develop more rigorous tools for clinical decision support.

These themes—the need for multiple explanation types, uncertainty visualization, and performance metrics—directly informed how RQ4 was addressed. Our work demonstrated that making AI-driven behavioral predictions truly interpretable and actionable required more than just accuracy; it demanded interactive visualization tools that translated complex model decisions into clinically meaningful explanations. While PRIMO achieved 81% early weight loss prediction accuracy and significantly increased clinician agreement (89.3% final agreement vs. 67.9% for traditional methods, $p < 0.01$), the key insight was that interpretability emerged through transforming random forest decisions into prime implicant explanations—visual ranges showing exactly which behavioral thresholds (e.g., fat intake, food logging frequency) would maintain or change predictions. Our qualitative findings revealed that transparency required layering these instance-level explanations with global performance metrics and behavioral rationales that connected statistical patterns to clinical knowledge. Most critically, actionability emerged through interactivity, allowing clinicians to test hypothetical scenarios and explore decision boundaries dynamically. This work established that effective clinical AI systems must move beyond static accuracy metrics to provide exploratory interfaces that respect domain expertise while making algorithmic reasoning both transparent and clinically relevant for real-world deployment.

These findings directly address RQ4 by demonstrating that explainability is not merely an auxiliary feature of machine learning systems, but a prerequisite for their effective adoption in clinical decision-making. Through PRIMO, we showed that translating complex model logic into structured, instance-level explanations grounded in prime implicants enables experts to better understand, evaluate, and trust early predictive models. Compared to feature-ranking and probability-based baselines, PRIMO significantly increased expert agreement with model predictions while reducing inappropriate switches from

correct judgments, highlighting its role as both a corrective and confirmatory decision support tool. Importantly, our qualitative analysis revealed that clinicians value explanations that expose model stability regions and uncertainty boundaries, rather than opaque importance scores alone. This work establishes that successful deployment of AI-driven decision support systems depends not only on predictive accuracy, but on aligning model explanations with expert mental models—bridging the gap between algorithmic inference and actionable clinical insight.

Chapter 6

Discussion

6.1 Synthesis of Findings

This dissertation demonstrated that effective behavioral health monitoring requires careful integration of technical innovation with human-centered design principles throughout the sensing-to-insight pipeline. Across our four research questions, several key themes emerged that advance our understanding of unobtrusive health sensing.

First, our work challenges the long-standing assumption that collecting rich behavioral data necessarily compromises privacy, showing instead that both goals can be achieved through privacy-aware sensing and representation. Through HabitSense, we showed that multimodal sensing systems could capture clinically meaningful data while maintaining user comfort and social acceptability. The key was not avoiding cameras or rich sensors, but rather designing systems with appropriate form factors, lens orientations, and clear value propositions. With over 80% of participants reporting no discomfort and 7.2 hours average daily wear times, we demonstrated that users will adopt potentially invasive sensing when benefits are transparent and privacy controls are evident.

Second, our findings on multimodal fusion (RTFVNet) established that behavioral

recognition benefits more from complementary modality integration than from increasing single-modality sophistication. The 84.8% F1-score achieved through RGB-thermal fusion, while requiring half the computational resources of decision fusion, suggests that future sensing systems should prioritize diverse, lightweight sensors over complex unimodal approaches. This finding has important implications for system design, particularly for resource-constrained mobile and wearable platforms.

Third, our privacy preservation work revealed that acceptability is not merely a technical problem but fundamentally a social one. The 72% preference for cartoon-based obfuscation was not solely due to its privacy-preserving properties, but because it balanced multiple stakeholder concerns—from wearer privacy to bystander anonymity to clinical utility. Critically, we found that privacy strategies must be context-adaptive; techniques that preserved privacy in high-risk scenarios could paradoxically compromise it in others by drawing unwanted attention.

Finally, PRIMO's success in increasing expert-model alignment to 89.3% demonstrated that clinical adoption of AI depends not on perfect accuracy but on interpretable reasoning that respects professional expertise. Weight management experts required multiple layers of explanation—from instance-level predictions to global performance metrics—delivered through interactive interfaces that allowed hypothesis testing. This finding challenges the common assumption that model accuracy is the primary barrier to clinical AI adoption.

6.2 Practical Implications

From a translational perspective, this dissertation provides concrete design guidelines for deploying behavioral sensing in real-world health interventions. For system design-

ers, we established that multimodal sensing platforms should prioritize complementary modalities over redundant ones, implement privacy preservation at the hardware and software levels, and provide clear feedback about system state and data use. The success of activity-oriented camera placement in reducing privacy concerns while maintaining gesture recognition accuracy offers a template for future wearable camera systems.

For intervention developers, our work demonstrates that continuous behavioral monitoring can achieve the adherence necessary for longitudinal studies and interventions. The key is that providing users with access to their data, control over privacy settings, and clear explanations of how data informs care recommendations can reframe perceived invasiveness into a sense of empowerment and support.

For clinical stakeholders, PRIMO illustrates that AI systems gain acceptance through transparency and interactivity rather than automation. Clinicians valued the ability to test scenarios and understand decision boundaries more than having decisions made for them. This suggests that clinical AI should be positioned as decision support rather than decision replacement, with interfaces that make reasoning transparent and contestable.

6.3 Limitations

Our behavioral detection focused primarily on eating and smoking—two behaviors with similar gestural patterns but distinct contextual cues. Extending these methods to other health behaviors such as medication adherence, substance use, or mental health symptoms may require different sensing modalities or fusion strategies. Behaviors with less distinctive physical gestures may prove more challenging for current approaches.

Second, our studies, while comprehensive, were conducted over relatively short

time periods (days to weeks rather than months to years). Longitudinal deployment may reveal habituation effects, privacy fatigue, or changing user needs that our current evaluations could not capture. As users become familiar with obfuscation techniques, their privacy perceptions may shift, potentially requiring adaptive strategies that evolve with user comfort and context.

Third, our evaluation of PRIMO focused on weight management experts examining hypothetical scenarios rather than live clinical deployment. Real-world clinical settings present additional challenges around liability, workflow integration, and time constraints that controlled studies cannot fully replicate.

Fourth, while recruitment for user studies was conducted broadly and without geographic constraints, the resulting participant sample was predominantly urban. Rural populations, older adults, and individuals with limited experience in technology may have different privacy concerns, varying comfort levels with wearable sensors, and varying levels of trust in AI systems.

6.4 Future Directions

Several promising research directions emerge from this work. The integration of large language models with multimodal sensing could enable a more sophisticated understanding of behavior, moving beyond activity detection to inferring behavioral intentions, emotional states, and contextual triggers. Such systems could provide not just "what" but "why" explanations for health behaviors, supporting more targeted interventions.

Federated learning approaches could allow behavioral models to improve across populations while preserving individual privacy. Rather than centralizing sensitive be-

havioral data, models could be trained locally, and only model updates shared, addressing both technical and ethical challenges in health data sharing. This could enable population-scale behavioral insights without compromising individual privacy.

The development of closed-loop intervention systems represents another frontier. By combining continuous sensing, real-time prediction, and just-in-time adaptive interventions, future systems could move beyond passive monitoring to active behavior change support. For example, a system that detects stress-induced smoking urges and automatically delivers personalized coping strategies before the behavior occurs.

Personalized privacy models that learn individual preferences and adapt obfuscation accordingly could address the context-dependency of privacy concerns we observed. Such systems could provide maximum privacy preservation when users are in sensitive contexts while relaxing constraints in comfortable, familiar settings.

Finally, extending explainable AI interfaces to support different stakeholders—from specialists to primary care providers to patients themselves—could democratize access to behavioral health insights. Patient-facing explanations could empower individuals to understand their behavioral patterns and participate more actively in their care.

Chapter 7

Conclusion

This dissertation presented a comprehensive framework for unobtrusive health behavior sensing that bridges the gap between real-world behavioral capture and clinical insight. Through the design and evaluation of HabitSense, RTFVNet, privacy-preserving obfuscation methods, and PRIMO, we addressed fundamental challenges in monitoring and understanding health-risk behaviors as they unfold in daily life.

Our work made four primary contributions to mobile and ubiquitous health sensing. First, we demonstrated that multimodal wearable systems could achieve the comfort and adherence necessary for continuous behavioral monitoring while maintaining clinical utility. Second, we established that complementary sensor fusion provides better performance-efficiency trade-offs compared to single-modality or late-fusion approaches. Third, we demonstrated that the privacy-utility trade-off can be balanced through context-aware generative obfuscation techniques. Fourth, we showed that explainable AI can bridge the gap between algorithmic predictions and clinical decision-making when designed as interactive exploration tools rather than static explanations.

These contributions converge on a central insight: successful health sensing systems must be designed as sociotechnical systems that respect human values while lever-

aging computational capabilities. Technical performance alone—whether in sensing accuracy, model precision, or privacy guarantees—is insufficient without corresponding attention to user comfort, social acceptability, and clinical interpretability.

This dissertation establishes a foundation for a future where continuous behavioral monitoring becomes routine in preventive healthcare—not as surveillance, but as collaborative tools that empower individuals and clinicians with timely, actionable insights. As sensors become more ubiquitous and AI more sophisticated, the frameworks developed here can help ensure these technologies enhance rather than compromise human agency and clinical judgment.

The path from sensing to insight is not merely technical but fundamentally human. By demonstrating that unobtrusive, privacy-preserving, clinically meaningful behavioral monitoring is both technically feasible and socially acceptable, we hope to accelerate the translation of ubiquitous computing innovations into real-world health impact. In doing so, we move closer to a future where health behaviors are understood not through sporadic self-reports or limited laboratory studies, but through continuous, context-aware observation that respects individual privacy and augments human understanding—ultimately helping individuals make healthier choices and clinicians provide more personalized, effective care.

References

- [1] *Obesity and Severe Obesity Prevalence in Adults: United States, August 2021-August 2023* - PubMed — [pubmed.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/39808758/?utm_source=chatgpt.com), https://pubmed.ncbi.nlm.nih.gov/39808758/?utm_source=chatgpt.com, [Accessed 07-01-2026].
- [2] CDC, *Cigarette Smoking* — [cdc.gov](https://www.cdc.gov/tobacco/about/index.html?utm_source=chatgpt.com), https://www.cdc.gov/tobacco/about/index.html?utm_source=chatgpt.com, [Accessed 07-01-2026].
- [3] L. N. Borrell and S. E. Echeverria, “The clustering effects of current smoking status, overweight/obesity, and physical inactivity with all-cause and cause-specific mortality risks in us adults,” *Preventive Medicine Reports*, vol. 42, p. 102742, 2024.
- [4] O. Perski et al., “Understanding health behaviours in context: A systematic review and meta-analysis of ecological momentary assessment studies of five key health behaviours,” *Health psychology review*, vol. 16, no. 4, pp. 576–601, 2022.
- [5] H. Hiraguchi, P. Perone, A. Toet, G. Camps, and A.-M. Brouwer, “Technology to automatically record eating behavior in real life: A systematic review,” *Sensors*, vol. 23, no. 18, p. 7757, 2023.
- [6] R. Alharbi, M. Tolba, L. C. Petito, J. Hester, and N. Alshurafa, “To Mask or Not to Mask? Balancing Privacy with Visual Confirmation Utility in Activity-Oriented Wearable Cameras,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, 72:1–72:29, Sep. 2019. DOI: [10.1145/3351230](https://doi.org/10.1145/3351230). Accessed: Mar. 24, 2022. [Online]. Available: <https://doi.org/10.1145/3351230>.
- [7] R. Alharbi, T. Stump, N. Vafaie, A. Pfammatter, B. Spring, and N. Alshurafa, “I Can’t Be Myself: Effects of Wearable Cameras on the Capture of Authentic Behavior in the Wild,” en, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–40, Sep. 2018, ISSN: 2474-9567. DOI: [10.1145/3351230](https://doi.org/10.1145/3351230).

3264900. Accessed: Aug. 31, 2022. [Online]. Available: <https://dl.acm.org/doi/10.1145/3264900>.
- [8] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, and P. Consortium, “Explainability for artificial intelligence in healthcare: A multidisciplinary perspective,” *BMC medical informatics and decision making*, vol. 20, no. 1, p. 310, 2020.
 - [9] C. V. Anikwe et al., “Mobile and wearable sensors for data-driven health monitoring system: State-of-the-art and future prospect,” *Expert Systems with Applications*, vol. 202, p. 117 362, 2022.
 - [10] M. Pedram et al., “Experience: Barriers and opportunities of wearables for eating research,” in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–8.
 - [11] B. M. Bell et al., “Automatic, wearable-based, in-field eating detection approaches for public health research: A scoping review,” *NPJ digital medicine*, vol. 3, no. 1, p. 38, 2020.
 - [12] R. Alharbi, S. Sen, A. Ng, N. Alshurafa, and J. Hester, “Actisight: Wearer foreground extraction using a practical rgb-thermal wearable,” in *2022 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, IEEE, 2022, pp. 237–246.
 - [13] R. Alharbi et al., “Smokemon: Unobtrusive extraction of smoking topography using wearable energy-efficient thermal,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 6, no. 4, Jan. 2023. DOI: [10.1145/3569460](https://doi.org/10.1145/3569460). [Online]. Available: <https://doi.org/10.1145/3569460>.
 - [14] G. J. Fernandes et al., “Habitsense: A privacy-aware, ai-enhanced multimodal wearable platform for mhealth applications,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 3, pp. 1–48, 2024.
 - [15] G. Fernandes, M. Lu, F. Shahabi, J. Zheng, A. Katsaggelos, and N. Alshurafa, “Multimodal hand-to-mouth gesture recognition in activity-oriented rgb-thermal footage (student abstract),” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 2025, pp. 29 368–29 370.
 - [16] G. Fernandes et al., “Is cartoonized life-vlogging the key to increasing adoption of activity-oriented wearable camera systems?” In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–8.

- [17] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, “Interpretability of machine learning-based prediction models in healthcare,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 5, e1379, 2020.
- [18] M. Jacobs, M. F. Pradier, T. H. McCoy, R. H. Perlis, F. Doshi-Velez, and K. Z. Gajos, “How machine-learning recommendations influence clinician treatment selections: The example of antidepressant selection,” en, *Translational Psychiatry*, vol. 11, no. 1, pp. 1–9, Feb. 2021, Number: 1 Publisher: Nature Publishing Group, ISSN: 2158-3188. DOI: [10.1038/s41398-021-01224-x](https://doi.org/10.1038/s41398-021-01224-x). Accessed: Jul. 31, 2022. [Online]. Available: <https://www.nature.com/articles/s41398-021-01224-x>.
- [19] F. Shahabi, S. L. Battalio, A. F. Pfammatter, D. Hedeker, B. Spring, and N. Alshurafa, “A machine-learned model for predicting weight loss success using weight change features early in treatment,” *npj Digital Medicine*, vol. 7, no. 1, p. 344, 2024.
- [20] G. J. Fernandes et al., “An explainable artificial intelligence software tool for weight management experts (primo): Mixed methods study,” *Journal of medical Internet research*, vol. 25, e42047, 2023.
- [21] CDC, *Obesity is a Common, Serious, and Costly Disease* — [cdc.gov](https://www.cdc.gov/obesity/data/adult.html), <https://www.cdc.gov/obesity/data/adult.html>, [Accessed 28-04-2024].
- [22] C. L. Ogden, M. D. Carroll, B. K. Kit, and K. M. Flegal, “Prevalence of obesity among adults: United states,” *NCHS data brief*, vol. 2013, no. 131, pp. 1–8, 2012.
- [23] CDCTobaccoFree, *Health Effects of Cigarette Smoking* — [cdc.gov](https://www.cdc.gov/tobacco/data_statistics/fact_sheets/health_effects/effects_cig_smoking/index.htm), https://www.cdc.gov/tobacco/data_statistics/fact_sheets/health_effects/effects_cig_smoking/index.htm, [Accessed 28-04-2024], 2024.
- [24] A. L. Association, *Tobacco Facts | State of Tobacco Control* — [lung.org](https://www.lung.org/research/sotc/facts), <https://www.lung.org/research/sotc/facts>, [Accessed 28-04-2024], 2024.
- [25] J. M. Samet, “Tobacco smoking: The leading cause of preventable disease worldwide,” *Thoracic surgery clinics*, vol. 23, no. 2, pp. 103–112, 2013.
- [26] D. E. Ramo, J. Thrul, E. A. Vogel, K. Delucchi, and J. J. Prochaska, “Multiple health risk behaviors in young adult smokers: Stages of change and stability over time,” *Annals of Behavioral Medicine*, vol. 54, no. 2, pp. 75–86, 2020.
- [27] U. D. of Health and H. Services, “The health consequences of smoking—50 years of progress: A report of the surgeon general,” *U.S. Department of Health and Human*

Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2014.

- [28] B. Stierman et al., “National health and nutrition examination survey 2017–march 2020 prepandemic data files development of files and prevalence estimates for selected health outcomes,” *National Center for Health Statistics (U.S.)*, 2021.
- [29] E. Kremer, A. Block, and M. Gaylor, “Behavioral approaches to treatment of chronic pain: The inaccuracy of patient self-report measures,” *Archives of Physical Medicine and Rehabilitation*, vol. 62, no. 4, pp. 188–191, 1981.
- [30] X. Fan et al., “An exploratory study about inaccuracy and invalidity in adolescent self-report surveys,” *Field methods*, vol. 18, no. 3, pp. 223–244, 2006.
- [31] T. F. Van de Mortel, “Faking it: Social desirability response bias in self-report research,” *Australian Journal of Advanced Nursing, The*, vol. 25, no. 4, pp. 40–48, 2008.
- [32] H. J. Arnold and D. C. Feldman, “Social desirability response bias in self-report choice situations,” *Academy of Management Journal*, vol. 24, no. 2, pp. 377–385, 1981.
- [33] B. Singh, E. M. Zopf, and E. J. Howden, “Effect and feasibility of wearable physical activity trackers and pedometers for increasing physical activity and improving health outcomes in cancer survivors: A systematic review and meta-analysis,” *Journal of sport and health science*, 2021.
- [34] Y. Vaizman, N. Weibel, and G. Lanckriet, “Context recognition in-the-wild: Unified model for multi-modal sensors and multi-label classification,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 4, Jan. 2018. DOI: [10.1145/3161192](https://doi.org/10.1145/3161192). [Online]. Available: <https://doi.org/10.1145/3161192>.
- [35] T. Ghosh et al., “Integrated image and sensor-based food intake detection in free-living,” *Scientific Reports*, vol. 14, no. 1, p. 1665, 2024.
- [36] M. Ulloa, B. Rothrock, F. S. Ahmad, and M. Jacobs, “Invisible clinical labor driving the successful integration of ai in healthcare,” *Frontiers in Computer Science*, vol. 4, p. 1045704, 2022.
- [37] D. R. Louie, M.-L. Bird, C. Menon, and J. J. Eng, “Perspectives on the prospective development of stroke-specific lower extremity wearable monitoring technology: A qualitative focus group study with physical therapists and individuals with stroke,” *Journal of neuroengineering and rehabilitation*, vol. 17, no. 1, pp. 1–11, 2020.

- [38] R. Alharbi, T. Stump, N. Vafaie, A. Pfammatter, B. Spring, and N. Alshurafa, "I can't be myself: Effects of wearable cameras on the capture of authentic behavior in the wild," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 2, no. 3, pp. 1–40, 2018.
- [39] J. H. Saltzer, "Protection and the control of information sharing in multics," *Communications of the ACM*, vol. 17, no. 7, pp. 388–402, 1974.
- [40] J. H. Saltzer and M. D. Schroeder, "The protection of information in computer systems," *Proceedings of the IEEE*, vol. 63, no. 9, pp. 1278–1308, 1975.
- [41] R. Alharbi, M. Tolba, L. C. Petito, J. Hester, and N. Alshurafa, "To mask or not to mask? balancing privacy with visual confirmation utility in activity-oriented wearable cameras," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 3, no. 3, pp. 1–29, 2019.
- [42] S. Shahi, M. Pedram, G. Fernandes, and N. Alshurafa, "Smartact: Energy efficient and real-time hand-to-mouth gesture detection using wearable rgb-t," in *2022 IEEE-EMBS International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, 2022, pp. 1–4. DOI: [10.1109/BSN56160.2022.9928492](https://doi.org/10.1109/BSN56160.2022.9928492).
- [43] M. B. Morshed et al., "A real-time eating detection system for capturing eating moments and triggering ecological momentary assessments to obtain further context: System development and validation study," *JMIR mHealth and uHealth*, vol. 8, no. 12, e20625, 2020.
- [44] M. Bin Morshed, H. K. Haresamudram, D. Bandaru, G. D. Abowd, and T. Plötz, "A personalized approach for developing a snacking detection system using earbuds in a semi-naturalistic setting," in *Proceedings of the 2022 ACM International Symposium on Wearable Computers*, 2022, pp. 11–16.
- [45] S. Zhang et al., "Necksense: A multi-sensor necklace for detecting eating activities in free-living conditions," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 4, no. 2, pp. 1–26, 2020.
- [46] A. Bedri, D. Li, R. Khurana, K. Bhuwarka, and M. Goel, "Fitbyte: Automatic diet monitoring in unconstrained situations using multimodal sensing on eyeglasses," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20, Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–12, ISBN: 9781450367080. DOI: [10.1145/3313831.3376869](https://doi.org/10.1145/3313831.3376869). [Online]. Available: <https://doi.org/10.1145/3313831.3376869>.

- [47] J. Shin et al., “Mydj: Sensing food intakes with an attachable on your eyeglass frame,” in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–17.
- [48] E. V. Añazco, P. R. Lopez, S. Lee, K. Byun, and T.-S. Kim, “Smoking activity recognition using a single wrist imu and deep learning light,” in *Proceedings of the 2nd international conference on digital signal processing*, 2018, pp. 48–51.
- [49] V. Y. Senyurek, M. H. Imtiaz, P. Belsare, S. Tiffany, and E. Sazonov, “Smoking detection based on regularity analysis of hand to mouth gestures,” *Biomedical signal processing and control*, vol. 51, pp. 106–112, 2019.
- [50] V. Senyurek, M. Imtiaz, P. Belsare, S. Tiffany, and E. Sazonov, “Cigarette smoking detection with an inertial sensor and a smart lighter,” *Sensors*, vol. 19, no. 3, p. 570, 2019.
- [51] Z. Gao et al., “Mmtsa: Multi-modal temporal segment attention network for efficient human activity recognition,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 7, no. 3, Sep. 2023. DOI: [10.1145/3610872](https://doi.org/10.1145/3610872). [Online]. Available: <https://doi.org/10.1145/3610872>.
- [52] S. Bi et al., “Auracle: Detecting eating episodes with an ear-mounted sensor,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–27, 2018.
- [53] M. T. Nyamukuru and K. M. Odame, “Tiny eats: Eating detection on a microcontroller,” *CoRR*, vol. abs/2003.06699, 2020. arXiv: [2003.06699](https://arxiv.org/abs/2003.06699). [Online]. Available: <https://arxiv.org/abs/2003.06699>.
- [54] A. Bedri et al., “Earbit: Using wearable sensors to detect eating episodes in unconstrained environments,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 3, Sep. 2017. DOI: [10.1145/3130902](https://doi.org/10.1145/3130902). [Online]. Available: <https://doi.org/10.1145/3130902>.
- [55] I. T. U. Echebarría, S. A. Imtiaz, M. Peng, and E. Rodriguez-Villegas, “Monitoring smoking behaviour using a wearable acoustic sensor,” in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2017, pp. 4459–4462. DOI: [10.1109/EMBC.2017.8037846](https://doi.org/10.1109/EMBC.2017.8037846).
- [56] E. Sazonov, P. Lopez-Meyer, and S. Tiffany, “A wearable sensor system for monitoring cigarette smoking,” *Journal of studies on alcohol and drugs*, vol. 74, no. 6, pp. 956–964, 2013.

- [57] P. Lopez-Meyer, S. Tiffany, Y. Patil, and E. Sazonov, "Monitoring of cigarette smoking using wearable sensors and support vector machines," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 7, pp. 1867–1872, 2013. DOI: [10.1109/TBME.2013.2243729](https://doi.org/10.1109/TBME.2013.2243729).
- [58] L.-C. Tai et al., "Nicotine monitoring with a wearable sweat band," *ACS sensors*, vol. 5, no. 6, pp. 1831–1837, 2020.
- [59] S. L. Battalio et al., "Sense2stop: A micro-randomized trial using wearable sensors to optimize a just-in-time-adaptive stress management intervention for smoking relapse prevention," *Contemporary Clinical Trials*, vol. 109, p. 106 534, 2021, ISSN: 1551-7144. DOI: <https://doi.org/10.1016/j.cct.2021.106534>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1551714421002706>.
- [60] S. Hodges, E. Berry, and K. Wood, "Sensecam: A wearable camera that stimulates and rehabilitates autobiographical memory," *Memory*, vol. 19, no. 7, pp. 685–696, 2011.
- [61] S. Bi and D. Kotz, "Eating detection with a head-mounted video camera," in *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*, 2022, pp. 60–66. DOI: [10.1109/ICHI54592.2022.00021](https://doi.org/10.1109/ICHI54592.2022.00021).
- [62] M. H. Imtiaz, D. Hossain, V. Y. Senyurek, P. Belsare, S. Tiffany, and E. Sazonov, "Wearable egocentric camera as a monitoring tool of free-living cigarette smoking: A feasibility study," *Nicotine and Tobacco Research*, vol. 22, no. 10, pp. 1883–1890, 2020.
- [63] L. Lu et al., "Wearable health devices in health care: Narrative systematic review," *JMIR mHealth and uHealth*, vol. 8, 2020. DOI: [10.2196/18907](https://doi.org/10.2196/18907).
- [64] A. Yetisen, J. L. Martinez-Hurtado, B. Ünal, A. Khademhosseini, and H. Butt, "Wearables in medicine," *Advanced Materials (Deerfield Beach, Fla.)*, vol. 30, 2018. DOI: [10.1002/adma.201706910](https://doi.org/10.1002/adma.201706910).
- [65] S. M. A. Iqbal, I. Mahgoub, E. Du, M. A. Leavitt, and W. Asghar, "Advances in health-care wearable devices," *npj Flexible Electronics*, vol. 5, pp. 1–14, 2021. DOI: [10.1038/s41528-021-00107-x](https://doi.org/10.1038/s41528-021-00107-x).
- [66] K. Bayoumy et al., "Smart wearable devices in cardiovascular care: Where we are and how to move forward," *Nature Reviews. Cardiology*, vol. 18, pp. 581–599, 2021. DOI: [10.1038/s41569-021-00522-7](https://doi.org/10.1038/s41569-021-00522-7).

- [67] A. Godfrey, V. Hetherington, H. P. H. Shum, P. Bonato, N. Lovell, and S. Stuart, "From a to z: Wearable technology explained.," *Maturitas*, vol. 113, pp. 40–47, 2018. DOI: [10.1016/j.maturitas.2018.04.012](https://doi.org/10.1016/j.maturitas.2018.04.012).
- [68] P. Bonato, "Advances in wearable technology and applications in physical medicine and rehabilitation," *Journal of NeuroEngineering and Rehabilitation*, vol. 2, pp. 2–2, 2005. DOI: [10.1186/1743-0003-2-2](https://doi.org/10.1186/1743-0003-2-2).
- [69] M. Chen, Y. Zhang, Y. Li, M. M. Hassan, and A. Alamri, "Aiwac: Affective interaction through wearable computing and cloud technology," *IEEE Wireless Communications*, vol. 22, pp. 20–27, 2015. DOI: [10.1109/MWC.2015.7054715](https://doi.org/10.1109/MWC.2015.7054715).
- [70] E. McAdams, C. Géhin, B. Massot, and J. McLaughlin, "The challenges facing wearable sensor systems," *Studies in health technology and informatics*, vol. 177, pp. 196–202, 2012. DOI: [10.3233/978-1-61499-069-7-196](https://doi.org/10.3233/978-1-61499-069-7-196).
- [71] H. Lewy, "Wearable technologies - future challenges for implementation in healthcare services.," *Healthcare technology letters*, vol. 2 1, pp. 2–5, 2015. DOI: [10.1049/htl.2014.0104](https://doi.org/10.1049/htl.2014.0104).
- [72] L. A. Simpson, C. Menon, A. Hodgson, W. B. Mortenson, and J. Eng, "Clinicians' perceptions of a potential wearable device for capturing upper limb activity post-stroke: A qualitative focus group study," *Journal of NeuroEngineering and Rehabilitation*, vol. 18, 2021. DOI: [10.1186/s12984-021-00927-y](https://doi.org/10.1186/s12984-021-00927-y).
- [73] A. Chattopadhyay and T. Boulton, "Privacyncam: A privacy preserving camera using uclinux on the blackfin dsp," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8. DOI: [10.1109/CVPR.2007.383413](https://doi.org/10.1109/CVPR.2007.383413).
- [74] T. Winkler and B. Rinner, "Trustcam: Security and privacy-protection for an embedded smart camera based on trusted computing," in *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2010, pp. 593–600. DOI: [10.1109/AVSS.2010.38](https://doi.org/10.1109/AVSS.2010.38).
- [75] T. Winkler, Á. Erdélyi, and B. Rinner, "Trusteye.m4: Protecting the sensor — not the camera," in *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2014, pp. 159–164. DOI: [10.1109/AVSS.2014.6918661](https://doi.org/10.1109/AVSS.2014.6918661).
- [76] L. Chan, V. D. Swain, C. Kelley, K. de Barbaro, G. D. Abowd, and L. Wilcox, "Students' experiences with ecological momentary assessment tools to report on emotional well-being," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2,

- no. 1, Mar. 2018. DOI: 10.1145/3191735. [Online]. Available: <https://doi.org/10.1145/3191735>.
- [77] N. Alshurafa, H. Kalantarian, M. Pourhomayoun, S. Sarin, J. J. Liu, and M. Sarrafzadeh, “Non-invasive monitoring of eating behavior using spectrogram analysis in a wearable necklace,” in *2014 IEEE healthcare innovation conference (HIC)*, IEEE, 2014, pp. 71–74.
 - [78] H. Kalantarian, N. Alshurafa, T. Le, and M. Sarrafzadeh, “Monitoring eating habits using a piezoelectric sensor-based necklace,” *Computers in biology and medicine*, vol. 58, pp. 46–55, 2015.
 - [79] N. Saleheen et al., “Puffmarker: A multi-sensor approach for pinpointing the timing of first lapse in smoking cessation,” in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2015, pp. 999–1010.
 - [80] S. Jero et al., “Practical principle of least privilege for secure embedded systems,” in *2021 IEEE 27th Real-Time and Embedded Technology and Applications Symposium (RTAS)*, IEEE, 2021, pp. 1–13.
 - [81] N. Alshurafa, J. Jain, T. K. Stump, B. Spring, and J. K. Robinson, “Assessing recall of personal sun exposure by integrating uv dosimeter and self-reported data with a network flow framework,” *Plos one*, vol. 14, no. 12, e0225371, 2019.
 - [82] T. K. Stump et al., “Daily minutes of unprotected sun exposure (muse) inventory: Measure description and comparisons to uvr sensor and sun protection survey data,” *Preventive medicine reports*, vol. 11, pp. 305–311, 2018.
 - [83] T. Stump et al., “Unprotected sun exposure and physical activity among melanoma survivors and first-degree relatives,” in *ANNALS OF BEHAVIORAL MEDICINE*, OXFORD UNIV PRESS INC JOURNALS DEPT, 2001 EVANS RD, CARY, NC 27513 USA, vol. 55, 2021, S491–S491.
 - [84] S. Xu, T. Stump, J. Jain, N. Alshurafa, and J. Robinson, “Variation in daily ultraviolet radiation exposure and sun protection behaviours of melanoma survivors: An observational single-arm pilot study with a wearable sensor,” *British Journal of Dermatology*, vol. 180, no. 2, pp. 413–414, 2019.
 - [85] HumanSignal, *Labelimg: Image annotation tool*, <https://github.com/HumanSignal/labelImg>, Accessed: 2023-11-11, 2023.

- [86] A. Vaswani et al., “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [87] G. Bertasius, H. Wang, and L. Torresani, “Is space-time attention all you need for video understanding?” In *ICML*, vol. 2, 2021, p. 4.
- [88] S. Abnar and W. Zuidema, “Quantifying attention flow in transformers,” *arXiv preprint arXiv:2005.00928*, 2020.
- [89] Y. Xu et al., *Timesformer-rolled-attention: Github repository*, <https://github.com/yiyixuxu/TimeSformer-rolled-attention>, Accessed: 2023-11-09, 2023.
- [90] G. Bertasius, H. Wang, and L. Torresani, *Timesformer: Github repository*, <https://github.com/facebookresearch/TimeSformer>, Accessed: 2023-11-09, 2023.
- [91] *Find Sunrise and Sunset Times Anywhere - SunriseSunset.io — sunrisesunset.io*, <https://sunrisesunset.io/>, [Accessed 29-04-2024].
- [92] S. Ma et al., “The era of 1-bit llms: All large language models are in 1.58 bits,” *arXiv preprint arXiv:2402.17764*, 2024.
- [93] A. Orhon, A. Wadhwa, Y. Kim, F. Rossi, and V. Jagadeesh. “Deploying transformers on the apple neural engine.” Accessed: 2023-11-15. [Online]. Available: <https://machinelearning.apple.com/research/neural-engine-transformers>.
- [94] Apple. “Apple introduces the advanced new apple watch series 9.” Accessed: 2023-11-15. [Online]. Available: <https://www.apple.com/newsroom/2023/09/apple-introduces-the-advanced-new-apple-watch-series-9/>.
- [95] A. of the study, “The prevalence of electronic cigarettes vaping globally: A systematic review and meta-analysis,” *Archives of Public Health*, vol. 80, 2022. [Online]. Available: <https://archpublichealth.biomedcentral.com/articles/10.1186/s13690-022-00998-w>.
- [96] <https://www.juul.com/>, [Accessed 27-04-2024].
- [97] *Posh Max 2.0 | Rechargeable Disposable Vape | Best Flavors — nowposh.com*, <https://nowposh.com/posh-max-2-0/>, [Accessed 28-04-2024].
- [98] *Iqos.com*, <https://www.iqos.com/global>, [Accessed 28-04-2024].

- [99] Y. Sun, W. Zuo, and M. Liu, “Rtfnnet: Rgb-thermal fusion network for semantic segmentation of urban scenes,” *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2576–2583, 2019.
- [100] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, “Attention bottlenecks for multimodal fusion,” *Advances in neural information processing systems*, vol. 34, pp. 14 200–14 213, 2021.
- [101] R. Girdhar et al., “Imagebind: One embedding space to bind them all,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 180–15 190.
- [102] S. Hodges, E. Berry, and K. Wood, “SenseCam: A wearable camera that stimulates and rehabilitates autobiographical memory,” in *Memory*, vol. 19, no. 7, pp. 685–696, Oct. 2011, ISSN: 0965-8211, 1464-0686. DOI: [10.1080/09658211.2011.605591](https://doi.org/10.1080/09658211.2011.605591). Accessed: Aug. 31, 2022. [Online]. Available: <http://www.tandfonline.com/doi/abs/10.1080/09658211.2011.605591>.
- [103] J. Chen et al., “Using the SenseCam as an objective tool for evaluating eating patterns,” in *Proceedings of the 4th International SenseCam & Pervasive Imaging Conference*, ser. SenseCam ’13, New York, NY, USA: Association for Computing Machinery, Nov. 2013, pp. 34–41, ISBN: 978-1-4503-2247-8. DOI: [10.1145/2526667.2526673](https://doi.org/10.1145/2526667.2526673). Accessed: Aug. 31, 2022. [Online]. Available: <https://doi.org/10.1145/2526667.2526673>.
- [104] E. Thomaz, A. Parnami, J. Bidwell, I. Essa, and G. D. Abowd, “Technological approaches for addressing privacy concerns when recognizing eating behaviors with wearable cameras,” in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, 2013, pp. 739–748.
- [105] A. J. Perez and S. Zeadally, “Privacy issues and solutions for consumer wearables,” *It Professional*, vol. 20, no. 4, pp. 46–56, 2017.
- [106] M. Koelle, S. Ananthanarayan, S. Czupalla, W. Heuten, and S. Boll, “Your smart glasses’ camera bothers me! exploring opt-in and opt-out gestures for privacy mediation,” in *Proceedings of the 10th Nordic Conference on Human-Computer Interaction*, 2018, pp. 473–481.
- [107] M. Koelle, S. Ananthanarayan, and S. Boll, “Social acceptability in hci: A survey of methods, measures, and design strategies,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–19.

- [108] E. T. Hassan, R. Hasan, P. Shaffer, D. Crandall, and A. Kapadia, “Cartooning for Enhanced Privacy in Lifelogging and Streaming Videos,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, ISSN: 2160-7516, Jul. 2017, pp. 1333–1342. DOI: [10.1109/CVPRW.2017.175](https://doi.org/10.1109/CVPRW.2017.175).
- [109] K. Brkic, I. Sikiric, T. Hrkac, and Z. Kalafatic, “I Know That Person: Generative Full Body and Face De-identification of People in Images,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, ISSN: 2160-7516, Jul. 2017, pp. 1319–1328. DOI: [10.1109/CVPRW.2017.173](https://doi.org/10.1109/CVPRW.2017.173).
- [110] Y. Li, N. Vishwamitra, B. P. Knijnenburg, H. Hu, and K. Caine, “Effectiveness and Users’ Experience of Obfuscation as a Privacy-Enhancing Technology for Sharing Photos,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. CSCW, 67:1–67:24, Dec. 2017. DOI: [10.1145/3134702](https://doi.org/10.1145/3134702). Accessed: Dec. 30, 2021. [Online]. Available: <https://doi.org/10.1145/3134702>.
- [111] Á. Erdélyi, T. Winkler, and B. Rinner, “Multi-Level Cartooning for Context-Aware Privacy Protection in Visual Sensor Networks,” in *MediaEval*, 2014.
- [112] Á. Erdélyi, T. Winkler, and B. Rinner, “Serious Fun: Cartooning for Privacy Protection,” in *MediaEval*, 2013.
- [113] E. Newton, L. Sweeney, and B. Malin, “Preserving privacy by de-identifying face images,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 2, pp. 232–243, Feb. 2005, Conference Name: IEEE Transactions on Knowledge and Data Engineering, ISSN: 1558-2191. DOI: [10.1109/TKDE.2005.32](https://doi.org/10.1109/TKDE.2005.32).
- [114] R. Gross, L. Sweeney, F. de la Torre, and S. Baker, “Model-Based Face De-Identification,” in *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW’06)*, ISSN: 2160-7516, Jun. 2006, pp. 161–161. DOI: [10.1109/CVPRW.2006.125](https://doi.org/10.1109/CVPRW.2006.125).
- [115] P. Korshunov and T. Ebrahimi, “Using face morphing to protect privacy,” in *2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*, Aug. 2013, pp. 208–213. DOI: [10.1109/AVSS.2013.6636641](https://doi.org/10.1109/AVSS.2013.6636641).
- [116] Q. Sun, A. Tewari, W. Xu, M. Fritz, C. Theobalt, and B. Schiele, “A Hybrid Model for Identity Obfuscation by Face Replacement,” in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2018, pp. 570–586, ISBN: 978-3-030-01246-5. DOI: [10.1007/978-3-030-01246-5_34](https://doi.org/10.1007/978-3-030-01246-5_34).

- [117] J. Lin, Y. Li, and G. Yang, “FPGAN: Face de-identification method with generative adversarial networks for social robots,” en, *Neural Networks*, vol. 133, pp. 132–147, Jan. 2021, ISSN: 0893-6080. DOI: [10.1016/j.neunet.2020.09.001](https://doi.org/10.1016/j.neunet.2020.09.001). Accessed: Dec. 30, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0893608020303245>.
- [118] Y. Chen, Y.-K. Lai, and Y.-J. Liu, “Cartoongan: Generative adversarial networks for photo cartoonization,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9465–9474.
- [119] X. Wang and J. Yu, “Learning to Cartoonize Using White-Box Cartoon Representations,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, ISSN: 2575-7075, Jun. 2020, pp. 8087–8096. DOI: [10.1109/CVPR42600.2020.00811](https://doi.org/10.1109/CVPR42600.2020.00811).
- [120] Y. Wu, F. Yang, Y. Xu, and H. Ling, “Privacy-Protective-GAN for Privacy Preserving Face De-Identification,” en, *Journal of Computer Science and Technology*, vol. 34, no. 1, pp. 47–60, Jan. 2019, ISSN: 1860-4749. DOI: [10.1007/s11390-019-1898-8](https://doi.org/10.1007/s11390-019-1898-8). Accessed: Dec. 30, 2021. [Online]. Available: <https://doi.org/10.1007/s11390-019-1898-8>.
- [121] Y. Shu et al., “GAN-based Multi-Style Photo Cartoonization,” *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2021, Conference Name: IEEE Transactions on Visualization and Computer Graphics, ISSN: 1941-0506. DOI: [10.1109/TVCG.2021.3067201](https://doi.org/10.1109/TVCG.2021.3067201).
- [122] T. Winkler, Á. Erdélyi, and B. Rinner, “TrustEYE.M4: Protecting the sensor — Not the camera,” in *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Aug. 2014, pp. 159–164. DOI: [10.1109/AVSS.2014.6918661](https://doi.org/10.1109/AVSS.2014.6918661).
- [123] J. R. Padilla-López, A. A. Chaaraoui, and F. Flórez-Revuelta, “Visual privacy protection methods: A survey,” en, *Expert Systems with Applications*, vol. 42, no. 9, pp. 4177–4195, Jun. 2015, ISSN: 0957-4174. DOI: [10.1016/j.eswa.2015.01.041](https://doi.org/10.1016/j.eswa.2015.01.041). Accessed: Dec. 30, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0957417415000561>.
- [124] Z. Ren, Y. J. Lee, and M. S. Ryoo, “Learning to Anonymize Faces for Privacy Preserving Action Detection,” en, in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2018, pp. 639–655, ISBN: 978-3-030-01246-5. DOI: [10.1007/978-3-030-01246-5_38](https://doi.org/10.1007/978-3-030-01246-5_38).

- [125] R. Hasan et al., “Can Privacy Be Satisfying? On Improving Viewer Satisfaction for Privacy-Enhanced Photos Using Aesthetic Transforms,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: Association for Computing Machinery, May 2019, pp. 1–13, ISBN: 978-1-4503-5970-2. Accessed: Dec. 29, 2021. [Online]. Available: <https://doi.org/10.1145/3290605.3300597>.
- [126] Y. Li, N. Vishwamitra, B. P. Knijnenburg, H. Hu, and K. Caine, “Blur vs. Block: Investigating the Effectiveness of Privacy-Enhancing Obfuscation for Images,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, ISSN: 2160-7516, Jul. 2017, pp. 1343–1351. DOI: [10.1109/CVPRW.2017.176](https://doi.org/10.1109/CVPRW.2017.176).
- [127] *Qualtrics xm - experience management software*, <https://www.qualtrics.com/>, Dec. 2022. [Online]. Available: <https://www.qualtrics.com/>.
- [128] *Prolific quickly find research participants you can trust*. <https://www.prolific.co/>, Dec. 2022. [Online]. Available: <https://www.prolific.co/>.
- [129] A. Biener, J. Cawley, and C. Meyerhoefer, “The High and Rising Costs of Obesity to the US Health Care System,” en, *Journal of General Internal Medicine*, vol. 32, no. 1, pp. 6–8, Apr. 2017, ISSN: 1525-1497. DOI: [10.1007/s11606-016-3968-8](https://doi.org/10.1007/s11606-016-3968-8). Accessed: Jul. 31, 2022. [Online]. Available: <https://doi.org/10.1007/s11606-016-3968-8>.
- [130] J. Cawley et al., “Direct medical costs of obesity in the United States and the most populous states,” *Journal of Managed Care & Specialty Pharmacy*, vol. 27, no. 3, pp. 354–366, Mar. 2021, ISSN: 2376-0540. DOI: [10.18553/jmcp.2021.20410](https://doi.org/10.18553/jmcp.2021.20410). Accessed: Jul. 31, 2022. [Online]. Available: <https://www.jmcp.org/doi/full/10.18553/jmcp.2021.20410>.
- [131] M. M. Islam, T. N. Poly, B. A. Walther, and Y.-C. (Li, “Use of Mobile Phone App Interventions to Promote Weight Loss: Meta-Analysis,” EN, *JMIR mHealth and uHealth*, vol. 8, no. 7, e17039, Jul. 2020. DOI: [10.2196/17039](https://doi.org/10.2196/17039). Accessed: Jul. 31, 2022. [Online]. Available: <https://mhealth.jmir.org/2020/7/e17039>.
- [132] X. Cai, S. Qiu, D. Luo, L. Wang, Y. Lu, and M. Li, “Mobile Application Interventions and Weight Loss in Type 2 Diabetes: A Meta-Analysis,” en, *Obesity*, vol. 28, no. 3, pp. 502–509, 2020, ISSN: 1930-739X. DOI: [10.1002/oby.22715](https://doi.org/10.1002/oby.22715). Accessed: Jul. 31, 2022. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/oby.22715>.
- [133] G. Farage et al., “Assessing the Contribution of Self-Monitoring Through a Commercial Weight Loss App: Mediation and Predictive Modeling Study,” EN, *JMIR*

- mHealth and uHealth*, vol. 9, no. 7, e18741, Jul. 2021. DOI: 10.2196/18741. Accessed: Jul. 31, 2022. [Online]. Available: <https://mhealth.jmir.org/2021/7/e18741>.
- [134] P. Bower and S. Gilbody, "Stepped care in psychological therapies: Access, effectiveness and efficiency: Narrative literature review," en, *The British Journal of Psychiatry*, vol. 186, no. 1, pp. 11–17, Jan. 2005, Publisher: Cambridge University Press, ISSN: 0007-1250, 1472-1465. DOI: 10.1192/bjp.186.1.11. Accessed: Jul. 31, 2022. [Online]. Available: <https://www.cambridge.org/core/journals/the-british-journal-of-psychiatry/article/stepped-care-in-psychological-therapies-access-effectiveness-and-efficiency/16CE065666C896FAAC8A39B80AE470E7>.
- [135] J. Finkelstein and I. C. Jeong, "Machine learning approaches to personalize early prediction of asthma exacerbations," *Annals of the New York Academy of Sciences*, vol. 1387, no. 1, pp. 153–165, 2017, Publisher: Wiley Online Library.
- [136] G. Tripathi and R. Kumar, "Early prediction of diabetes mellitus using machine learning," in *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, IEEE, 2020, pp. 1009–1014.
- [137] D. A. Martinez et al., "Early Prediction of Acute Kidney Injury in the Emergency Department With Machine-Learning Methods Applied to Electronic Health Record Data," en, *Annals of Emergency Medicine*, vol. 76, no. 4, pp. 501–514, Oct. 2020, ISSN: 0196-0644. DOI: 10.1016/j.annemergmed.2020.05.026. Accessed: Nov. 5, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0196064420303978>.
- [138] M. Moor, B. Rieck, M. Horn, C. R. Jutzeler, and K. Borgwardt, "Early Prediction of Sepsis in the ICU Using Machine Learning: A Systematic Review," *Frontiers in Medicine*, vol. 8, 2021, ISSN: 2296-858X. Accessed: Nov. 5, 2022. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fmed.2021.607952>.
- [139] K. O. Asare, Y. Terhorst, J. Vega, E. Peltonen, E. Lagerspetz, and D. Ferreira, "Predicting Depression From Smartphone Behavioral Markers Using Machine Learning Methods, Hyperparameter Optimization, and Feature Importance Analysis: Exploratory Study," EN, *JMIR mHealth and uHealth*, vol. 9, no. 7, e26540, Jul. 2021. DOI: 10.2196/26540. Accessed: Jul. 31, 2022. [Online]. Available: <https://mhealth.jmir.org/2021/7/e26540>.
- [140] M. Kim, J. Yang, W.-Y. Ahn, and H. J. Choi, "Machine Learning Analysis to Identify Digital Behavioral Phenotypes for Engagement and Health Outcome Efficacy of an mHealth Intervention for Obesity: Randomized Controlled Trial," EN, *Journal of Medical Internet Research*, vol. 23, no. 6, e27218, Jun. 2021. DOI: 10.2196/27218.

- Accessed: Jul. 31, 2022. [Online]. Available: <https://www.jmir.org/2021/6/e27218>.
- [141] I. Akour, M. Alshurideh, B. A. Kurdi, A. A. Ali, and S. Salloum, "Using Machine Learning Algorithms to Predict People's Intention to Use Mobile Learning Platforms During the COVID-19 Pandemic: Machine Learning Approach," EN, *JMIR Medical Education*, vol. 7, no. 1, e24032, Feb. 2021. DOI: [10.2196/24032](https://doi.org/10.2196/24032). Accessed: Jul. 31, 2022. [Online]. Available: <https://mededu.jmir.org/2021/1/e24032>.
 - [142] J. Yu, C. Chiu, Y. Wang, E. Dzibur, W. Lu, and J. Hoffman, "A Machine Learning Approach to Passively Informed Prediction of Mental Health Risk in People with Diabetes: Retrospective Case-Control Analysis," EN, *Journal of Medical Internet Research*, vol. 23, no. 8, e27709, Aug. 2021. DOI: [10.2196/27709](https://doi.org/10.2196/27709). Accessed: Jul. 31, 2022. [Online]. Available: <https://www.jmir.org/2021/8/e27709>.
 - [143] T. Mori and N. Uchihiro, "Balancing the trade-off between accuracy and interpretability in software defect prediction," en, *Empirical Software Engineering*, vol. 24, no. 2, pp. 779–825, Apr. 2019, ISSN: 1573-7616. DOI: [10.1007/s10664-018-9638-1](https://doi.org/10.1007/s10664-018-9638-1). Accessed: Jul. 31, 2022. [Online]. Available: <https://doi.org/10.1007/s10664-018-9638-1>.
 - [144] J. Yoo, S.-H. Kim, S. Hur, J. Ha, K. Huh, and W. C. Cha, "Candidemia Risk Prediction (CanDETEC) Model for Patients With Malignancy: Model Development and Validation in a Single-Center Retrospective Study," EN, *JMIR Medical Informatics*, vol. 9, no. 7, e24651, Jul. 2021. DOI: [10.2196/24651](https://doi.org/10.2196/24651). Accessed: Jul. 31, 2022. [Online]. Available: <https://medinform.jmir.org/2021/7/e24651>.
 - [145] S. Surodina, C. Lam, S. Grbich, M. Milne-Ives, M. v. Velthoven, and E. Meinert, "Machine Learning for Risk Group Identification and User Data Collection in a Herpes Simplex Virus Patient Registry: Algorithm Development and Validation Study," EN, *JMIRx Med*, vol. 2, no. 2, e25560, Jun. 2021. DOI: [10.2196/25560](https://doi.org/10.2196/25560). Accessed: Jul. 31, 2022. [Online]. Available: <https://xmed.jmir.org/2021/2/e25560>.
 - [146] R. O'Driscoll et al., "Comparison of the Validity and Generalizability of Machine Learning Algorithms for the Prediction of Energy Expenditure: Validation Study," EN, *JMIR mHealth and uHealth*, vol. 9, no. 8, e23938, Aug. 2021. DOI: [10.2196/23938](https://doi.org/10.2196/23938). Accessed: Jul. 31, 2022. [Online]. Available: <https://mhealth.jmir.org/2021/8/e23938>.
 - [147] F. Cacheda, D. Fernandez, F. J. Novoa, and V. Carneiro, "Early Detection of Depression: Social Network Analysis and Random Forest Techniques," EN, *Journal of*

- Medical Internet Research*, vol. 21, no. 6, e12554, Jun. 2019. DOI: 10.2196/12554. Accessed: Jul. 31, 2022. [Online]. Available: <https://www.jmir.org/2019/6/e12554>.
- [148] A. Chatzimpampas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren, “The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations,” en, *Computer Graphics Forum*, vol. 39, no. 3, pp. 713–756, 2020, _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14034>, ISSN: 1467-8659. DOI: 10.1111/cgf.14034. Accessed: Jan. 5, 2023. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14034>.
- [149] K. E. Henry et al., “Human-machine teaming is key to AI adoption: Clinicians’ experiences with a deployed machine learning system,” en, *npj Digital Medicine*, vol. 5, no. 1, pp. 1–6, Jul. 2022, Number: 1 Publisher: Nature Publishing Group, ISSN: 2398-6352. DOI: 10.1038/s41746-022-00597-7. Accessed: Jan. 5, 2023. [Online]. Available: <https://www.nature.com/articles/s41746-022-00597-7>.
- [150] A. Bussone, S. Stumpf, and D. O’Sullivan, “The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems,” in *2015 International Conference on Healthcare Informatics*, Oct. 2015, pp. 160–169. DOI: 10.1109/ICHI.2015.26.
- [151] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017. Accessed: Jul. 31, 2022. [Online]. Available: <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>.
- [152] M. T. Ribeiro, S. Singh, and C. Guestrin, ““Why Should I Trust You?”: Explaining the Predictions of Any Classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’16, New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 1135–1144, ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778. Accessed: Jul. 31, 2022. [Online]. Available: <https://doi.org/10.1145/2939672.2939778>.
- [153] M. T. Ribeiro, S. Singh, and C. Guestrin, “Anchors: High-Precision Model-Agnostic Explanations,” en, *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018, Number: 1, ISSN: 2374-3468. DOI: 10.1609/aaai.v32i1.11491. Accessed: Jul. 31, 2022. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/11491>.
- [154] S. Mohseni, N. Zarei, and E. D. Ragan, “A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems,” *ACM Transactions on Interactive Intelligent Systems*, vol. 11, no. 3-4, 24:1–24:45, Aug. 2021, ISSN: 2160-6455. DOI:

- 10.1145/3387166. Accessed: Jul. 31, 2022. [Online]. Available: <https://doi.org/10.1145/3387166>.
- [155] D. A. Melis, H. Kaur, H. D. Iii, H. Wallach, and J. W. Vaughan, “From Human Explanation to Model Interpretability: A Framework Based on Weight of Evidence,” en, *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 9, pp. 35–47, Oct. 2021, ISSN: 2769-1349. Accessed: Jul. 31, 2022. [Online]. Available: <https://ojs.aaai.org/index.php/HCOMP/article/view/18938>.
 - [156] A. Shih, A. Choi, and A. Darwiche, “Compiling Bayesian Network Classifiers into Decision Graphs,” en, *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 7966–7974, Jul. 2019, Number: 01, ISSN: 2374-3468. DOI: 10.1609/aaai.v33i01.33017966. Accessed: Jul. 31, 2022. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/4797>.
 - [157] W. Shi, A. Shih, A. Darwiche, and A. Choi, *On Tractable Representations of Binary Neural Networks*, arXiv:2004.02082 [cs], Jul. 2020. DOI: 10.48550/arXiv.2004.02082. Accessed: Jul. 31, 2022. [Online]. Available: <http://arxiv.org/abs/2004.02082>.
 - [158] A. Choi, A. Shih, A. Goyanka, and A. Darwiche, *On Symbolically Encoding the Behavior of Random Forests*, arXiv:2007.01493 [cs], Jul. 2020. DOI: 10.48550/arXiv.2007.01493. Accessed: Jul. 31, 2022. [Online]. Available: <http://arxiv.org/abs/2007.01493>.
 - [159] E. Kim et al., “Application of machine learning to predict weight loss in overweight, and obese patients on korean medicine weight management program,” *The Journal of Korean Medicine*, vol. 41, no. 2, pp. 58–79, 2020, Publisher: The Society of Korean Medicine.
 - [160] C. A. Pellegrini, S. A. Hoffman, L. M. Collins, and B. Spring, “Optimization of remotely delivered intensive lifestyle treatment for obesity using the Multiphase Optimization Strategy: Opt-IN study protocol,” en, *Contemporary Clinical Trials*, vol. 38, no. 2, pp. 251–259, Jul. 2014, ISSN: 1551-7144. DOI: 10.1016/j.cct.2014.05.007. Accessed: Jul. 31, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1551714414000755>.
 - [161] B. Spring et al., “A Factorial Experiment to Optimize Remotely Delivered Behavioral Treatment for Obesity: Results of the Opt-IN Study,” en, *Obesity*, vol. 28, no. 9, pp. 1652–1662, 2020, ISSN: 1930-739X. DOI: 10.1002/oby.22915. Accessed: Jul. 31, 2022. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/oby.22915>.

- [162] W. C. Knowler et al., “Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin,” 2002.
- [163] M. L. Patel, C. M. Hopkins, and G. G. Bennett, “Early weight loss in a standalone mHealth intervention predicting treatment success,” en, *Obesity Science & Practice*, vol. 5, no. 3, pp. 231–237, 2019, ISSN: 2055-2238. DOI: 10.1002/osp4.329. Accessed: Nov. 23, 2022. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/osp4.329>.
- [164] B. Spring, C. A. Pellegrini, and I. Nahum-Shani, “How long does it take to identify non-responders to weight loss treatment,” in *annual meeting of The Obesity Society, Boston, MA*, 2014.
- [165] M. E. Waring et al., “Early-treatment weight loss predicts 6-month weight loss in women with obesity and depression: Implications for stepped care,” en, *Journal of Psychosomatic Research*, vol. 76, no. 5, pp. 394–399, May 2014, ISSN: 0022-3999. DOI: 10.1016/j.jpsychores.2014.03.004. Accessed: Nov. 23, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0022399914000737>.
- [166] L. Hooper, A. Abdelhamid, H. J. Moore, W. Douthwaite, C. M. Skeaff, and C. D. Summerbell, “Effect of reducing total fat intake on body weight: Systematic review and meta-analysis of randomised controlled trials and cohort studies,” en, *BMJ*, vol. 345, e7666, Dec. 2012, Publisher: British Medical Journal Publishing Group Section: Research, ISSN: 1756-1833. DOI: 10.1136/bmj.e7666. Accessed: Nov. 23, 2022. [Online]. Available: <https://www.bmj.com/content/345/bmj.e7666>.
- [167] C. M. Champagne et al., “Dietary intakes associated with successful weight loss and maintenance during the Weight Loss Maintenance Trial,” *Journal of the American Dietetic Association*, vol. 111, no. 12, pp. 1826–1835, Dec. 2011, ISSN: 0002-8223. DOI: 10.1016/j.jada.2011.09.014. Accessed: Nov. 23, 2022. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3225890/>.
- [168] P.-H. Lin et al., “The Association Between Engagement and Weight Loss Through Personal Coaching and Cell Phone Interventions in Young Adults: Randomized Controlled Trial,” EN, *JMIR mHealth and uHealth*, vol. 6, no. 10, e10471, Oct. 2018. DOI: 10.2196/10471. Accessed: Nov. 23, 2022. [Online]. Available: <https://mhealth.jmir.org/2018/10/e10471>.
- [169] A. Carey, Q. Yang, L. DeLuca, T. Toro-Ramos, Y. Kim, and A. Michaelides, “The Relationship Between Weight Loss Outcomes and Engagement in a Mobile Behavioral Change Intervention: Retrospective Analysis,” EN, *JMIR mHealth and uHealth*,

- vol. 9, no. 11, e30622, Nov. 2021. DOI: [10.2196/30622](https://doi.org/10.2196/30622). Accessed: Nov. 23, 2022. [Online]. Available: <https://mhealth.jmir.org/2021/11/e30622>.
- [170] A. Darwiche and P. Marquis, “A Knowledge Compilation Map,” en, *Journal of Artificial Intelligence Research*, vol. 17, pp. 229–264, Sep. 2002, ISSN: 1076-9757. DOI: [10.1613/jair.989](https://doi.org/10.1613/jair.989). Accessed: Jul. 31, 2022. [Online]. Available: <https://www.jair.org/index.php/jair/article/view/10311>.
 - [171] A. Shih, *RF_sdd*, original-date: 2018-12-15T00:13:35Z, Jul. 2021. Accessed: Aug. 18, 2022. [Online]. Available: https://github.com/AndyShih12/RF_SDD.
 - [172] A. Shih, A. Choi, and A. Darwiche, *A Symbolic Approach to Explaining Bayesian Network Classifiers*, arXiv:1805.03364 [cs], May 2018. DOI: [10.48550/arXiv.1805.03364](https://doi.org/10.48550/arXiv.1805.03364). Accessed: Jul. 31, 2022. [Online]. Available: <http://arxiv.org/abs/1805.03364>.
 - [173] A. Shih, A. Choi, and A. Darwiche, “Formal Verification of Bayesian Network Classifiers,” en, in *Proceedings of the Ninth International Conference on Probabilistic Graphical Models*, ISSN: 2640-3498, PMLR, Aug. 2018, pp. 427–438. Accessed: Jul. 31, 2022. [Online]. Available: <https://proceedings.mlr.press/v72/shih18a.html>.
 - [174] A. Darwiche and A. Hirth, *On The Reasons Behind Decisions*, arXiv:2002.09284 [cs], Feb. 2020. DOI: [10.48550/arXiv.2002.09284](https://doi.org/10.48550/arXiv.2002.09284). Accessed: Jul. 31, 2022. [Online]. Available: <http://arxiv.org/abs/2002.09284>.
 - [175] O. Coudert, J. C. Madre, H. Fräisse, and H. Touati, “Implicit prime cover computation: An overview,” in *Proc. SASIMI’93*, 1993.
 - [176] O. Coudert and J. C. Madre, “Fault tree analysis: 10/sup 20/prime implicants and beyond,” in *Annual Reliability and Maintainability Symposium 1993 Proceedings*, IEEE, 1993, pp. 240–245.
 - [177] B. Cahour and J.-F. Forzy, “Does projection into use improve trust and exploration? An example with a cruise control system,” en, *Safety Science*, Research in Ergonomic Psychology in the Transportation Field in France, vol. 47, no. 9, pp. 1260–1270, Nov. 2009, ISSN: 0925-7535. DOI: [10.1016/j.ssci.2009.03.015](https://doi.org/10.1016/j.ssci.2009.03.015). Accessed: Jul. 31, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925753509000587>.
 - [178] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec, “Faithful and Customizable Explanations of Black Box Models,” in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES ’19, New York, NY, USA: Association for Computing Machinery, Jan. 2019, pp. 131–138, ISBN: 978-1-4503-6324-2. DOI: [10.1145/3292500.3292501](https://doi.org/10.1145/3292500.3292501).

- 3306618.3314229. Accessed: Aug. 17, 2022. [Online]. Available: <https://doi.org/10.1145/3306618.3314229>.
- [179] M. Yin, J. Wortman Vaughan, and H. Wallach, “Understanding the Effect of Accuracy on Trust in Machine Learning Models,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’19, New York, NY, USA: Association for Computing Machinery, May 2019, pp. 1–12, ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300509. Accessed: Aug. 17, 2022. [Online]. Available: <https://doi.org/10.1145/3290605.3300509>.
 - [180] L. Al-Abdullah, P. Welsh, and J. Logue, “Development of a predictive model for short and medium-term weight loss in people with type 2 diabetes attending a weight management programme,” in *DIABETOLOGIA*, Issue: SUPPL 1, vol. 65, SPRINGER ONE NEW YORK PLAZA, SUITE 4600, NEW YORK, NY, UNITED STATES, 2022, S266–S266.
 - [181] N. ALSHURAF, J. JAIN, R. ALHARBI, G. IAKOVLEV, B. SPRING, and A. PFAM-MATTER, “Is More Always Better?: Discovering Incentivized mHealth Intervention Engagement Related to Health Behavior Trends,” *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 2, no. 4, p. 153, Dec. 2018, ISSN: 2474-9567. DOI: 10.1145/3287031. Accessed: Nov. 23, 2022. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7173729/>.
 - [182] S. Särnblad, U. Ekelund, and J. Åman, “Dietary Fat Intake Predicts 1-Year Change in Body Fat in Adolescent Girls With Type 1 Diabetes,” *Diabetes Care*, vol. 29, no. 6, pp. 1227–1230, Jun. 2006, ISSN: 0149-5992. DOI: 10.2337/dc05-2318. Accessed: Nov. 23, 2022. [Online]. Available: <https://doi.org/10.2337/dc05-2318>.
 - [183] C. M. Grilo, M. A. White, R. M. Masheb, V. Ivezaj, P. T. Morgan, and R. Gueorguieva, “Randomized controlled trial testing the effectiveness of adaptive “SMART” stepped-care treatment for adults with binge-eating disorder comorbid with obesity,” *American Psychologist*, vol. 75, no. 2, pp. 204–218, 2020, Place: US Publisher: American Psychological Association, ISSN: 1935-990X. DOI: 10.1037/amp0000534.
 - [184] F. Shahabi, S. Battalio, A. Pfammatter, N. Alshurafa, and B. Spring, “Weight-Loss Prediction: Building and Interpreting a Machine-learned Model That Predicts Weight-Loss Success,” *ANNALS OF BEHAVIORAL MEDICINE*, Sep. 2022.
 - [185] J.-H. Eom, S.-C. Kim, and B.-T. Zhang, “AptaCDSS-E: A classifier ensemble-based clinical decision support system for cardiovascular disease level prediction,” in *Expert Systems with Applications*, vol. 34, no. 4, pp. 2465–2479, May 2008, ISSN: 0957-

4174. DOI: [10.1016/j.eswa.2007.04.015](https://doi.org/10.1016/j.eswa.2007.04.015). Accessed: Aug. 1, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S095741740700139X>.
- [186] A. M. Antoniadi et al., “Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review,” *en, Applied Sciences*, vol. 11, no. 11, p. 5088, Jan. 2021, Number: 11 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2076-3417. DOI: [10.3390/app11115088](https://doi.org/10.3390/app11115088). Accessed: Aug. 1, 2022. [Online]. Available: <https://www.mdpi.com/2076-3417/11/11/5088>.
- [187] J. Pearl and D. Mackenzie, *The book of why: the new science of cause and effect*. Basic books, 2018.
- [188] A. Darwiche, “Human-level intelligence or animal-like abilities?” *Communications of the ACM*, vol. 61, no. 10, pp. 56–67, 2018, Publisher: ACM New York, NY, USA.
- [189] W. K. Diprose, N. Buist, N. Hua, Q. Thurier, G. Shand, and R. Robinson, “Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator,” *Journal of the American Medical Informatics Association*, vol. 27, no. 4, pp. 592–600, Apr. 2020, ISSN: 1527-974X. DOI: [10.1093/jamia/ocz229](https://doi.org/10.1093/jamia/ocz229). Accessed: Jul. 31, 2022. [Online]. Available: <https://doi.org/10.1093/jamia/ocz229>.