



NORTHWESTERN UNIVERSITY

Computer Science Department

Technical Report
Number: NU-CS-2026-42

September 2026

AI-Enabled Digital Phenotyping and Multimodal Sensing for Precision Behavioral Health

Farzad Shahabi

Abstract

Health-risk behaviors that contribute to chronic disease unfold continuously in daily life, yet they remain difficult to measure, model, and personalize at scale. This dissertation addresses these challenges by developing computational methods for early risk prediction, multimodal behavior sensing, digital phenotyping, and label-efficient learning. It first introduces an interpretable machine-learning framework that uses early weight-monitoring trajectories to predict long-term weight-loss outcomes and identify individuals at risk for poor treatment response. Next, it presents SenseWhy, a multimodal sensing framework that combines wearable cameras, ecological momentary assessment, dietary self-report, and clinical recall data to detect personalized meal-level overeating in free-living settings. Building on these data, the dissertation applies contrastive representation learning to identify distinct overeating patterns associated with psychological, contextual, and behavioral factors. Finally, it introduces an active learning framework that combines cluster-based entropy sampling with uncertainty derived from training dynamics to reduce annotation requirements while maintaining predictive performance. Together, these contributions advance scalable, interpretable, and personalized behavioral AI systems for health monitoring and intervention.

Multimodal sensing, Digital phenotyping, Machine learning, Wearable sensing

NORTHWESTERN UNIVERSITY

AI-Enabled Digital Phenotyping and Multimodal Sensing for Precision Behavioral Health

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Computer Science

By

Farzad Shahabi

EVANSTON, ILLINOIS

September 2026

© Copyright by Farzad Shahabi, 2026

All Rights Reserved

Abstract

Health-risk behaviors that shape chronic disease risk often unfold continuously in daily life, yet most clinical and behavioral assessments capture them only episodically, retrospectively, or through burdensome self-report. Wearable and mobile sensing create an opportunity to measure these behaviors objectively, longitudinally, and in context, enabling earlier risk identification and more personalized behavioral intervention. This gap limits the ability to identify emerging risk, understand the behavioral mechanisms that drive poor health outcomes, and deliver timely, personalized support. Precision behavioral health therefore requires computational systems that can detect risk early, objectively measure proximal behaviors, transform longitudinal data into interpretable phenotypes, and scale across real-world deployments without requiring prohibitive manual annotation. This research develops an AI-enabled framework for digital phenotyping and multimodal sensing in behavioral health, progressing from early risk prediction to behavior detection, phenotype discovery, and label-efficient model development.

In *Chapter 2*, this research first examines whether early signals from continuous self-monitoring can identify individuals at risk for suboptimal weight loss. Using longitudinal weight trajectories from behavioral weight management trials, an explainable machine learning model is developed by combining baseline characteristics with dynamic features extracted from the first two weeks of treatment. The model demonstrates that early weight-loss trajectory features, including weight change, variability, and distributional patterns, can predict longer-term treatment response and generalize across external cohorts. This chapter establishes the first scientific premise of the dissertation: continuously collected behavioral data can provide actionable early indicators of treatment risk before conventional clinical decision points.

While early weight trajectories can identify *who* may be at risk, effective intervention also requires understanding *why*, *when*, and *under what conditions* high-risk eating behaviors occur. In *Chapter 3*, this research addresses this gap through the SenseWhy study, which integrates wearable camera sensing, ecological momentary assessment, mobile dietary reporting, and dietitian-administered recalls to characterize overeating in free-living conditions. Overeating is operationalized at the personalized meal level, and supervised machine learning models are evaluated using psychological, contextual, and passive sensing features. The results show that multimodal features provide complementary information for detecting overeating episodes, linking subjective states such as hunger, loss of control, and perceived overeating with objective behavioral markers such as chews, bites, eating duration, and chewing dynamics.

In *Chapter 4*, overeating detection is extended toward digital phenotyping. Rather than treating overeating as a homogeneous behavior, this chapter uses multimodal meal-level data to learn latent representations of eating episodes and identify distinct overeating phenotypes. The approach combines supervised contrastive learning, dimensionality reduction, clustering, and feature-level interpretation to uncover behavioral subtypes defined by different configurations of psychological state, eating context, and sensor-derived eating microstructure. These phenotypes provide a mechanistic bridge between detection and intervention by showing that overeating can arise through multiple pathways, including stress-related eating, socially situated eating, distracted eating, and altered eating pace.

Finally, *Chapter 5* addresses a central scalability barrier in mobile and wearable health sensing: the high cost of annotating large volumes of sensor data. An active learning framework is developed to combine model uncertainty with intrinsic uncertainty de-

rived from training dynamics, enabling the selection of informative samples for annotation. By identifying samples that are both useful for improving decision boundaries and representative of difficult-to-learn behavioral patterns, this approach reduces labeling burden while maintaining model performance. This chapter positions active learning as a practical mechanism for scaling AI-enabled behavioral sensing systems beyond small, manually annotated research datasets.

Together, these chapters form a coherent sensing-to-insight pipeline for precision behavioral health. *Chapter 2* demonstrates that continuous monitoring can support early risk stratification; *Chapter 3* shows how multimodal sensing can detect proximal overeating episodes in daily life; *Chapter 4* transforms these episodes into interpretable behavioral phenotypes; and *Chapter 5* advances the scalability of such systems through label-efficient learning. Collectively, this dissertation contributes computational methods and empirical evidence for building adaptive, interpretable, and scalable AI systems that move behavioral health research from retrospective assessment toward timely, personalized intervention.

Acknowledgment

I would like to express my deepest gratitude to my advisor, Professor Nabil Alshurafa, for his mentorship, support, and guidance throughout my PhD. His commitment to rigorous thinking and scientific excellence shaped the way I approach research and problem-solving. Beyond his academic mentorship, his kindness, encouragement, and belief in his students helped me grow both as a researcher and as a person. I am deeply grateful for the opportunity to learn from him.

I am also sincerely grateful to my dissertation committee members, Professor Maia Jacobs, Professor Jacob Schauer, and Professor Karan Ahuja, for their time, thoughtful feedback, and valuable perspectives. Their insights strengthened this work and helped me think more critically about its technical, clinical, and human-centered contributions.

I am thankful to my labmates and collaborators, including Glenn Fernandes, Soroush Shahi, Boyang Wei, and Mahdi Pedram, for their support, collaboration, and many insightful discussions throughout this journey. Their contributions, feedback, and friendship made this research journey both more productive and more meaningful.

I am also grateful to my close friends, Mehdi Moshafi, Babak Saleck Pay, Soroush Karami, and Javad Sajjadi. Whether from Chicago or across the world, they always had my back through the challenges and uncertainties of this journey.

List of Abbreviations

AI	Artificial Intelligence
AUM	Area Under the Margin
AUPRC	Area Under the Precision–Recall Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
BMI	Body Mass Index
CME	Cluster-Based Maximum Entropy
DBI	Davies–Bouldin Index
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DNN	Deep Neural Network
EMA	Ecological Momentary Assessment
F1	F1 Score
GMM	Gaussian Mixture Model
IMU	Inertial Measurement Unit
IQR	Interquartile Range
LOC	Loss of Control
LOESS	Locally Weighted Scatterplot Smoothing

mHealth	Mobile Health
ML	Machine Learning
MLP	Multilayer Perceptron
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
SASWV	Sparsity-Adjusted Successive Weight Variability
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
UBS	User Burden Scale
UMAP	Uniform Manifold Approximation and Projection
XGBoost	Extreme Gradient Boosting

Dedication

*To my mother, Mehri, and my father, Houshang, whose love,
sacrifices, and belief in me made this possible.*

*To my brothers, Mehrdad and Farhad, and to Roshanak and Elmira,
for their constant encouragement, and support throughout the years.*

*To my partner Kelsey, for walking beside me through every challenge
and success with patience, love, and unwavering support.*

*And to the memory of my grandmother, Vahdat, whose love and
selflessness will always remain with me.*

Table of Contents

Abstract	3
Acknowledgment	6
List of Abbreviations	7
Table of Contents	10
List of Figures	13
List of Tables	19
1 Introduction	21
1.1 Motivation	21
1.2 From Continuous Monitoring to Early Risk Identification	23
1.3 From Risk Identification to Proximal Behavior Detection	24
1.4 From Behavior Detection to Digital Phenotyping	25
1.5 From Manual Annotation to Scalable Behavioral AI	26
1.6 Dissertation Overview and Major Contributions	27
2 Early Risk Prediction	29
2.1 Technical Challenge: Early Identification of Treatment Risk	29
2.2 Longitudinal Self-Monitoring Data and Prediction Task	31

	11
2.3 Early-Trajectory Feature Engineering and Risk Modeling	35
2.4 Evaluation and External Validation	39
2.5 Discussion	46
2.6 Summary	53
2.7 Supplementary	55
3 SenseWhy	64
3.1 Technical Challenge: Objective Measurement of Proximal Eating Behaviors in Free-Living Settings	64
3.2 SenseWhy Study Design and Personalized Overeating Labels	66
3.3 Multimodal Feature Construction and Overeating Detection	69
3.4 Evaluation and Results	77
3.5 Discussion and Limitations	88
3.6 Summary	96
3.7 Supplementary	99
4 Digital Phenotyping	119
4.1 Technical Challenge: Digital Phenotyping Under Sparse and Heterogeneous Free-Living Behavioral Data	119
4.2 Multimodal Meal-Level Data and Behavioral Feature Construction	121
4.3 Representation Learning and Phenotype Discovery	124
4.4 Results and Discussion	126
4.5 Conclusion and Future Work	129
5 Active Learning	130
5.1 Technical Challenge: Scalable Wearable Behavioral Sensing Under Limited Annotation Budget	130

	12
5.2 Active Learning Framework for Informative Sample Selection	134
5.3 Evaluation and Results	142
5.4 Conclusion	145
6 Discussion	146
6.1 Overview	146
6.2 From Episodic Assessment to Continuous Behavioral Risk Modeling . . .	147
6.3 From Risk Prediction to Proximal Overeating Detection	148
6.4 From Detection to Interpretable Digital Phenotypes	149
6.5 Scalability and the Annotation Bottleneck	150
6.6 Scientific Implications	151
6.7 Practical Implications for Precision Behavioral Health	151
6.8 Limitations	152
6.9 Future Directions	154
7 Conclusion	156
References	159

List of Figures

2.3.1	Developing a generalizable and explainable weight loss prediction machine learning model for healthcare providers. a The SMART weight loss trial data was cleaned before extracting measures, which were used to train an ML model for accurate and explainable predictions. The model was continually refined, and its performance was assessed on external datasets, such as Opt-IN and ENGAGED to evaluate generalizability. The goal is to develop a reliable decision-support tool for healthcare providers through ongoing model refinement and data preparation. b Raw weight data were obtained from the SMART study, then cleaned and pre-processed. Static measures, including demographics and baseline clinical characteristics, as well as dynamic measures, consisting of statistical features, were extracted for use in training the models. Following this, the machine-learned model was trained on the extracted features and tuned its hyperparameters. The prediction model was then evaluated using well-validated evaluation metrics. This iterative cycle ensures continuous improvement of the model by re-evaluating and tuning until optimal performance is achieved. Using an explainability tool, e.g., SHAP, the refined model is analyzed to identify the features that drive its predictions, enhancing the transparency and interpretability of the model's decisions. The final model was deployed on other datasets, Opt-IN and ENGAGED, to evaluate the model's generalizability. .	38
2.4.1	Explainability plots by SHAP for the features used in the study. a Feature importance of the variables used in the random forest model trained on the SMART dataset. b Correlation between the feature magnitude and SHAP values, indicating the direction and impact of features on the model's predictions.	44
2.4.2	Generalizability assessment of a random forest model for weight loss prediction using external datasets (Opt-IN and ENGAGED). (Full description continued on next page.)	45

2.5.1 Comparative analysis of weight loss predictions: random forest vs. logistic regression models across initial two-week trajectory. (Description continued on next page.)	49
2.7.1 Study profile for SMART study. The figure provides a visual representation of the number of participants in the SMART Weight Management Study. It illustrates the total number of study participants, the count of individuals excluded from our analysis, and the overall number of participants considered in our analytical assessment.	56
2.7.2 Study profile for Opt-IN study. The figure provides a visual representation of the number of participants in the Opt-IN Weight Management Study. It illustrates the total number of study participants, the count of individuals excluded from our analysis, and the overall number of participants considered in our analytical assessment.	56
2.7.3 Study profile for ENGAGED study. The figure provides a visual representation of the number of participants in the ENGAGED Weight Management Study. It illustrates the total number of study participants, the count of individuals excluded from our analysis, and the overall number of participants considered in our analytical assessment.	57
2.7.4 Examples of individuals with high and low weight loss variability. This figure presents the variation in weight change over 14 days from baseline weight for four participants. Successive weight-loss variability was calculated based on the daily weight loss percentages (in blue). To better visualize the overall trend and highlight the variation in weight loss over time, a red linear trendline is superimposed on each plot. This allows for a clearer comparison of weight-loss consistency among the participants, distinguishing those with high variability from those with low variability. (A) Two individuals with high weight variability are shown, calculated using the proposed successive weight loss variability method. (B) Two individuals with low weight loss variability are presented.	58
3.3.1 Analytical pipeline, a step-by-step process of the study's analytical approach. a Conceptual framework for an integrated ML pipeline aimed at identifying and addressing overeating phenotypes in practice. (Continued on next page.)	73

3.4.1	Consort diagram. This figure presents the participant flow through the study, from initial recruitment to final analysis, showing the number of participants at each stage (screened, enrolled, and included in the analysis) and the reasons for exclusions at various points.	78
3.4.2	AUROC, AUPRC, and Brier Score for XGBoost compared to SVM and Naïve Bayes. Performance of XGBoost, SVM, and Naïve Bayes models across three data scenarios: a EMA-only data, b Passive sensing-only data, and c Feature-complete data. Each graph shows the AUROC, AUPRC, and Brier score loss for each model to allow comparison of performance across different data inputs and modeling approaches.	82
3.4.3	SHAP plots illustrating the feature importance and impact on XGBoost model output for different datasets. a EMA-only, b Passive sensing-only, and c Feature-complete data combining EMA and passive sensing.	83
3.4.4	Polar bar plot illustrating the z-scores of features associated with distinct overeating phenotypes. The radial axis represents the magnitude of the z-scores, with bold labels indicating features with a $ z \geq 1$. The circumference is divided into five phenotype clusters, each represented by a unique color, highlighting the key features that differentiate these clusters. This visualization provides insights into the characteristic features of each overeating phenotype, such as psychological factors, contextual influences, and behavioral patterns, enabling a deeper understanding of their unique profiles.	85
3.7.1	Distribution of meal counts across participants for EMA-only and feature-complete datasets. (a) Meal count distribution for the EMA-only dataset, with a mean of 48 meals per participant. (b) Meal count distribution for the feature-complete dataset, with a mean of 17 meals per participant. . .	105
3.7.2	Hourly distribution of meals across the day for all participants. (a) Distribution based on EMA-only data. (b) Distribution for meals with feature-complete set (EMA and passive sensing features). The blue bars represent meal counts per hour, while the red dashed line shows the cumulative percentage.	106
3.7.3	Mean responses to UBS inventory for the SenseWhy Camera wearable. Sub-scale composite means are given in color, i.e., the Difficulty of Use subscale (comprising questions 1–4) composite mean is represented by the height of the red box on the y -axis and is also given in the red text above the box. .	107

- 3.7.4 **Performance evaluation of the XGBoost model on validation and test sets using ROC curves (left), Precision–Recall curves (center), and Calibration curves (right) for EMA-only (a), passive sensing-only (b), and feature complete (c) datasets.** The similar AUROC, AUPRC, and Brier score loss values between the validation and test datasets indicate generalization and minimal overfitting. Shaded regions represent the standard deviation across iterations. 108
- 3.7.5 **Calibration curve for the XGBoost classifier after applying Platt’s scaling, a post-calibration method using a sigmoid function to refine predicted probabilities.** This approach fits a logistic regression model to the classifier’s outputs, improving the alignment of predicted probabilities with actual class distributions. The analysis reveals a reduction in the Brier score from 0.114 ± 0.017 (pre-calibration) to 0.103 ± 0.015 (post-calibration), indicating enhanced model calibration and reduced error in probability estimates. 109
- 3.7.6 **Silhouette plot illustrating the silhouette coefficient values for all meals across all clusters derived from K-Means clustering.** Each bar represents the silhouette score for individual samples within a cluster, sorted by decreasing score. The red dashed line indicates the average silhouette score, providing a measure of overall clustering quality. 109
- 3.7.7 **Overeating clusters visualization in 2 dimensions using UMAP.** Each cluster is represented by a unique color, with clusters 1 through 5 labeled. Overeating meals are highlighted as black points, showing their distribution across clusters. This visualization highlights the clustering patterns and the association of overeating behaviors within the feature space. 110
- 3.7.8 **SHAP value analysis for features associated with overeating behavior (i.e., number of chews, number of bites, and chew-bite ratio).** (a) SHAP values for the number of chews, colored by meal duration, indicating the influence of prolonged chewing patterns on overeating. (b) SHAP values for the number of chews, colored by bite rate, highlighting the interaction between chewing intensity and bite rate. (c) SHAP values for the number of bites, demonstrating their contribution to predicting overeating. (d) SHAP values for the chew-to-bite ratio, showing its inverse relationship with overeating. 111

- 3.7.9 The design and usage of the sensing suite, optimized for user comfort and functionality.** The device has dimensions of 88 mm × 46.86 mm × 45.22 mm (length × width × height) and a weight of 106 grams, making it compact and lightweight for everyday use. Its total volume is approximately 102.18 cm³. To ensure consistent placement, the sensing suite is equipped with an adjustable neckband, allowing for precise calibration to maintain stable positioning on the user's body. Additionally, the device features a magnetic back plate for secure attachment and flexibility in adjustments. These design elements enable users to fine-tune the device's position, ensuring optimal alignment with the body and minimizing movement artifacts during data collection. This design allows the device to remain stable and functional in various real-world scenarios, contributing to the robustness and reliability of the collected data. The right panel of the figure demonstrates how the device is worn on the body, while the left panel provides detailed dimensions and a technical schematic of the sensing suite. 112
- 3.7.10 Eating profile for 6 participants over 14 days with each dot representing a single meal.** The blue dashed line indicates the mean caloric intake, while the red dashed line represents the overeating threshold. The plot captures the variability and frequency of meals exceeding the overeating threshold, displaying individual differences in eating patterns. The overeating thresholds for each panel are as follows: (a) 869.04 kCal, (b) 1027.24 kCal, (c) 944.29 kCal, (d) 653.10 kCal, (e) 975.63 kCal, and (f) 1104.77 kCal. 113
- 4.2.1 Multimodal pipeline for overeating phenotype discovery:** (A) Passive sensing, EMA, and dietitian-verified meals enable rich data collection. (B) Supervised contrastive learning and UMAP uncover latent clusters. (C) Key features ($|z\text{-score}| \geq 1$) profile behavioral differences. 122
- 5.2.1 Uncertainty Visualization through Data Map for the Necksense Dataset.** Here we see that the easy-to-learn segments are those with high confidence, low variability, and high correctness in model prediction. Hard-to-learn segments are typically those with lower confidence and correctness in model prediction. As shown in the distribution plots, the majority of the segments are considered easy-to-learn. 135

5.2.2 Uncertainty Visualization by Data Map and AUM statistic distribution for hard and easy chewing intervals. a) Here we see a snippet from long chewing signals, one that is hard-to-learn (blue) and one that's easy-to-learn (red); b) Here the algorithm extracted 144 candidate segments from the hard-to-learn and 80 segments from the easy-to-learn, and we see the distribution of the segments; c) Here we see the Data Map visual for the easy-to-learn signal (represented by stars), and hard-to-learn signal (represented by squares).	136
5.2.3 Part <i>i</i> is the traditional active learning pipeline with uncertainty sampling strategy (CME). Part <i>ii</i> is the AUM-only method which selects α instances with AUM values. Part <i>iii</i> shows the proposed active learning framework (CME+AUM)	138
5.3.1 ActiveSense Result on Necksense dataset	144

List of Tables

2.4.1 Baseline characteristics of the individuals who participated in the SMART study (development dataset), Opt-IN and ENGAGED studies (generalizability datasets)	40
2.4.2 Comparison of the proposed random forest model with three reference methods	42
2.4.3 Results of generalizability for the model trained using the SMART dataset when applied to the external cohorts, Opt-IN, and ENGAGED	42
2.7.1 Descriptions of Body Weight Variability Operational Definitions in Difference Studies	59
2.7.2 Performance of ML Models at Baseline	60
2.7.3 Feature Description Table	61
3.4.1 Baseline Demographics and EMA Response Summary	79
3.4.2 Phenotype interpretation	87
3.7.1 Summary statistics for the analytic subsample	114

3.7.2 Thematic analysis of open-ended post-study question asking participants for suggestions to improve the SenseWhy Camera wearable. Codes represent specific improvements proposed for the device. Sets of codes with similar user/design considerations are categorized together under broader themes. Themes are adapted from UBS subscales, helping us interpret quantitative UBS results to inform design priorities. Note on adaptation: UBS subscale scopes are reduced in accordance with participant provided codes, e.g., UBS subscale “Time and Social” became the theme “Social” because no participant response related to time-based device burden.	116
3.7.3 Feature descriptions	117
4.4.1 Phenotype Profiles of Overeating Clusters	127

Chapter 1

Introduction

1.1 Motivation

Health-risk behaviors are among the most important modifiable contributors to long-term disease burden. Behaviors such as poor dietary intake, physical inactivity, tobacco use, excess alcohol consumption, and inconsistent engagement with behavioral treatment do not occur as isolated events; they unfold through repeated decisions, routines, affective states, environmental exposures, and social contexts in daily life [1], [2], [3]. These behaviors are often clustered, meaning that risk is not only a function of one behavior but of patterns of co-occurring behaviors that emerge over time [3]. For precision behavioral health, this creates a fundamental measurement challenge: the behaviors most relevant to disease prevention and treatment are longitudinal, contextual, and dynamic, while many clinical and research assessments remain episodic, retrospective, or burdensome.

Traditional behavioral assessment methods, including clinic visits, retrospective surveys, dietary recalls, self-monitoring logs, and laboratory-based observation, have generated substantial knowledge about health behavior. However, they are limited in

their ability to capture behavior as it naturally occurs. Retrospective self-report is vulnerable to recall bias, social desirability bias, and missing contextual information, while laboratory studies often sacrifice ecological validity for experimental control. Ecological Momentary Assessment (EMA) was developed to address some of these limitations by repeatedly sampling behaviors, experiences, and contexts in real time and in naturalistic settings [4]. EMA improves temporal resolution, but it remains an active form of measurement that depends on participant response and can impose burden over long deployments.

Wearable and mobile sensing technologies offer a complementary approach. Smartphones, wearable cameras, inertial sensors, physiological sensors, and other passive sensing systems can capture behavioral signals continuously and in context, reducing sole reliance on recall-based measurement [5], [6]. Digital phenotyping extends this idea by using data from personal digital devices to quantify individual-level behavior and health-related states in situ [5]. In health behavior research, wearable cameras and image-assisted dietary assessment have shown the potential to improve objective measurement of lifestyle behaviors and dietary intake, while also exposing new challenges related to privacy, data management, annotation burden, and interpretation [7], [8], [9].

This dissertation is motivated by the gap between what sensing systems can collect and what behavioral health systems need to know. Continuous data streams are valuable only if they can be transformed into actionable information: early indicators of treatment risk, objective measures of proximal behaviors, interpretable behavioral phenotypes, and scalable models that can be trained without excessive manual annotation. This requires a computational pipeline that links sensing, prediction, interpretation, and scalability.

1.2 From Continuous Monitoring to Early Risk Identification

A first step toward precision behavioral health is identifying risk early enough for intervention to be adapted. In behavioral weight management, early response is especially important because initial treatment trajectories often predict later outcomes. Stepped-care interventions are built around this principle: treatment intensity can be adjusted when early signs suggest that a participant is unlikely to respond adequately to the current intervention [10]. However, many stepped-care approaches rely on simple decision rules, such as whether a participant has crossed a fixed weight-loss threshold by a given time point. These rules may be interpretable, but they often fail to capture the richer structure of early behavioral trajectories.

Continuous self-monitoring provides a way to move beyond single-point thresholds. Daily or near-daily weight data can encode not only the amount of early weight change but also variability, slope, distributional features, missingness, and consistency of engagement. Prior work has shown that early weight loss is associated with longer-term weight outcomes in both conventional and mobile health interventions [11], [12]. These findings support the idea that early behavioral signals can serve as decision-relevant markers for adaptive treatment.

In *Chapter 2*, this dissertation develops an explainable machine learning framework for early risk prediction in behavioral weight management. The chapter uses dynamic features extracted from early longitudinal weight trajectories, combined with baseline characteristics, to predict longer-term treatment response. This work establishes the first component of the dissertation’s pipeline: continuous monitoring can support earlier and

more individualized risk stratification than delayed or retrospective assessment alone.

1.3 From Risk Identification to Proximal Behavior

Detection

Early risk prediction can identify who may be at risk, but it does not fully explain why risk emerges. Behavioral intervention requires insight into proximal mechanisms: the momentary behaviors, contexts, and psychological states that contribute to later outcomes. In the context of weight management, overeating is one such proximal behavior. It is influenced by hunger, stress, affect, hedonic motivation, loss of control, social context, environmental cues, and eating microstructure [13], [14], [15]. Because these factors fluctuate across meals and contexts, overeating is difficult to characterize through retrospective self-report alone.

Wearable sensing has created new opportunities for objective assessment of eating behavior. Prior work has shown that wrist-worn inertial sensors can detect eating moments and model in-meal eating microstructure, including food intake cycles and wrist micromovements [16], [17]. Image-assisted and wearable-camera-based methods can also improve dietary assessment by providing objective visual evidence of eating episodes and food context [7], [9]. However, no single modality fully captures the complexity of overeating. Self-report can capture subjective states such as hunger, stress, and loss of control, while passive sensing can capture objective features such as bites, chews, eating duration, and eating pace.

In *Chapter 3*, this dissertation addresses this measurement problem through the SenseWhy study. The study integrates wearable camera sensing, EMA, mobile dietary

reporting, and dietitian-administered dietary recalls to characterize overeating in free-living conditions. Overeating is operationalized at the personalized meal level rather than through a fixed calorie threshold, allowing each meal to be interpreted relative to an individual's usual intake. Supervised machine learning models are then used to examine how psychological, contextual, and passive sensing features contribute to overeating detection. This chapter establishes the second component of the dissertation's pipeline: multimodal sensing can detect proximal behavioral episodes and reveal the behavioral and contextual factors associated with them.

1.4 From Behavior Detection to Digital Phenotyping

Detecting overeating episodes is necessary but insufficient for personalized intervention. The same behavioral outcome may arise through different mechanisms. For example, overeating may occur in response to stress, evening hunger, social dining, restaurant-sourced meals, task-oriented distraction, or rapid eating. Treating overeating as a homogeneous category can obscure these differences and limit the specificity of intervention strategies. A more useful approach is to identify interpretable behavioral phenotypes that describe how and under what conditions overeating occurs.

Digital phenotyping provides a computational framework for this task. By using dense, longitudinal, and multimodal data, digital phenotyping aims to characterize individual-level behavior in naturalistic settings [5]. Representation learning methods can support this process by transforming high-dimensional behavioral data into latent spaces where meaningful structure becomes easier to identify. Supervised contrastive learning, for example, uses label information to pull together samples from the same class while separating samples from different classes in embedding space [18]. Dimensionality reduc-

tion methods such as UMAP can then support visualization and downstream clustering of learned representations [19].

In *Chapter 4*, this dissertation extends overeating detection toward meal-level digital phenotyping. The chapter uses multimodal meal-level data to learn latent representations of eating episodes and identify distinct overeating phenotypes. These phenotypes are interpreted using psychological, contextual, and sensor-derived features, linking computational clusters back to clinically meaningful behavioral patterns. This chapter establishes the third component of the dissertation’s pipeline: sensing systems should not only classify behavior, but also reveal interpretable behavioral subtypes that can inform personalized intervention design.

1.5 From Manual Annotation to Scalable Behavioral AI

A central barrier to scalable sensing-based behavioral health systems is the cost of labeled data. Wearable and mobile systems can produce large volumes of video, inertial, physiological, and contextual data, but supervised models often require detailed annotation. In eating behavior research, this may include labeling meal boundaries, bites, chews, gestures, or contextual features. Such annotation is time-intensive, expensive, and difficult to scale across participants, behaviors, devices, and deployment settings.

Active learning addresses this bottleneck by selecting the most informative samples for annotation rather than labeling data randomly or exhaustively [20]. In behavioral sensing, this is particularly relevant because many segments are redundant, easy to classify, or uninformative, while others are ambiguous, difficult, or valuable for improving model decision boundaries. Training dynamics provide another way to characterize examples by examining how a model behaves on each sample across training epochs.

Dataset cartography has shown that confidence and variability across training can identify easy-to-learn, hard-to-learn, and ambiguous examples, offering a principled way to reason about data quality and sample utility [21].

In *Chapter 5*, this dissertation develops an active learning framework for wearable sensing data. The framework combines model uncertainty with intrinsic uncertainty derived from training dynamics to select samples that are informative for annotation. This chapter establishes the fourth component of the dissertation’s pipeline: scalable behavioral AI requires not only accurate models, but also efficient strategies for deciding which data should be labeled.

1.6 Dissertation Overview and Major Contributions

Together, the chapters of this dissertation form a sensing-to-insight pipeline for precision behavioral health. *Chapter 2* demonstrates how continuous self-monitoring data can support early risk prediction in behavioral weight management. *Chapter 3* shows how multi-modal sensing and EMA can detect proximal overeating episodes in daily life. *Chapter 4* transforms meal-level behavioral data into interpretable overeating phenotypes. *Chapter 5* addresses the annotation bottleneck that limits the scalability of wearable sensing systems.

This dissertation makes the following major contributions:

1. **An explainable early-risk prediction framework for behavioral weight management.**

This work shows that dynamic features extracted from early longitudinal weight trajectories can support prediction of longer-term treatment response and enable earlier identification of individuals at risk for suboptimal outcomes.

2. A multimodal sensing framework for detecting overeating in free-living conditions.

Through the SenseWhy study, this dissertation integrates wearable camera sensing, EMA, dietary reporting, and dietitian-administered recalls to characterize overeating at the personalized meal level.

3. A digital phenotyping approach for identifying heterogeneous overeating patterns.

This work moves beyond binary overeating detection by learning latent meal-level representations and identifying interpretable overeating phenotypes defined by psychological, contextual, and sensor-derived behavioral features.

4. A label-efficient active learning framework for scalable wearable sensing.

This dissertation develops an active learning strategy that uses model uncertainty and training-dynamics-based uncertainty to reduce annotation burden while supporting behavioral model development.

5. A unified computational framework for precision behavioral health.

Across these contributions, this dissertation shows how continuous monitoring, multimodal sensing, interpretable machine learning, digital phenotyping, and active learning can be integrated to move behavioral health research from retrospective assessment toward timely, personalized, and scalable intervention.

Chapter 2

Early Risk Prediction

2.1 Technical Challenge: Early Identification of Treatment Risk

Research Question 1

How can early longitudinal self-monitoring signals be transformed into generalizable and explainable indicators of behavioral treatment risk?

Obesity is a growing population health concern and is among the leading preventable causes of premature death globally.[22] Gold-standard treatment for obesity involves intensive behavioral treatment that can be costly and burdensome to administer, impeding the scalability needed to redress obesity on a global scale.[23] Stepped-care treatment models may be a viable solution. A rational resource allocation strategy, stepped-care, involves starting with a less resource-intensive intervention, and, for those who respond suboptimally, stepping them up by adding more vigorous, expensive intervention components. A major challenge facing obesity stepped-care intervention, however, is lack of a predictive treatment algorithm (policy) that accurately identifies non-responders early in

treatment. Improved early detection of non-response to weight loss treatment will guide stepped-care decisions for treatment that can be scaled for population-level impact. This would result in a treatment allocation algorithm (policy) that gives people the treatment resources they need—not more and not less.[24]

Prior stepped-care obesity treatment approaches have had two main limitations: a) poor accuracy of the algorithm predicting nonresponse, or b) prolonged delay (2–3 months) before augmenting treatment,[10] a problematic design feature because initial weight loss predicts long-term weight loss. There is disagreement as to whether weight loss variability and other statistical measures increase predictive power of weight loss. Moreover, it is unknown how we operationalize initial weight loss, weight loss variability, trends, and other statistical measures. Given the recent adoption of mobile health tools, such as wireless weight scales, we can further identify, with greater granularity to daily changes, critical predictors for stepped-care treatment interventions. Creating adequate tools for early prediction of weight loss treatment success prevents researchers from resorting to inaccurate weight loss prediction algorithms or subjective judgment. SHapley Additive exPlanations (SHAP)[25] is an explainable AI tool that helps us understand the direction and magnitude of the marginal contribution of each feature to the weight loss outcome. Explainable AI is increasingly enhancing clinician understanding of machine-learning (ML) models and augmenting their decision-making processes. By improving transparency in model predictions, it increases clinicians’ trust in the outcome of ML models and advances their integration into clinical practice.

Although conventional statistical methods remain a powerful predictive tool, recent weight loss research has developed ML-based prediction models using longitudinal data that uncover new knowledge that can improve the models’ predictive performance.[26],

[27] Our study had two primary aims. First, we aimed to build a supervised ML model that incorporates both dynamic features obtained during the initial treatment period combined with baseline features to improve the predictive utility of a clinical decision rule that identifies non-responders in the first 2 weeks of a stepped-care weight loss trial. Second, we aimed to evaluate the generalizability of our model by testing it on populations from other weight loss trials.

2.2 Longitudinal Self-Monitoring Data and Prediction Task

2.2.1 Overview

Our data analysis pipeline aimed to demonstrate whether dynamic features (weight-loss information obtained early in the intervention) improve the early prediction of weight loss in an interventional weight management program, select the best-performing predictive model, and test its explainability and generalizability. Participants' weights in the SMART study were objectively measured using the Fitbit Aria wireless smart scale. They were instructed to weigh themselves every morning immediately after waking up, post-urination, and without any clothing. The weight loss outcome, operationalized as a dichotomous variable, is assessed at month 6, with success defined as participants achieving a clinically meaningful weight loss of 5% or more from their baseline weight by the 6-month endpoint. We compared our machine-learned model with several reference methods. The first was the clinical decision rule in which non-responders were defined as participants who lost <0.5 pounds/week (0.227 kg/week). The clinical decision rule is operationalized by evaluating participants at the end of week 2 and flagging them as non-responders if they did not lose at least 1 pound from their baseline weight at the end of the 2-week period. Because most prior work suggests the best predictor of weight loss success

is the degree of initial (early) weight loss, our second reference method included a logistic regression based on weight loss at week 2. This allowed us to move towards a more tailored personalized approach to assess whether collecting a feature such as weight loss from baseline at week 2 (expressed as a percentage)– temporally closer to the 6-month endpoint –increased the predictive utility of our model. Our third reference method was the baseline machine-learned model, which tests the predictive power of static features obtained from participants (e.g., demographics, psychological measures, etc.) at baseline using a supervised machine-learned model. This was essential to assess whether data prior to beginning the intervention was sufficient for prediction of weight loss. Our final proposed weight loss machine-learned model combined baseline static variables with dynamic statistical variables obtained early (first 2 weeks) during the weight loss treatment period to highlight the predictive power of data-driven statistical features. To evaluate the in-distribution generalizability of the final model, we utilized the randomly selected holdout test set from the SMART study. With the goal of building a trustworthy machine-learned model that could reliably augment clinical practice, we used explainability methods to uncover not only the most predictive variables but also the magnitude and direction of a variable’s effect on weight loss. A clinician needs to know the direction and impact by which a model’s features predict weight loss to understand how the model “works,” and thereby to decide whether it is credible and trustworthy. To evaluate the out-of-distribution generalizability of the final model (SMART-trained), we assessed its performance on the Opt-IN and ENGAGED datasets and compared it to models trained separately on these datasets, allowing us to gain insights into its ability to generalize despite differences in study design and measurement procedures. Moreover, given our interest in understanding behavior change and the existing uncertainty regarding weight variability and its association with long-term weight loss, we propose a new definition for body weight variability incorporating the factor missing rate and gauge its discriminative

power and relationship with weight loss at the 6-month mark.

2.2.2 SMART study design

The SMART study[28] was a sequential multiple assignment randomized trial that aimed to optimize an mHealth intensive lifestyle obesity stepped-care treatment package by identifying the optimal starting condition and the optimal augmentation (“step-up”) strategies for treatment non-responders. The study included a 3-month weight loss treatment program with a weight loss outcome measure at 6 months. The sample included 400 adults who were overweight or obese at baseline ([BMI], 27–45 kg/m²). Participant recruitment spanned 2016 to 2021 in the Chicagoland area.

2.2.3 Opt-IN study design

The Opt-IN study[29] was a clinical trial that used a full factorial design to identify the combination of remotely delivered, technology-supported weight loss treatment components that maximized weight loss over a 6-month period. The study included 562 adult participants with overweight or obesity. All participants received a base treatment intervention including a custom-built smartphone app that facilitated self-management of diet and activity behaviors, online lessons, and a coach. In addition to the base package, participants were randomly assigned to 32 experimental conditions, covering all possible permutations of five treatment components: moderate vs intense (12 vs 24) number of coaching call sessions, text messaging, buddy (social support) training, primary care provider engagement, and meal replacement. This allowed researchers to identify the optimal combination of these treatment components in terms of the primary outcome, weight loss achieved at 6 months. Participant recruitment spanned 2013 to 2017 in the Chicagoland area, with the study’s protocol and design previously documented.

2.2.4 ENGAGED study design

The ENGAGED study[30], conducted between 2009 to 2013, was a randomized controlled trial involving 96 adults with obesity, comparing three different 6-month weight loss treatments: Self-Guided (SELF), Standard (STND), and Technology-Supported (TECH). The STND and TECH groups attended eight in-person group sessions, while the SELF and STND groups used paper diaries for self-monitoring. In contrast, the TECH group utilized a smartphone app with social networking features and a wireless accelerometer for self-monitoring. The primary goal of the study was to evaluate the effectiveness of each treatment, measured by the amount of weight loss achieved after 6 months.

2.2.5 Development and External Cohorts

We selected the SMART study as the development cohort due to its accurate and objective weight measurement, facilitated by the Fitbit Aria wireless smart scale, compared to the self-reported weight measurements in the other studies. Additionally, the SMART study was the most recent among the weight loss studies, ensuring more up-to-date methodologies and technologies. Since part of the SMART study was run during the COVID-19 pandemic, a post hoc sensitivity analysis was conducted, showing no significant difference in weight loss between groups assessed before and during the lockdown.[31] Opt-IN and ENGAGED were chosen as external cohorts to deploy the SMART-trained model and test its generalizability.

2.2.6 Outlier Detection Architecture

We removed duplicate records and eliminated erroneous observations with invalid formats in each study separately. In the first step, we leveraged a traditional statistical ap-

proach to remove explicit outliers (i.e., unreasonably high/low weight records) by applying a within-person threshold (lower bound = $Q1 - 1.5 * \text{interquartile range [IQR]}$; upper bound = $Q3 + 1.5 * \text{IQR}$, where $Q1 = 25\text{th percentile}$ and $Q3 = 75\text{th percentile}$). We supplemented this approach with another validated method, time-windowed geometric path analysis to identify within-person errors in weight measurements in time series data.[32] We measured the time interval and weight difference between three consecutive weight measurements for each participant. Subsequently, for every three contiguous weight measurements, we computed the path ratio. We created a personalized distribution of path ratios for each individual and excluded weight values that had a path ratio that exceeded five z-scores, empirically set, and verified through visual confirmation by two authors (FS and NA) to ensure accuracy and reliability in the analysis. This method was applied uniformly to all datasets (SMART, Opt-IN, and ENGAGED). Consequently, all baseline models and ML methods utilized the same cleaned dataset, ensuring consistency and comparability in the analysis.

2.3 Early-Trajectory Feature Engineering and Risk Modeling

2.3.1 Measures and Features

We extracted data-driven features to assist the ML models in discriminating between individuals who achieved weight loss success and those with suboptimal weight loss results. Features were added to create generalizable and robust machine-learned models that identify early non-response in the first 2 weeks of a weight loss intervention. Two baseline psychological features (self-efficacy for diet and for exercise) and four baseline demographic variables (height, age, gender, race/ethnicity) were included in the model.

The data-driven features included daily weight deviation ratio from baseline weight for days 7 and 14, weight variability, ranges and means of weight loss during the first 2 weeks, and slope/intercept (obtained from linear regression), zero crossing, kurtosis, and skewness, which reveal weight distribution characteristics over the 14-day course. Formal definitions of the features are in Supplementary Table 2.7.3. To handle missing data, we used linear regression as a curve-fitting method[33] to estimate and impute missing weight data from the known records over 2 weeks. For instance, if a weight on day 14 was missing or removed due to the outlier detection architecture, we used the imputed value.

2.3.2 Weight Variability

Previous research has used different methods to define/measure weight variability; however, their impact on weight/health outcomes remains unknown. Some researchers have derived weight variability as the coefficient of variation known as the relative standard deviation.[34] Others have defined weight variability as the mean successive weight change.[35] Another weight variability metric is derived from non-linear mean deviation estimated from locally weighted scatterplot smoother (LOESS) regression.[36] However, weight variability is also widely used as root mean square error (RMSE) calculated from the estimated weight records' distance from the best-fitted linear regression line.[37], [38] All derivations are in Supplementary Table 2.7.1. We propose a sparsity-adjusted successive weight loss variability estimated by equation (1), where we measure the within-person successive (day-to-day) weight loss variability and adjust by the count of missingness over 2 weeks, mathematically formalized as follows:

$$\sigma_i^2 = \frac{\sum_{j=1}^{M-1} (y_{i,j} - y_{i,j+1} - \mu_i)^2}{M}, \quad \mu_i = \frac{\sum_{j=1}^{M-1} (y_{i,j} - y_{i,j+1})}{M}, \quad SASWV_i = \sigma_i^2 * M_i \quad (2.1)$$

where $M = 14$ is the number of observations, $j \in \{1, 2, \dots, M\}$, and N is the number of participants, $i \in \{1, 2, \dots, N\}$, μ_i is the within-person mean of the successive weight loss differences, σ_i^2 is the variability of the within-person successive weight loss difference and M_i is the count of missingness (i.e., the number of imputed observations) over 2 weeks (Supplementary Figure 2.7.4).

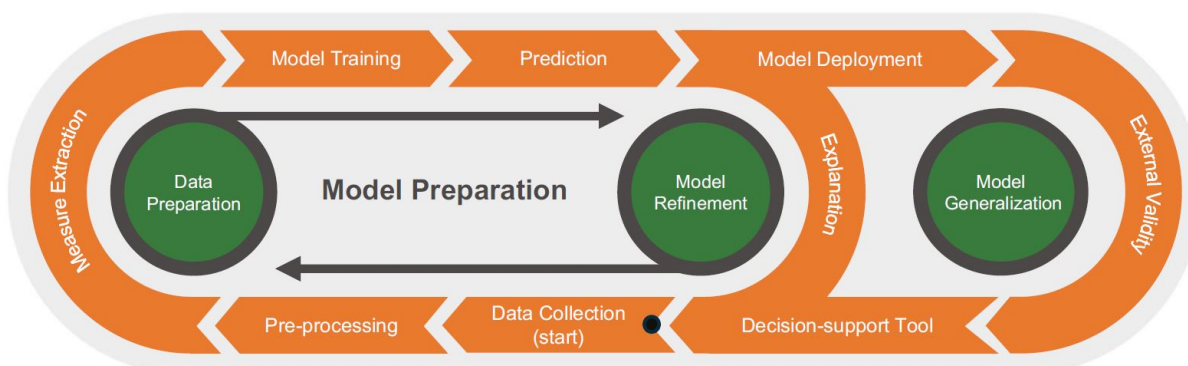
2.3.3 Model preparation

We tested supervised ML algorithms to identify early non-response and predict suboptimal weight loss at month 6 in the SMART study at baseline and by week 2. To identify the optimal model, we ran discriminative (e.g., XGBoost, random forest, support-vector machines, logistic regression, and k-nearest neighbors) and generative (e.g., naive bayes) supervised ML classifiers. Figure 2.3.1 displays the analytic pipeline, with each component described in detail in the following sections.

2.3.4 Model generalizability and explainability

To test in-distribution generalizability of the machine-learned models, we created a train/-validation/test split (60%/20%/20%), ensuring that the data was randomly partitioned without cross-contamination between sets. In the model training and refinement phase, we used Bayesian optimization[39] to select the optimal set of parameters for each algorithm on the validation set in an efficient and informed manner. To address sample diversity and class imbalance, we used sample stratification (based on gender, race/ethnicity, and outcome variable) to eliminate bias and model overfitting. Finally, we evaluated in-distribution generalizability by assessing the performance of the fine-tuned model (trained on the combined train and validation sets) on the holdout test set. Based on these evaluations, we selected the best-performing machine-learned model and proceeded to the model de-

(a)



(b)

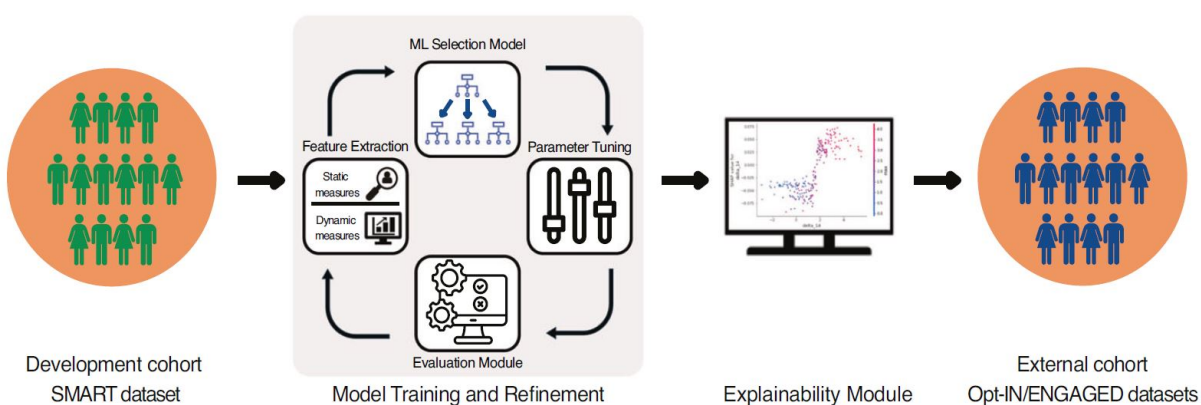


Figure 2.3.1: Developing a generalizable and explainable weight loss prediction machine learning model for healthcare providers. **a** The SMART weight loss trial data was cleaned before extracting measures, which were used to train an ML model for accurate and explainable predictions. The model was continually refined, and its performance was assessed on external datasets, such as Opt-IN and ENGAGED to evaluate generalizability. The goal is to develop a reliable decision-support tool for healthcare providers through ongoing model refinement and data preparation. **b** Raw weight data were obtained from the SMART study, then cleaned and pre-processed. Static measures, including demographics and baseline clinical characteristics, as well as dynamic measures, consisting of statistical features, were extracted for use in training the models. Following this, the machine-learned model was trained on the extracted features and tuned its hyperparameters. The prediction model was then evaluated using well-validated evaluation metrics. This iterative cycle ensures continuous improvement of the model by re-evaluating and tuning until optimal performance is achieved. Using an explainability tool, e.g., SHAP, the refined model is analyzed to identify the features that drive its predictions, enhancing the transparency and interpretability of the model's decisions. The final model was deployed on other datasets, Opt-IN and ENGAGED, to evaluate the model's generalizability.

ployment phase. To assess the generalizability of this final model (the SMART-trained model), we evaluated its performance on external datasets, specifically ENGAGED and Opt-IN. Additionally, we independently trained models on each of these external datasets using k-fold cross-validation. We then compared the cross-validation metrics from these independently trained models to the generalizability metrics obtained from applying the SMART-trained model to the same datasets. We used SHAP to improve the explainability of our model by uncovering the magnitude, directionality (positive or negative), and predictive order of the features.

2.3.5 Evaluation

We evaluated our model using various metrics, including sensitivity, specificity, F1 scores for weight loss success, F1 scores for suboptimal weight loss, weighted F1 scores, the area under the receiver operating curve (AUROC), the area under the precision-recall curve (AUPRC), and Brier score loss. F1 score is a precise measure of performance used in ML to capture precision of the algorithm and recall of both weight loss success and suboptimal weight loss or as a combined weighted F1 score that assigns weights based on each class's support.

2.4 Evaluation and External Validation

Data from 400 participants with obesity were obtained from the SMART study, of which 281 participants were eligible for inclusion in the model development dataset (15 withdrew, 43 missed 6-month follow-up, and 61 did not produce sufficient weight data from the Fitbit Aria wireless weight scale during the first 2 weeks of the study). We allocated 60% (169/281) of the data in the training dataset, 20% (55/281) in the validation set, and 20% (57/281) in the holdout test set. Approximately 64% (179 participants) of the whole

development dataset reached suboptimal weight loss (encoded as zero), and 36% (102 participants) achieved sufficient weight loss (encoded as one) based on the 5% weight outcome at month 6. Table 2.4.1 shows baseline characteristics of patients in the training, validation, and test sets. The profiles for the SMART, Opt-IN, and ENGAGED studies are presented in Supplementary 2.7.1, 2.7.2, and 2.7.3.

Table 2.4.1: Baseline characteristics of the individuals who participated in the SMART study (development dataset), Opt-IN and ENGAGED studies (generalizability datasets)

Sample size	Development dataset [SMART] ($n = 281$)			Generalizability datasets [Opt-IN & ENGAGED] ($n = 472$)	
	Train ($n = 169$)	Validation ($n = 55$)	Test ($n = 57$)	Opt-IN ($n = 442$)	ENGAGED ($n = 30$)
Age range, years	20–60	23–60	18–60	18–60	22–57
Sex, n (%)					
Female	123 (72.8)	40 (72.7)	43 (75.4)	357 (80.8)	27 (90)
Male	43 (25.5)	15 (27.3)	14 (24.6)	85 (19.2)	3 (10)
Other	3 (1.7)	0 (0)	0 (0)	0 (0)	0 (0)
Race/ethnicity, n (%)					
Non-Hispanic White	103 (60.9)	34 (61.8)	35 (61.4)	335 (75.8)	14 (47.7)
Non-Hispanic Black	32 (18.9)	12 (21.8)	11 (19.3)	51 (11.5)	12 (40)
Hispanic	21 (12.5)	5 (9.2)	6 (10.5)	40 (9.1)	4 (13.3)
Non-Hispanic Other	13 (7.7)	4 (7.2)	5 (8.7)	16 (3.6)	0 (0)
Weight loss success, n (%)					
No	104 (61.5)	34 (61.8)	34 (59.6)	209 (47.3)	19 (63.3)
Yes	65 (38.5)	21 (38.2)	23 (40.4)	233 (52.7)	11 (36.7)

Data are n (%) and (range) for baseline characteristics.

2.4.1 Proposed model compared to reference methods

Results for the three reference methods show imbalanced sensitivity and specificity as indicated in Table 2.4.2. The average of the three models yields a low sensitivity of 26.1% (95% CI, 14.1-38.1) and high specificity of 86.2% (95% CI, 82.9-89.4) suggesting a tendency towards higher false negatives (misidentifying actual sufficient weight loss as suboptimal) and lower false positives (misidentifying actual inadequate weight loss as sufficient), implying that the models predominantly predict a single outcome, suboptimal weight loss, akin to a negatively biased decision maker. The baseline ML models, both discriminative and generative, which were trained solely using static features as indicated in the Supplementary Table 2.7.2, achieved an average weighted F1 score of 59.0% (95% CI, 57.6-60.4),

and AUROC of 62.8% (95% CI, 59.7-65.9). The generative and discriminative ML models, when trained on dynamic and static features at week 2, achieved an average F1 weighted score of 68.9% (95% CI, 66.2-71.6) and AUROC 76.0% (95% CI, 72.2-79.9), demonstrating the value of dynamic features in enhancing prediction accuracy shown in Supplementary Table ???. By incorporating dynamic features in addition to static features in the model, we observed a 9.9% improvement in the F1 weighted score and a 13.2% increase in the AUROC. Consequently, we selected the best-performing ML model, the proposed random forest, which uses the complete feature set. This model outperformed all reference methods yielding a weighted F1 score of 75.7%, AUROC of 84.5%, AUPRC of 86.3%, and Brier score loss of 0.173, as indicated in Table 2.4.2. To ensure an unbiased approach to modeling, further analysis was performed on 15 other random train/validation/test splits, yielding similar performance and explainability outcomes (Supplementary Note 2.7.1).

2.4.2 Explainability

Through SHAP, we generated a ranking of feature importance and corresponding explanations using SHAP values and feature magnitudes. Our analysis highlights the significance of certain features. For instance, SHAP magnitudes and directionality support the expectation that a higher magnitude of weight loss on the 14th day increases the likelihood of successful weight loss shown in Figure 2.4.1. Furthermore, we found that features derived from the weight trajectory over a two-week period, such as high minimum, maximum, mean, 75th percentile, IQR, and slope, played a crucial role in weight loss prediction based on SHAP plots. Additionally, by examining the SASWV feature, we observed that higher SASWV was indicative of weight gain. Our examination also revealed multiple baseline features such as age (being older), sex (being male), height (being tall), and high self-efficacy in exercise are predictive of weight loss.

Table 2.4.2: Comparison of the proposed random forest model with three reference methods

Model	Sensitivity, %	Specificity, %	F1 weighted, %	AUROC, %	AUPRC, %	Brier score loss	Time endpoint
Historic clinical decision rule (0.5 lb/week)	5.5	81.4	44.4	NA	NA	NA	Two weeks
Logistic regression (weight loss on day 14)	40.9	85.7	66.5	76.1	70.7	0.217	Two weeks
Logistic regression (Static features)	31.8	91.4	64.8	68.4	70.0	0.239	Baseline
Proposed random forest (full feature set)	82.2	71.1	75.7	84.5	86.3	0.173	Two weeks

The clinical decision rule, which selects responders based on weight loss exceeding 0.5 lb/week, and the logistic regression model trained solely on weight loss data from day 14 on SMART data. The third reference model was trained on static baseline features. Logistic regression outperformed all other ML models, both generative and discriminative, that were trained using static features at the baseline. The fourth model is the proposed random forest model trained on the full feature set by the two-week endpoint.

Table 2.4.3: Results of generalizability for the model trained using the SMART dataset when applied to the external cohorts, Opt-IN, and ENGAGED

SMART-trained model results	Opt-IN	ENGAGED
Sensitivity (%)	77.7	90.9
Specificity (%)	60.3	78.9
F1 weighted (%)	69.2	83.6
AUROC (%)	70.9	80.5
AUPRC (%)	68.7	82.0
Brier score loss	0.218	0.137

2.4.3 Generalizability

The results of our generalizability analysis are indicated in Figure 2.4.2, and Table 2.4.3. The results for Opt-IN revealed that the 5-fold cross-validation procedure yielded an average AUROC of 72.6% (95% CI, 70.3-74.9), AUPRC of 72.3% (95% CI, 68.3-76.3), and a Brier score loss of 0.216 (95% CI, 0.207-0.225). For generalizability of Opt-IN, AUROC, AUPRC, and Brier score loss were 70.9%, 68.7%, and 0.218, respectively. Applying a similar methodology to the ENGAGED dataset, for 3-fold cross-validation, we achieved an AUROC of 84.4% (95% CI, 75.6-93.2), an AUPRC of 79.9% (95% CI, 70.9- 88.9), and a Brier score loss of 0.166 (95% CI, 0.126-0.206). For ENGAGED generalizability, AUROC, AUPRC, and Brier score loss were 80.5%, 82.0%, and 0.137, respectively.

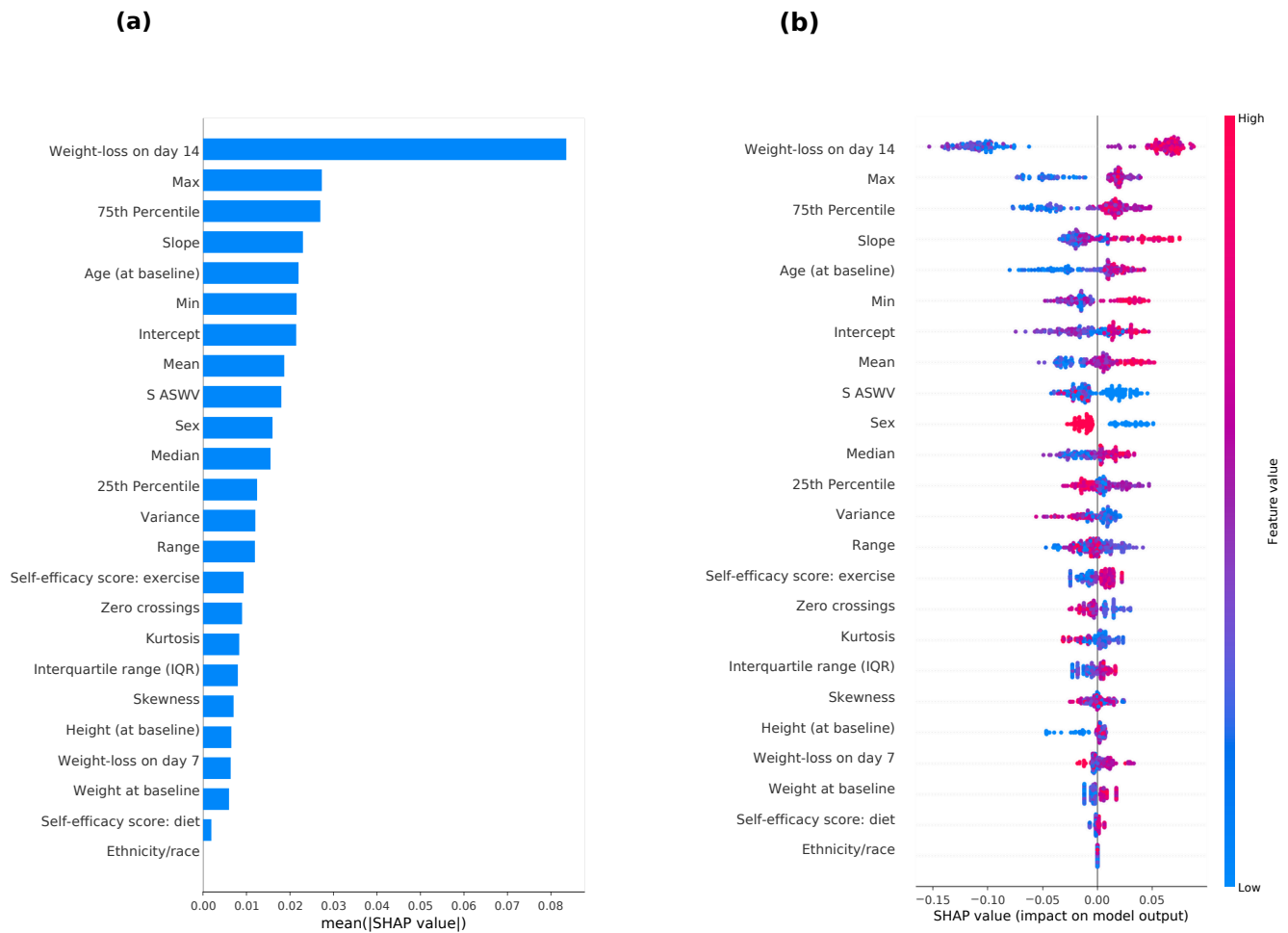


Figure 2.4.1: Explainability plots by SHAP for the features used in the study. **a** Feature importance of the variables used in the random forest model trained on the SMART dataset. **b** Correlation between the feature magnitude and SHAP values, indicating the direction and impact of features on the model's predictions.

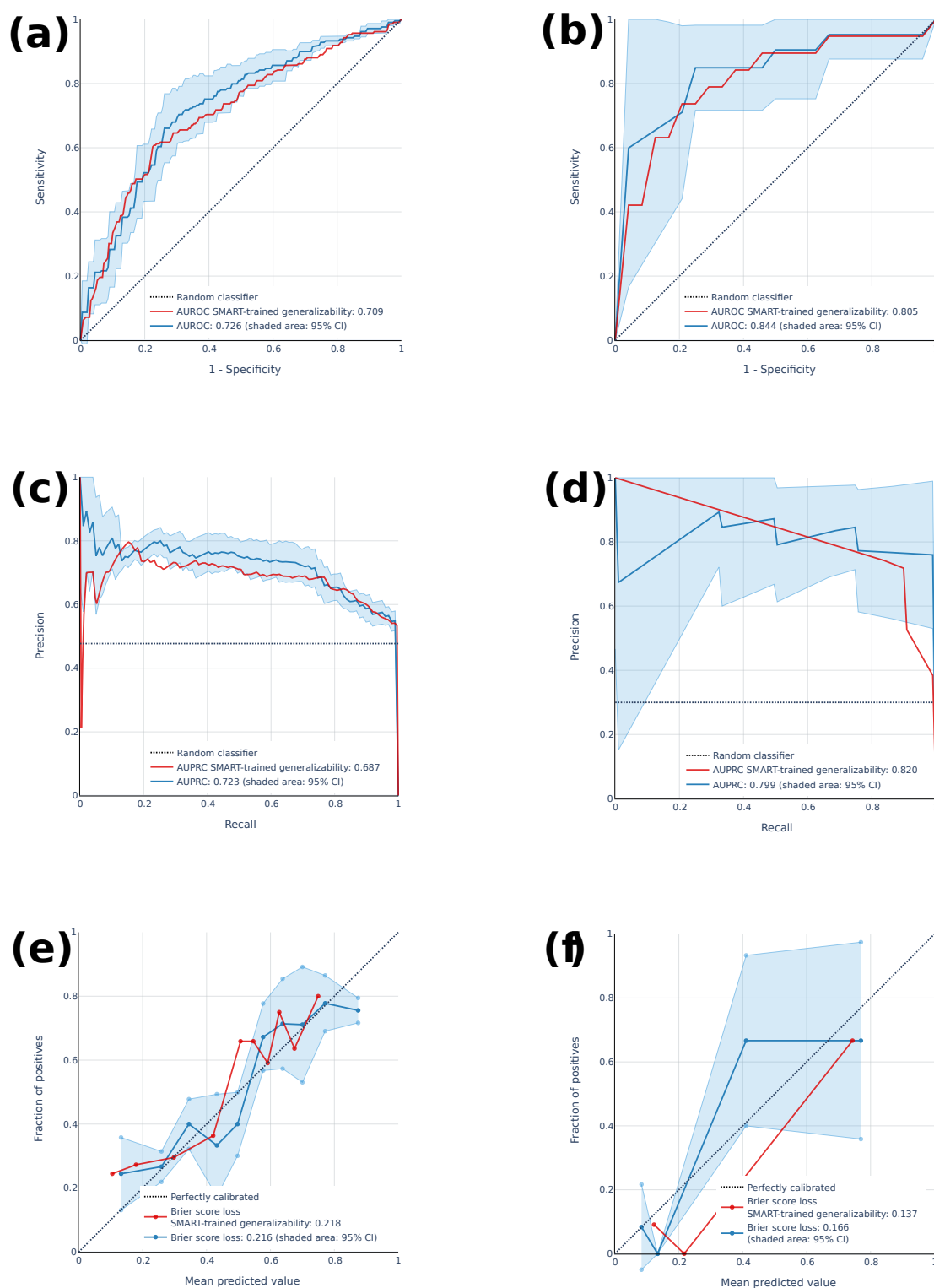


Figure 2.4.2: Generalizability assessment of a random forest model for weight loss prediction using external datasets (Opt-IN and ENGAGED). (Full description continued on next page.)

Figure 2.4.2: (Continued) **a, b** ROC curves for both models, with the blue curve (and a shaded 95% CI) representing the model trained/tested on Opt-In and ENGAGED. The red curve represents the ROC curve for the generalizability. The generalizability curve consistently fell within the 95% CI, achieving an AUROC of 0.709 and 0.805, which closely aligns with the AUROC of the model trained/tested on Opt-IN and ENGAGED, respectively (mean: Opt-IN, 0.726; ENGAGED, 0.844).

c, d Correspondingly, the Precision-Recall curves for the Opt-IN and ENGAGED, based on models trained and tested on these cohorts individually, were compared with those from the SMART-trained model's generalizability performance. The generalizability curves consistently fell within the 95% confidence intervals, achieving AUPRC values of 0.687 and 0.820, respectively. These values are in close alignment with the AUPRC of the models trained and tested on Opt-IN and ENGAGED (mean: 0.723 and 0.799, respectively).

e, f In a similar analysis for Brier score loss, the total losses recorded were 0.218 and 0.137 for the generalizability assessment, closely corresponding to the results from models trained and tested on Opt-IN and ENGAGED, with mean values of 0.216 and 0.166, respectively. Therefore, the SMART model generalized effectively to the external datasets without substantial loss in performance on Opt-IN and ENGAGED.

2.5 Discussion

In this study, we developed a machine-learned model that leverages recent advances in explainability and generalizability to increase validity, reliability, and trust in our ability to detect non-responders to weight loss programs with a 6-month endpoint and trigger adjustment in treatment. Importantly, our algorithm outperformed the clinical decision rule that has been used in past stepped-care trials. The algorithm was also generalizable, given that its effectiveness was replicated in two other distinct weight loss trials. For these reasons, our model holds promise as a valuable tool for personalized treatment allocation and improved outcomes in stepped-care weight management treatments.

The healthcare sector is witnessing a growing concern surrounding the adoption of ML algorithms that suffer from low interpretability.^[40] Clinicians and healthcare professionals hesitate to rely on decisions rendered by these models, given the lack of insight and explanations driving the outcome prediction and the high stakes of clinical decision-making. Although computer scientists and researchers often report the most highly per-

formant model, clinicians and experts will likely not use a tool they do not understand.[41] Fortunately, recent advances in interpreting complex models, make it more feasible to incorporate explainability into the decision-making process. Explainability methods are known to enhance trust in the model by allowing clinicians to understand that the model's accurate predictions are grounded in valid reasoning and that any errors made by the model are based on understandable and justifiable factors found in the data.[42]

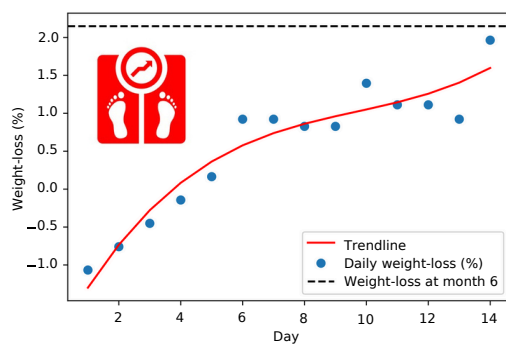
The SMART-trained model exhibited stable performance with no substantial difference compared to the models trained and tested on the Opt-IN and ENGAGED datasets separately, as indicated in Table 2.4.3 and Figure 2.4.2. This finding suggests features derived from objectively measured weights from wireless weight scales were discriminative and generalized effectively across external data from Opt-IN and ENGAGED studies, that used self-reported weights. Furthermore, the SMART-trained model outperformed the model trained and tested on ENGAGED in terms of AUPRC and Brier score loss. This might be due to the smaller sample size of 30 participants in ENGAGED, which can make it challenging to train a well-calibrated classifier capable of delivering stable predictions. Therefore, a generalizable model trained on a larger sample size with objective measurement is better equipped to withstand minor changes in study design and measurement procedures.

Simple rule-based approaches, such as the clinical decision rule (responders identified as losing ≥ 1 pounds in 2 weeks) result in low sensitivity (5.5%) and an F1 weighted score of 44.4%. The simple model generates a high number of false negatives, predicting non-responders when they were actually responders. Such a model would result in an increase of treatment components or care intensity for participants who do not need them, resulting in increased cost and care team burden. Given that initial weight loss has been

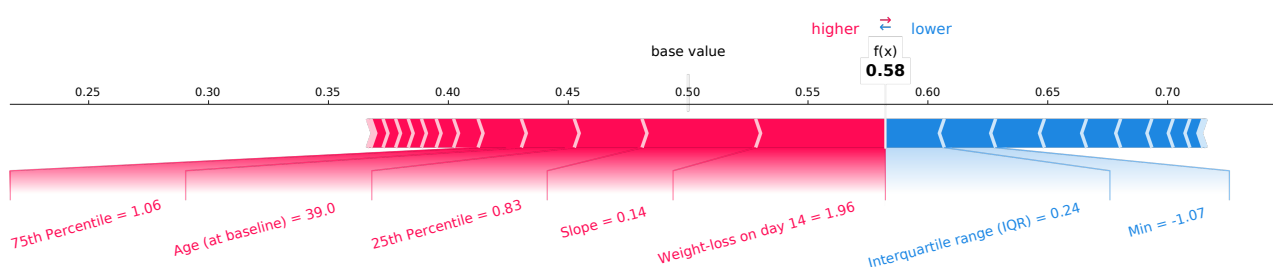
shown to predict long-term weight loss[43], we designed a logistic regression model that uses weight loss at the end of the second week as the sole predictor. Although this model improved the clinical decision rule (F1 weighted score, 66.5%) and provided greater interpretability, it comes at the expense of low accuracy and low sensitivity (40.9%). We then opted for building a model using the random forest classifier, which is known to prevent overfitting compared with typical decision tree models, generalizes well, and has shown success in various domains including finance and healthcare.[44], [45] Figure 2.5.1 shows a comparative analysis between the selected random forest model and logistic regression model.

We then set out to build a model using data only collected at baseline to test prediction before treatment initiation. The results were promising and uncovered some important features; however, the model had low accuracy, comparable sensitivity (40.9%), and a slightly lower F1 weighted score (59.0%). Our findings suggest that the statistical features extracted from the 2-week weight loss trajectory possess discriminative power, leading to a notable 16.7% increase in the F1 weighted score when compared with a random forest model trained solely on demographic and psychological features. Notably, certain demographic attributes such as age, gender, and height exhibit considerable discriminatory potential, implying that older and taller individuals, as well as males, are more likely to succeed in losing weight. These observations are in line with findings from prior work describing the reasons females and younger adults have greater difficulty losing weight.[46], [47] Increased body weight variability is a result of fluctuations in weight loss and gain and is influenced by a diverse range of factors, encompassing variations in dietary, physical activity, and other behavioral patterns (e.g., time of daily weighing).[48] The consequential impact of weight variability extends to numerous health outcomes, including conditions like depression,[49] non-alcoholic fatty liver,[50] and cardiovascular disease.[51]

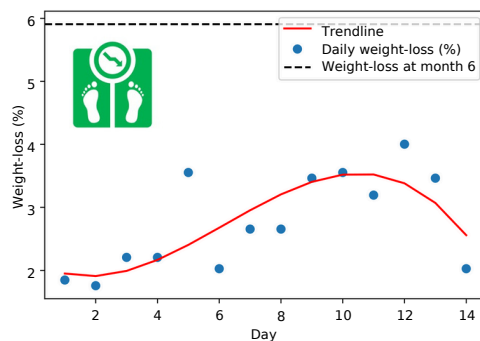
(a)



(b)



(c)



(d)

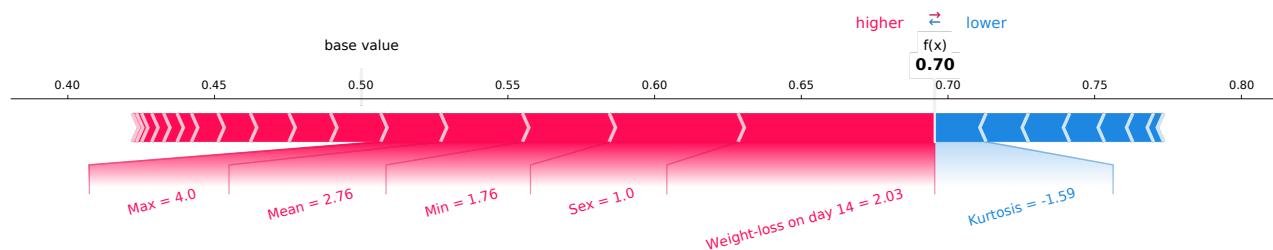


Figure 2.5.1: Comparative analysis of weight loss predictions: random forest vs. logistic regression models across initial two-week trajectory. (Description continued on next page.)

Figure 2.5.1: (Continued) The plots **a** and **c** delineate the trajectory of weight change over a 14-day period from the baseline for two participants, one achieving suboptimal weight loss and the other achieving sufficient weight loss at the 6-month mark. Accompanying the weight change data are SHAP force plots (**b** and **d**), illustrating the features contributing to the predictive models' outcomes for these individuals. Additionally, daily weight loss percentages are depicted using blue dots, while a red line signifies the trend in weight loss trajectories through orthogonal distance regression (ODR), for both participants. A comparative analysis is conducted between the predictive capabilities of the proposed random forest model (trained on the full set of features), and a logistic regression model (trained exclusively on the data pertaining to weight loss on day 14). **a** The logistic regression model functions as a tailored decision rule, categorizing participants who exhibit a weight loss of $\sim 2.44\%$ by the end of the second week as likely to achieve weight loss success at the 6-month mark. The random forest model considers a broader set of variables, including the trajectory of weight loss as determined by the slope, the participant's age (**b**), and the weight loss observed on the 14th day. However, the logistic regression model, relying solely on the singular metric of weight loss on day 14—approximately 2%—predicts suboptimal weight loss by the six-month marker. **c** In contrast, for this participant, the random forest model leverages a broader array of variables (**d**), including maximum, minimum, mean weights, and gender, enhancing the precision of its prediction. Conversely, the logistic regression model's reliance on a single variable—weight loss on day 14, estimated at 2%—renders it susceptible to misclassification of this participant's outcome.

Measuring weight variability is challenging due to outliers in weight data, inconsistent recording of weights by individuals, and non-linear trends in weight over time. These factors make accurate quantification of weight variability complex. Researchers vary in how they operationalize or measure weight variability, adopting distinct definitions tailored to their specific research objectives. However, the rationale driving selection of these definitions often remains unexplained, contributing to lack of consensus about the optimal characterization of weight variability to inform the investigation of weight loss prediction. Researchers have shown that short-term elevated weight variability is predictive of poor long-term outcome.^[52] Additionally, variation between weekdays and weekends, such that high compensatory weight loss during weekdays counteracts the upward trajectory of weight gain during weekends, predicts long-term weight loss and maintenance over time.^[53] This may be attributed to the operationalization of weekly weight variability, which focuses on comparing non-consecutive days of the week rather than

emphasizing variations in consecutive day-to-day changes. Although weekly weight variability is an interesting area of inquiry, our desire to predict weight loss as early in the intervention as possible renders it unfeasible to estimate weekday/weekend weight variability. The prevailing approach for long-term weight loss prediction has typically relied on the use of RMSE through linear regression methods. However, this method is not without limitations, primarily due to its inability to capture non-linearities in weight trajectory fluctuations. A different approach proposed employs LOESS regression, yet like other least squares methods, this technique is sensitive to outliers.[54] The estimation of body weight variability, regardless of the chosen definition, frequently encounters challenges stemming from the presence of missing data.[55] As the frequency of missing weight records increases, weight variability can decrease depending on the imputation method employed. Whether using self-report or technology-based smart devices to more objectively operationalize the measure of weight, there is a need to address missing data. We therefore scale our definition of a successive weight variability feature by the count of missing weight records spanning the 2-week interval. This underscores that higher successive weight variability and/or suboptimal utilization of technology-based tools (higher missing rate) are associated with reduced success in weight loss.

Our findings suggest another important aspect to consider: an elevated level of missing data likely indicates a decrease in adherence to the study protocol, perhaps reflecting faltering motivation that would be expected to result in failing to achieve weight loss in 6 months. However, isolated total missingness failed to demonstrate predictive capacity. This becomes more predictive when coupled with successive weight variability and suggests that the predictive factor for long-term weight gain may not solely stem from elevated weight variability but rather arise from the interplay between heightened weight variability and diminished adherence to prescribed study protocols.

Misclassification in behavioral weight loss trials can have significant implications, particularly concerning the risks associated with both under- and over-treatment. Incorrectly classifying an individual as a responder may result in the failure to provide necessary interventions, leading to suboptimal weight loss outcomes and potentially exacerbating obesity-related health conditions.[56] Conversely, the consequences of misclassifying an individual as a non-responder, leading to more intensive treatment than necessary, are not as well understood and have not been extensively studied. Although previous studies have found that the risks associated with intensive behavioral interventions, including weight loss maintenance programs for adults with obesity, are minimal, over-treatment may introduce risks at multiple levels, at both the individual and population level. At the population level, a prominent adverse consequence of overtreatment is the inefficient allocation of resources to superfluous care, which can result in the depletion of critical resources[57], subsequently limiting their availability for those who really need it. On an individual level, providing support that exceeds what is necessary can compromise autonomy. Individuals who perceive their weight loss as being reliant on external assistance may fail to cultivate the self-confidence and self-efficacy that could be gleaned by losing weight of their own accord, without external support. Achieving the right balance between providing support and encouraging autonomy is essential in weight loss interventions.[58]

Weight loss is a multifaceted process influenced by diverse physiological, psychological, and environmental factors. Our analysis is limited in that our feature set does not encompass elements capturing dietary patterns or variations in physical activity, which could significantly enhance the study's comprehensiveness. However, our choice of static/dynamic features aligns with those commonly found in most standard clinical settings. Furthermore, it is essential to acknowledge that ML models are inherently opaque. Al-

though SHAP aids in mitigating this opacity, it does not address potential algorithmic biases embedded in the model, often due to underlying assumptions in the algorithm's design. It is important to acknowledge that excessive reliance on SHAP interpretations may lead to overinterpretation, potentially resulting in the emergence of narrative fallacies.[59]

Our primary outcome variable is measured at 6 months, which presents a limitation in predicting the long-term sustainability of weight loss. Future research should aim to predict weight loss maintenance by extending the analysis to 12-month outcomes and beyond. Given the complexity and uncertainty of human behavior, predicting weight loss is an inherently challenging problem. Although initial weight loss is known as a strong predictor of weight loss maintenance, we shed light on the utility of low weight loss variability in predicting weight loss success through the predictive lens of machine-learned models. Our findings show that we can reliably predict weight loss as early as 2 weeks into a weight loss attempt, laying the foundation for future dynamic stepped-care weight loss models. However, the model is based on a limited universally measured set of features in common in clinical settings and a limited sample size from 3 studies; as such, it is meant as a proof-of-concept analysis. Before this kind of model can be used as a decision-support tool, the effectiveness of early-prediction models needs to be confirmed in a prospective study encompassing a diverse array of populations, a broad spectrum of treatment methodologies, and a range of different temporal intervals.

2.6 Summary

This chapter presented a machine learning framework for the early prediction of long-term weight loss success in behavioral obesity interventions. The study was motivated

by a central limitation in stepped-care weight management: existing decision rules can identify non-responders only with limited accuracy or after a delay that may reduce the effectiveness of treatment adaptation. To address this gap, this work examined whether early weight change patterns, collected during the first two weeks of treatment, could be used to predict clinically meaningful weight loss at six months.

The analysis used data from three behavioral weight management trials: SMART, Opt-IN, and ENGAGED. The SMART trial was used as the development cohort because it included objectively measured daily weights from wireless scales, while Opt-IN and ENGAGED were used as external validation cohorts. A random forest classifier was developed using both baseline features, such as demographic and psychosocial characteristics, and dynamic features extracted from early weight trajectories. These dynamic features included early weight loss, slope, distributional statistics, and measures of weight variability.

The proposed model outperformed existing clinical decision rules and baseline-only prediction approaches. The final random forest model achieved an AUROC of 0.845 and an AUPRC of 0.863 for predicting six-month weight loss success. Explainability analyses using SHAP showed that weight loss at day 14, trajectory-based features, weight variability measures, and age were among the strongest predictors of treatment outcome. The model also generalized effectively to the Opt-IN and ENGAGED cohorts, suggesting that early weight trajectory features can remain predictive across different intervention designs and measurement procedures.

This work contributes to the weight loss and digital health literature by demonstrating that meaningful prediction of long-term treatment success can be achieved as early as two weeks into an intervention. Rather than relying only on a simple threshold of early

weight loss, the study showed that richer statistical representations of early weight trajectories provide additional predictive value. The chapter also introduced a sparsity-adjusted measure of successive weight variability, linking both weight fluctuation and missingness in self-monitoring data to later outcomes. Overall, this work provides evidence that explainable and generalizable machine learning models can support more adaptive, efficient, and personalized stepped-care weight management interventions.

2.7 Supplementary

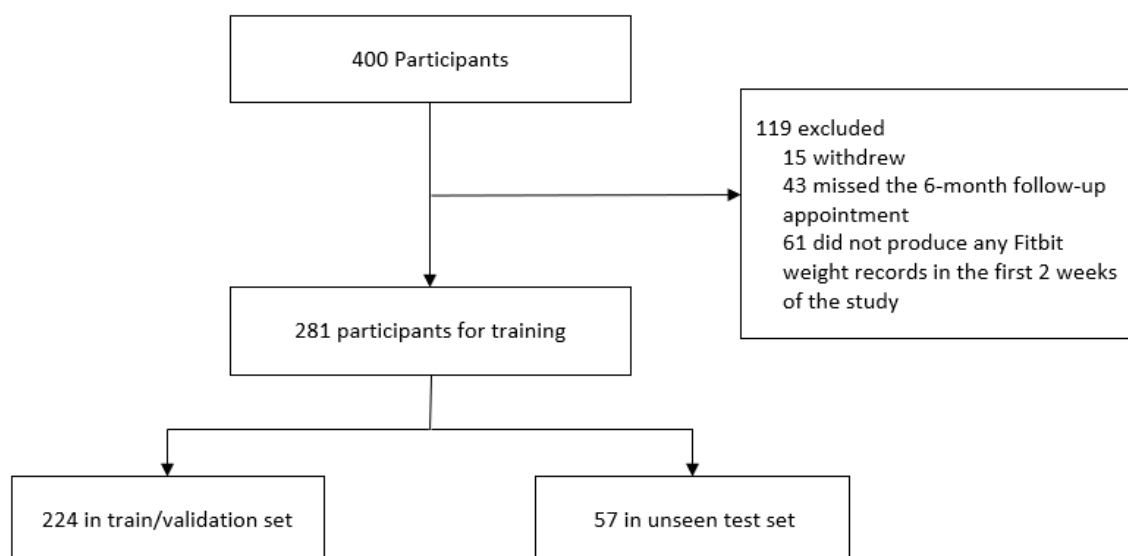


Figure 2.7.1: **Study profile for SMART study.** The figure provides a visual representation of the number of participants in the SMART Weight Management Study. It illustrates the total number of study participants, the count of individuals excluded from our analysis, and the overall number of participants considered in our analytical assessment.

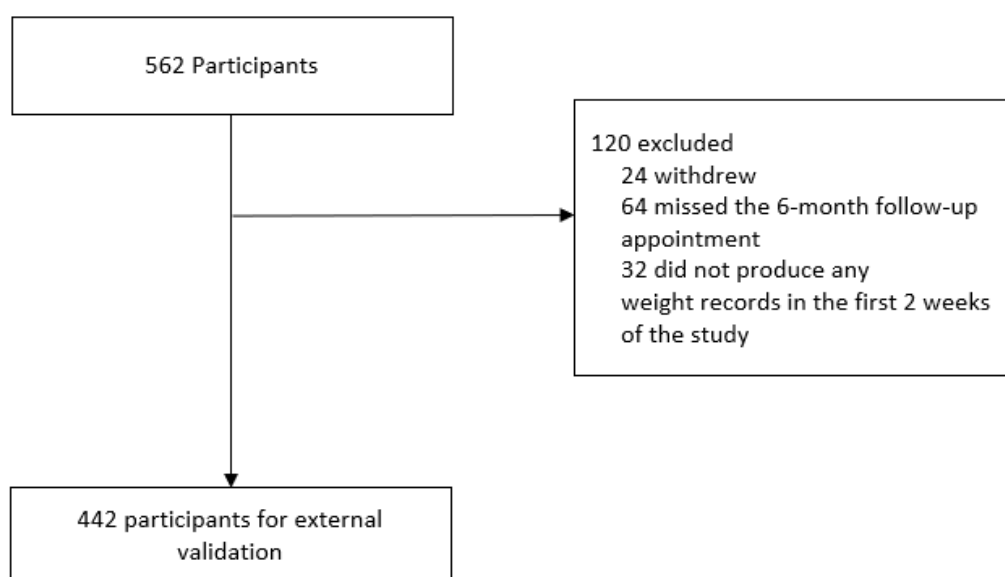


Figure 2.7.2: **Study profile for Opt-IN study.** The figure provides a visual representation of the number of participants in the Opt-IN Weight Management Study. It illustrates the total number of study participants, the count of individuals excluded from our analysis, and the overall number of participants considered in our analytical assessment.

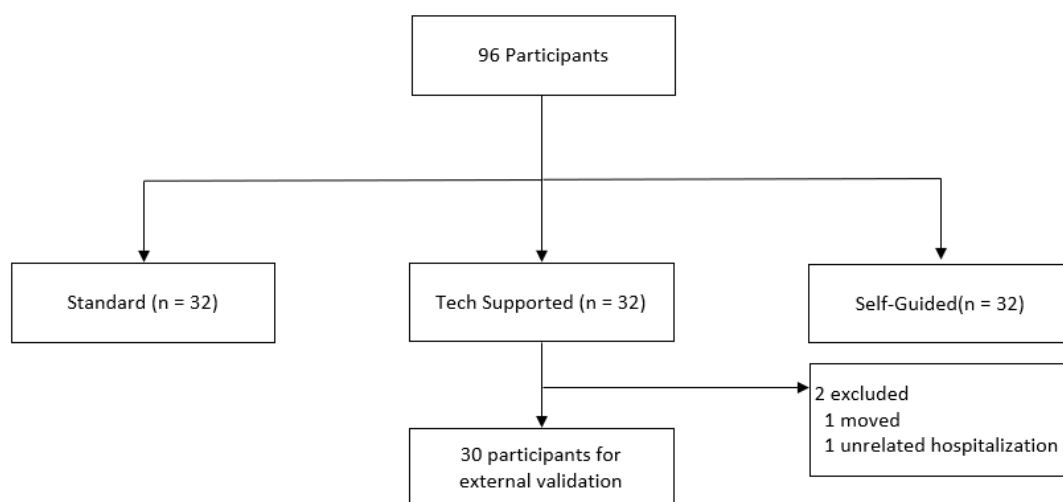


Figure 2.7.3: **Study profile for ENGAGED study.** The figure provides a visual representation of the number of participants in the ENGAGED Weight Management Study. It illustrates the total number of study participants, the count of individuals excluded from our analysis, and the overall number of participants considered in our analytical assessment.

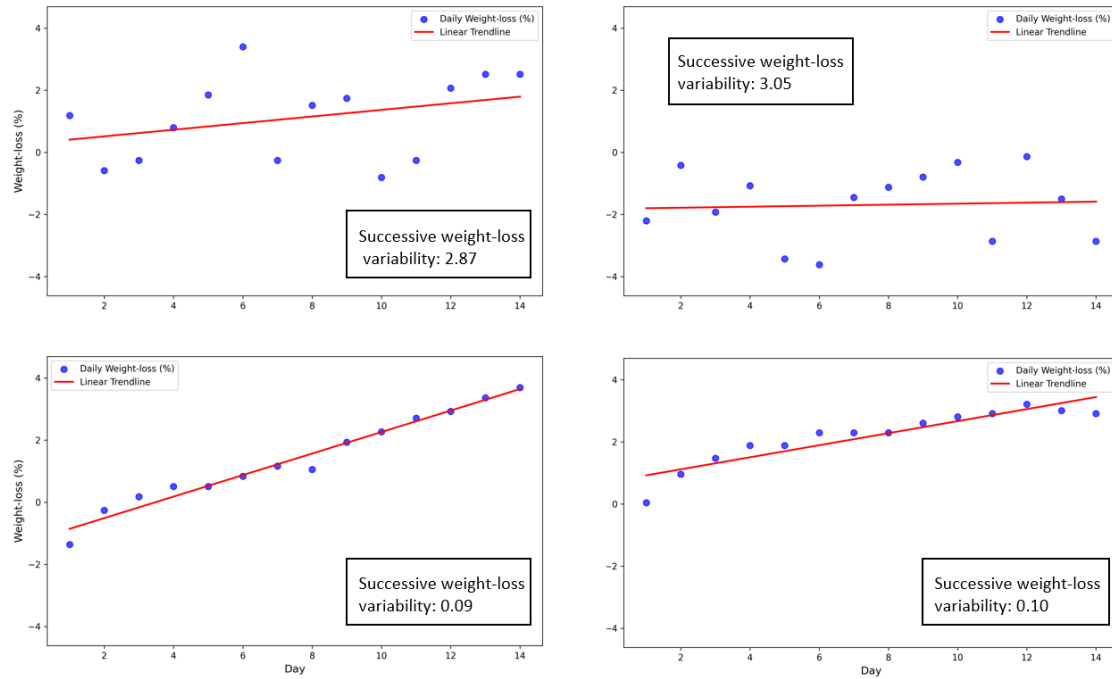


Figure 2.7.4: Examples of individuals with high and low weight loss variability. This figure presents the variation in weight change over 14 days from baseline weight for four participants. Successive weight-loss variability was calculated based on the daily weight loss percentages (in blue). To better visualize the overall trend and highlight the variation in weight loss over time, a red linear trendline is superimposed on each plot. This allows for a clearer comparison of weight-loss consistency among the participants, distinguishing those with high variability from those with low variability. (A) Two individuals with high weight variability are shown, calculated using the proposed successive weight loss variability method. (B) Two individuals with low weight loss variability are presented.

Weight Variability Measures	Formal Operational Definition	Derivation	Domain/Application	References
Coefficient of Variation	The standard deviation of the observed records over the average of the observed records	$CV = \frac{\sigma}{\mu}$ where: μ = observed records' mean, σ = observed records' standard deviation	All-cause mortality, Type 2 diabetes	Nam et al.[34], Okada et al.[60]
Mean Successive Variability	The mean absolute difference between successive weight records	$ASV = \frac{1}{n-1} \sum_{i=1}^{n-1} y_i - y_{i+1} $ where: n = the number of records, y_i = observed record	Coronary disease, Diabetes, Heart failure with preserved ejection fraction (HFpEF)	Banglore et al.[35], Rhee et al.[61], Li et al.[62]
Root Mean Squared Error	The square root of the average of the squared residuals from linear regression method	$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$ where: y_i = observed record, $\hat{y}_i$ = fitted record, n = number of records	Long-term weight-loss success, Cardiovascular, All-cause mortality	Benson et al.[38], Cologne et al.[37], Dyer et al.[63]
Non-linear Mean Deviation	The average of the absolute residuals error from LOESS regression method	$NLMD = \frac{1}{n} \sum_{i=1}^n y_i - \hat{y}_i $ where: y_i = observed record, $\hat{y}_i$ = fitted record	Weight-loss maintenance	Turicchi et al.[55]

Table 2.7.1: Descriptions of Body Weight Variability Operational Definitions in Difference Studies

ML Model	Sensitivity	Specificity	F1-score (6-month weight-loss success)	F1-score (suboptimal 6-month weight-loss)	F1 weighted	ROC-AUC	Balanced Accuracy
k-NN	0.364	0.714	0.400	0.676	0.569	0.519	0.539
Logistic Regression	0.318	0.914	0.438	0.780	0.648	0.684	0.616
Naïve Bayes	0.455	0.743	0.488	0.712	0.626	0.701	0.599
Random Forest	0.409	0.714	0.439	0.685	0.590	0.652	0.562
SVM	0.273	0.714	0.316	0.658	0.526	0.497	0.494
Gradient Boosting	0.227	0.886	0.323	0.747	0.583	0.713	0.556
Mean (STD)	0.341 (0.085)	0.781 (0.093)	0.401 (0.069)	0.710 (0.046)	0.590 (0.043)	0.628 (0.095)	0.561 (0.044)

We conducted training and testing of both discriminative and generative models, utilizing static features such as psychological and demographic attributes collected at the baseline. This table provides an overview of the baseline results. Notably, the Logistic Regression algorithm demonstrated F1 weighted scores of 64.8% and balanced accuracy of 61.6%. In general, the machine learning models trained at baseline exhibited an average F1 weighted score of 59.0% with a standard deviation of 4.3%.

Table 2.7.2: Performance of ML Models at Baseline

Variable	Feature Description
<i>Statistical Features</i>	
Mean	Average weight loss from a two-week weight trajectory
Variance	Variance of the weight loss from a two-week weight trajectory
Median	Median of the weight loss from a two-week weight trajectory
Maximum	Maximum weight loss from a two-week weight trajectory
Minimum	Minimum weight loss from a two-week weight trajectory
75th Percentile	75th percentile weight loss from a two-week weight trajectory
25th Percentile	25th percentile weight loss from a two-week weight trajectory
Range	Range of weight loss from a two-week weight trajectory
Inter-quartile range	Distance between 75th percentile and 25th percentile
Slope	Slope obtained from the best-fitted regression line for the two-week weight trajectory
Intercept	Intercept obtained from the best-fitted regression line for the two-week weight trajectory
Zero-crossing	Number of times in change of sign (e.g., positive to negative) in the mean-centered two-week weight loss trajectory
Kurtosis	A measure to calculate whether the two-week weight loss trajectory is heavy-tailed or light-tailed
Skewness	A measure to calculate the skewness of the weight loss trajectory
Weight loss on the 14th day (%)	The percent of weight loss on the 14th day
Weight loss on the 7th day (%)	The percent of weight loss on the 7th day
Sparsity-adjusted successive weight variability (SASWV)	As described in the main body of the paper

Table continued on the next page.

Table 2.7.3: Feature Description Table

Variable	Feature Description
<i>Baseline Features</i>	
Self-efficacy Score (diet)	As described in Pfammatter et al.[28]
Self-efficacy Score (exercise)	As described in Pfammatter et al.[28]
Sex	Sex
Race/ethnicity	Combined race and ethnicity variable
Age	Age
Height	Height
Baseline weight (lb)	Weight measured in the beginning of the study

In this table, we provide a detailed description of the features utilized in the machine learning models, encompassing both baseline features and statistical characteristics.

Table 2.7.3: Feature Description Table (continued)

2.7.1 Supplementary Note 1

Stratified Splitting and Model Evaluation

To maintain an unbiased modeling approach, we randomly split the data into train, validation, and test sets, stratified by sex, race/ethnicity, and the outcome variable (weight loss success) to ensure diversity within each set while avoiding cross-contamination. We repeated this 60/20/20 split 15 times using different random states, with each iteration selecting distinct samples for the train, validation, and test sets while preserving the same stratification. We then compared the results (with a fixed random state) to the 95% confidence interval range for the F1 weighted score. We observed that the results consistently fell within this confidence interval, confirming that the F1 weighted score of 75.7% was within the confidence interval range (95% CI, 74.8%–77.1%).

Chapter 3

SenseWhy

3.1 Technical Challenge: Objective Measurement of Proximal Eating Behaviors in Free-Living Settings

Research Question 2

How can multimodal sensing and self-report be integrated to detect, characterize, and phenotype proximal eating behaviors such as overeating in free-living settings?

Obesity remains a significant public health challenge despite various pharmacological and behavioral interventions. Since 1975, the global prevalence of obesity has tripled in number. As of 2022, 2.5 billion adults aged 18 years and older were considered overweight, including over 890 million who were classified with obesity.[22] Traditional behavioral weight loss interventions often fail to provide long-term results, with many individuals experiencing weight regain within 12 months post-intervention.[64], [65] One common target of obesity interventions and treatments is overeating.[66] However, these efforts are often unsuccessful, potentially because the specific patterns and behaviors that contribute to overeating are not well understood.[67] One critical reason for this

deficit in knowledge is the lack of a multifaceted approach to the study of eating behavior, excluding an investigation into the complex and interrelated factors that contribute to overeating. Consequently, most current treatments do not account for the dynamic interplay of psychological, contextual, and physiological factors, highlighting the need for more personalized and adaptive intervention strategies.[68]

Considerable effort has been devoted to identifying predictors of overeating, primarily employing self-reported data in combination with linear statistical approaches. Most of these studies have focused on single proximal determinants such as stress, cravings, and loss of control (LOC) as potential predictors of overeating.[13], [69], [70], [71] For example, recent studies have identified various single predictors of overeating, such as emotional eating and impulsive responses to food cues.[72], [73], [74] While Ecological Momentary Assessment (EMA), a research methodology that involves repeatedly sampling participants' behaviors, experiences, and moods in real-time and in their natural environments, provides valuable insights into eating behaviors, it can sometimes suffer from limitations related to the accuracy of meal timing and portion size reporting. Factors such as recall bias or delayed entries may affect the precision of self-reported data; thus, objective measures are needed to improve our understanding of overeating.

Wearable sensors are a promising source of objective data concerning overeating behaviors and predictors. Wearables can collect data passively and continuously, allowing researchers to obtain behavioral measurements that are both richer and more frequently collected than those collected with self-report measures. Applying wearable sensors to eating studies can increase the reliability of our data and introduce new ways of analyzing the occurrence and co-occurrence of overeating predictors through the provision of richer, finer-grained datasets.

In this study, our aim was to first apply machine-learning algorithms to identify features that predict overeating using passive sensing data, both with and without ecological momentary assessment (EMA) inputs. Secondly, we used these identified features to construct distinct clusters of overeating and eating episodes, allowing us to differentiate theoretically and clinically relevant patterns of problematic overeating behaviors. This approach enables the provision of individualized data for personalized, adaptive interventions, overcoming the limitations of current one-size-fits-all strategies.

3.2 SenseWhy Study Design and Personalized Overeating

Labels

To improve the accuracy of dietary intake assessments, we incorporated passive data collection methods using wearable cameras. Wearable cameras provide objective direct meal verification through recorded footage with validated timestamps, capturing both meal timing and dietary composition. Additionally, participants completed 24-hour dietitian-administered dietary recalls, which included a photo-assisted feature to minimize reporting biases and ensure professional oversight of caloric intake estimations.^[75] To analyze the collected data, we applied advanced machine learning techniques to identify patterns in eating behavior. Specifically, we used wearable device data and machine learning models to detect meal timing and examine the relationships between behavioral, psychological, and contextual predictors of overeating episodes. This methodological approach enables a comprehensive analysis of co-occurring factors contributing to overeating.

3.2.1 Study design

The SenseWhy study involved 65 adult participants with obesity ($\text{BMI} \geq 30 \text{ kg/m}^2$) residing in the Chicago Metropolitan Area, and was conducted over a 14-day (two-week) period.[76] Each participant received a sensing suite, including a chest-worn, activity-oriented wearable camera secured with a neck-worn lanyard and an under-the-shirt magnetic pad. The camera featured an upward-facing sensor array that captured video and thermal images of the face and surrounding areas, enabling visual tracking of eating behaviors in naturalistic settings. An infrared sensor enhanced data quality, particularly for nighttime or low-light eating conditions. Participants could pause recording to protect bystander privacy when others were present. Further details on device placement and orientation are illustrated in Supplementary Figure 3.7.9. Participants used a customized mobile application, installed on their personal smartphones, to record both dietary intake and event-based EMA surveys. The daily number of EMAs varied according to individual eating frequency, averaging approximately 4 entries per day. For each eating event logged, participants completed two pre-meal surveys and one post-meal survey. The pre-meal surveys—administered at the "Decided To" and "About To" stages—collectively consisted of 8 EMA items along with a meal image capture component for food content reporting. Following the meal, participants completed a post-meal survey that similarly comprised 8 EMA items, a meal image capture, and content reporting. This self-initiated, event-based design aimed to capture real-time, context-specific data while minimizing unnecessary notifications or disruptions. We used the validated multiple-pass method with the Nutritional Data System for Research (NDSR)[77] to collect 24-hour dietary recalls. This approach includes five steps: (1) listing all foods and beverages, (2) reviewing for forgotten items, (3) gathering additional details, (4) probing for forgotten foods, and (5) review of entire recall.[78] The before- and after-meal photographs and brief descrip-

tions for each eating event were provided to the dietitians in a web-portal to assist during recalls. Image-assisted dietary recall methods are shown to reduce underreporting and recall bias.[79] Images were processed by trained dietary interviewers as an initial step, with subsequent steps completed during the recall interview. Although these images were not used for calorie estimation, they helped confirm foods consumed, thereby enhancing recall accuracy. This study was approved by the Institutional Review Board (IRB) at Northwestern University under protocol number STU00204564, ensuring compliance with ethical standards.

3.2.2 Operationalization and validation of meal data collection

Self-reported EMA data on meals were collected using a mobile application. To validate caloric intake and meal composition, data from 24-hour dietary recalls and self-reported app data were merged by matching meals from both approaches within 60-minute windows. This process resulted in an EMA-only dataset containing validated caloric intake for each meal. From the camera footage, trained annotators labeled the precise start and end times of each meal following a standardized protocol. An eating episode was defined as a series of consecutive feeding gestures with intervals not exceeding 15 minutes, based on definitions from previous research.[80] Drinking gestures were also labeled separately. Additionally, fine-grained annotations of chews and bites were conducted. A bite was defined as a large jaw open/close movement when food touches the mouth, followed by chews defined as consecutive small jaw open/close movements. To accurately associate meals with the camera footage labels, we merged each meal using the start and end time labels from all data sources. We visually confirmed these associations by comparing the contextual self-reported data and the food content from the dietitian-administered 24-hour dietary recall data with the camera footage. These validated meals formed a dataset

comprising EMA response and passive sensing. We defined an overeating episode as any meal that met or exceeded 1 z-score above an individual's usual daily energy intake.[76], [81], [82] By analyzing the calorie distribution of all meals an individual consumed, we estimated a personalized 1 z-score cutoff for typical meal-level consumption. Overeating was then operationalized as a dichotomous variable, with 1 indicating an overeating meal and 0 representing a non-overeating meal. Supplementary Figure 3.7.10 displays 14-day eating profiles for several individuals, showing energy intake for all consumed meals.

3.3 Multimodal Feature Construction and Overeating Detection

3.3.1 Preprocessing and feature extraction

We extracted psychological, contextual, and behavioral features known to be associated with obesity for each meal from various data sources. Supplementary Table 3.7.3 provides detailed descriptions of EMA questions, camera-derived features, and the timing of data collection (before or after eating event). Psychological features were collected using EMA questionnaires administered through the study's mobile application. Participants responded to Likert-scale questions related to stress, emotions, affect, and hunger before and after each meal. Contextual features were derived from EMA questions focusing on social and environmental triggers during meals. To streamline the raw responses and address correlations among different answers, we preprocessed the data to consolidate responses into specific categories. Meals cooked from scratch, prepared using ingredient delivery services, or made from frozen prepackaged foods were classified as self-prepared meals. Meals obtained from restaurants were labeled as restaurant-sourced meals, while meals consisting solely of snacks, cereals, or beverages were designated as light refresh-

ments. Participants' activities during meals were categorized based on whether they were engaged in other tasks. If participants were socializing, watching TV, working, studying, or driving while eating, these instances were classified as task-oriented distractions. If they were not engaged in any of the activities listed above, this was termed focused eating. Meals consumed alone were labeled as solo dining, whereas those eaten with others were termed social dining experiences. Evening eating was defined as any meal consumed between 5 PM and 6 AM, based on meal distribution patterns observed in our dataset. Behavioral features were extracted from fine-grained annotations obtained through detailed analysis of camera footage for each meal, using a standardized annotation protocol. These features included the total number of chews (chewing motions during a meal), the total number of bites (instances where food was brought to the mouth), the chew rate (number of chews divided by the meal duration in minutes), the bite rate (number of bites divided by the meal duration in minutes), the chew-bite ratio (ratio of total chews to total bites), and the meal duration itself, validated through camera recordings. Annotation was performed by trained raters from a third-party professional labeling service, following a structured training program developed by the research team. Inter- and intra-rater agreement was monitored through an internal auditing system, periodic quality checks, and final validation by the research team. Further details on the annotation protocol, rater training, and quality control measures are provided in Supplementary Note 3.7.4. Building upon the extracted features, we conducted our analysis at three levels to evaluate the predictive capabilities of different data subsets. The first level, termed the passive-sensing-only analysis, included only the behavioral features obtained from the camera footage. The second level, the EMA-only analysis, utilized the psychological and contextual features derived from the EMA questionnaires. The third level, the feature-complete analysis, incorporated all available features. Detailed description of the analytical pipeline is displayed in Figure 3.3.1.

3.3.2 Overeating detection: Supervised machine learning

Prior to training machine learning models, we first performed feature preprocessing, including standardization of continuous variables (subtracting the mean and dividing by the standard deviation) to ensure zero mean and unit variance. For each subset of features (EMA-only, passive-sensing-only, and feature-complete), we implemented a 5-fold cross-validation procedure, stratified by the classes (overeating vs. non-overeating) to maintain class proportions within each fold at the meal level. Within each fold, the data was randomly split into training (60%), validation (20%), and test (20%) sets, ensuring no meal data cross-contamination between the sets. To assess the impact of meal-based splitting, we conducted an additional participant-level split analysis, where participants were assigned exclusively to one of the training, validation, or test sets. This evaluation was conducted to assess whether the meal-based and participant-based splitting strategies yielded comparable results. Full details of this analysis are provided in Supplementary Note 3.7.5. To address the imbalance between overeating and non-overeating labeled meals, we applied the Synthetic Minority Oversampling Technique (SMOTE)[83] to the training set. This ensured that the model learned from an equal representation of overeating and non-overeating meals, enhancing its ability to generalize across both categories. We evaluated several machine learning models, including XGBoost, support vector machines (SVM), and naive bayes in order to predict overeating episodes. Model training was performed on the upsampled training set using SMOTE. We performed hyperparameter optimization using Bayesian optimization,[39] utilizing the validation set to evaluate model performance during the optimization process. Final performance metrics were derived by averaging the results across the five folds. We also reported the standard deviation of these results computed from the between-fold results to assess the variability and robustness of the model's performance. Ultimately, we selected the best-performing

model based on the final performance metrics.

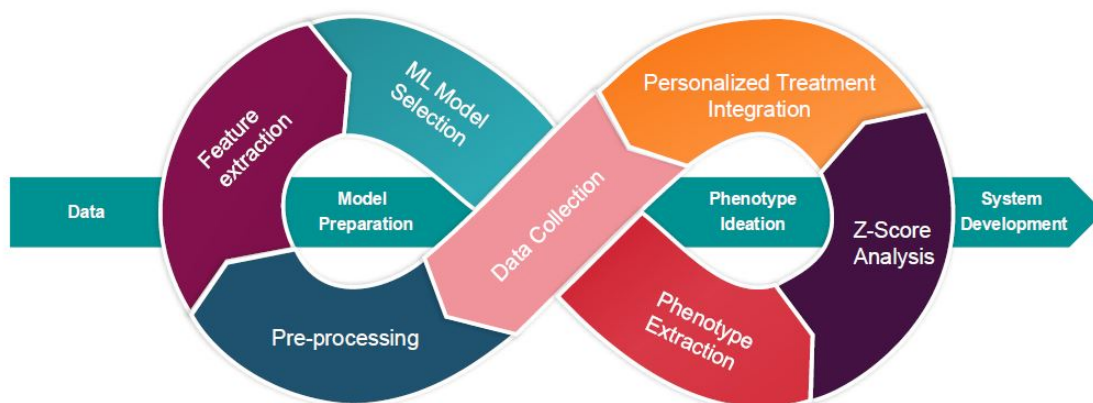
3.3.3 Supervised evaluation and explainability

Model performance was evaluated on the test set within each fold of the cross-validation procedure, using multiple evaluation metrics, including the Area Under the Receiver Operating Characteristic Curve (AUROC), the Area Under the Precision-Recall Curve (AUPRC), and Brier loss score. We report the mean and standard deviation of these metrics across the five cross-validation folds. Furthermore, we employed SHapley Additive exPlanations (SHAP)[25] to interpret the models and identify the importance and directionality of individual features contributing to the prediction of overeating outcomes. SHAP analysis was conducted on the best-performing fold for each dataset (EMA-only, passive-sensing-only, and feature-complete), allowing us to understand how each feature influenced the model's predictions and to uncover the underlying factors associated with overeating episodes. The results from SHAP also provided insights into the relative importance of behavioral, psychological, and contextual factors in predicting overeating behavior.

3.3.4 Overeating phenotype extraction: Semi-supervised machine learning

We focus on EMA-based data in phenotype generation primarily to ensure comparability with past and future studies, facilitating broader implementation in future interventions. This approach also provided the largest number of recorded meals, thereby increasing statistical power to capture a wider range of overeating episodes and enabling the identification of more distinct behavioral patterns. To uncover the intrinsic structure of our data and enhance interpretability, we implemented a semi-supervised learning pipeline that combined dimensionality reduction and clustering techniques. The dimensionality

(a)



(b)

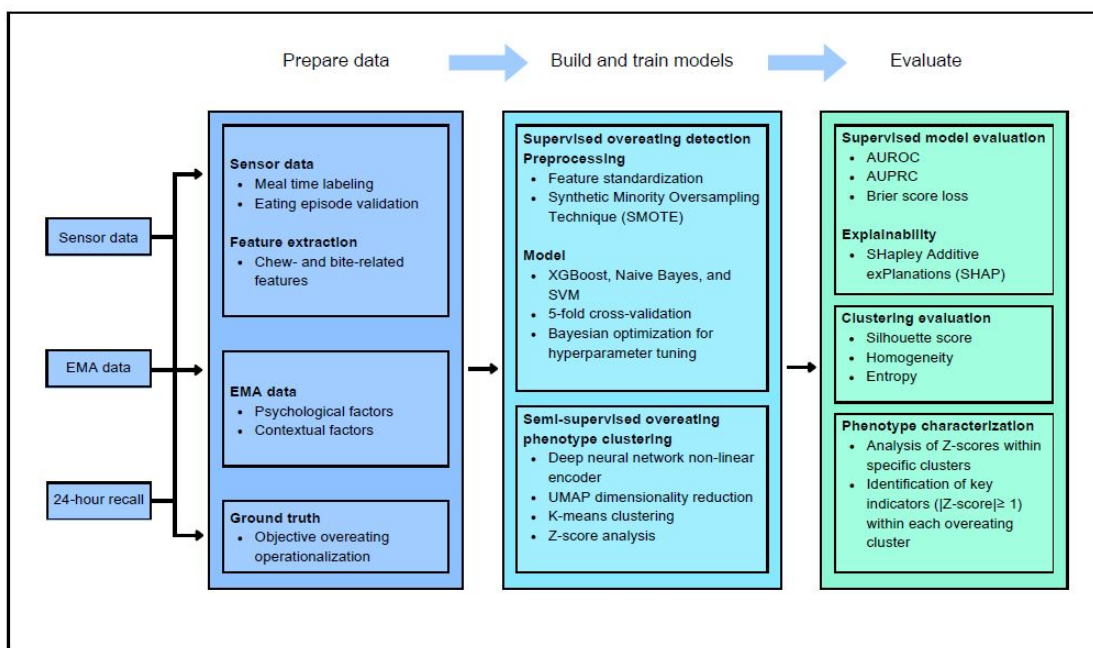


Figure 3.3.1: Analytical pipeline, a step-by-step process of the study's analytical approach. a Conceptual framework for an integrated ML pipeline aimed at identifying and addressing overeating phenotypes in practice. (Continued on next page.)

Figure 3.3.1: (Continued) The pipeline begins with data collection, integrating multimodal sources such as sensor data, EMA data, and dietary recalls to capture a comprehensive view of eating behaviors. This is followed by preprocessing to clean and harmonize the data, ensuring consistency across modalities, and feature extraction to derive key indicators such as chew rate, bite frequency, and emotional states associated with eating episodes. The next stage involves the development of ML models for overeating detection and phenotype ideation. Supervised learning models identify key features predictive of overeating episodes, leveraging behavioral and psychological features, while clustering techniques group individuals into distinct overeating phenotypes based on shared behavioral and contextual patterns. Once phenotypes are characterized, they can be integrated into personalized treatment strategies, tailoring interventions to address specific overeating patterns (context- and behavior-driven overeating). These treatments may include real-time feedback systems to prompt users to reflect on their behaviors, along with recommendations for sustainable behavioral changes. The framework culminates in system deployment, where real-time feedback and monitoring enable continuous assessment of eating behaviors and treatment efficacy. Data from deployed systems can feed back into the pipeline, enabling refinement and validation of models and interventions over time. This iterative process supports the practical application of overeating phenotype identification and management in real-world settings, creating a closed-loop system for adaptive and effective health interventions. **b** Methodological approach used in this study. The process begins with data preparation, including the labeling and validation of meal times from sensor data, integration of psychological and contextual factors from EMA data and 24-hour dietary recalls. Overeating detection is performed using supervised models, incorporating SMOTE for imbalanced data and Bayesian optimization for fine-tuning. A semi-supervised clustering approach identifies overeating phenotypes, leveraging a non-linear encoder, UMAP for dimensionality reduction, and K-means clustering with z-score analysis for phenotype characterization. Evaluation metrics include AUROC, AUPRC, and Brier score loss for model performance, SHAP for interpretability, and clustering metrics such as silhouette score, homogeneity, and entropy.

reduction was performed through a two-step process. Firstly, we employed a deep neural network architecture consisting of a feedforward multilayer perceptron (MLP) serving as a non-linear encoder. This encoder transformed the high-dimensional input data $X = \mathbb{R}^{n \times d}$ into a lower-dimensional latent representation $Z = \mathbb{R}^{n \times h}$, where n is the number of samples, d is the number of original features, and h is the reduced feature spaced dimension. The encoder compresses the input by learning a compact latent representation, ensuring that $h < d$ while retaining the most informative aspects of the data. The encoder captured non-linear relationships within the data, passively positioning similar data points closer together in the latent space while pushing dissimilar samples apart.

The transformation can be found in Equation (1):

$$Z = f_{encoder}(X; \Theta) \quad (3.1)$$

where $f_{encoder}$ denotes the non-linear function parameterized by weights Θ learned during training.

Subsequently, the latent representations Z obtained from the encoder were further processed using the Uniform Manifold Approximation and Projection (UMAP),[\[19\]](#) a non-linear manifold learning technique. UMAP reduced the data to a two-dimensional space $Y = \mathbb{R}^{n \times 2}$, preserving both global data structures while eliminating redundant and irrelevant features. This step facilitated visualization and interpretability of the high-dimensional data by mapping it onto a two-dimensional space in Equation (2):

$$Y = \text{UMAP}(Z) \quad (3.2)$$

Following dimensionality reduction, we applied the k-means clustering algorithm to the transformed data Y . To validate the robustness of our clustering results, we performed a sensitivity analysis using an alternative clustering method, Gaussian Mixture Modeling (GMM) that allows for soft assignments where each point has a probability of belonging to multiple clusters. We selected the number of clusters that maximized the silhouette score,[\[84\]](#) indicating the most appropriate number of clusters. Visual confirmation was performed using the two-dimensional UMAP projection to ensure well-separated clusters corresponding to distinct overeating meals.

3.3.5 Z-Score analysis for cluster interpretation

To interpret clusters, we computed cluster-level z-scores for each variable. This involved calculating cluster means of a given variable, and then computing a z-score across cluster means for that variable. The resulting value characterizes the deviation of a given cluster from the average value of a given variable on the scale of standard deviation units. Specifically, for each feature j and within each cluster i , we computed the z-score in Equation (3):

$$z_{ij} = \frac{\mu_{ij} - \mu_j}{\sigma_j}, \quad \mu_j = \frac{1}{k} \sum_{i=1}^k \mu_{ij}, \quad \sigma_j = \sqrt{\frac{1}{k} \sum_{i=1}^k (\mu_{ij} - \mu_j)^2} \quad (3.3)$$

Here, k is the total number of clusters, μ_{ij} represents the mean of feature j within cluster i , μ_j is the overall mean of cluster-specific means for feature j , and σ_j is the standard deviation of the cluster means for feature j .

This z-score represents how many standard deviations the cluster mean μ_{ij} deviates from the mean of cluster means μ_j for that feature. Features with high absolute z-scores (e.g., $|z_{ij}| \geq 1$) were considered substantially different from the average cluster mean, indicating their substantial influence on overeating behaviors within that cluster.

3.3.6 Clustering evaluation metrics

Overeating was not included as a feature in the clustering analysis; instead, we evaluated clusters based on the proportion of overeating episodes within each cluster, following clear criteria outlined in the following. We evaluated the resulting overeating clusters using per-cluster purity,^[85] which measures the extent to which each cluster contains data

points from a single class. The silhouette score assesses how well-separated and cohesive the clusters are by comparing intra-cluster and inter-cluster distances. Entropy[86] quantifies the uncertainty within each cluster, with lower values indicating higher purity. For overall clustering assessment, homogeneity[87] measures how uniform the clusters are with respect to the true class labels across the entire dataset. Lastly, we calculated the proportion of correctly assigned overeating meals to assess the clustering approach's effectiveness in identifying relevant clusters.

3.4 Evaluation and Results

Out of the initial 65 participants, 48 adults were included in the subsequent analyses as seven participants dropped out of the study, five lacked dietitian-administered recalls, and five recorded fewer than 10 meals during the study period as displayed in Figure 3.4.1. The final EMA-only dataset comprised 2,302 meal-level observations from participants with obesity (an average of 48 meals per participant). Participants had a mean age of 41 (range 21–66), and 77.1% were female. Baseline demographics and the EMA response summary are presented in Table 3.4.1. Supplementary Table 3.7.1 provides a summary of the EMA and passive sensing data, including demographic information, EMA response details, and statistics related to passive sensing features for 700 meals (an average of 17 meals per participant). To further illustrate the distribution of observations across participants, Supplementary Figure 3.7.1 displays the distribution of EMA and feature-complete data (incorporating all available features) per participant. Supplementary Figure 3.7.2 shows the hourly distribution of meals for EMA and feature-complete data across all participants. Details on wear time, adherence, and participant feedback for the SenseWhy wearable camera are available in Supplementary Note 3.7.1, Supplementary Table 3.7.2, and Supplementary Figure 3.7.3.

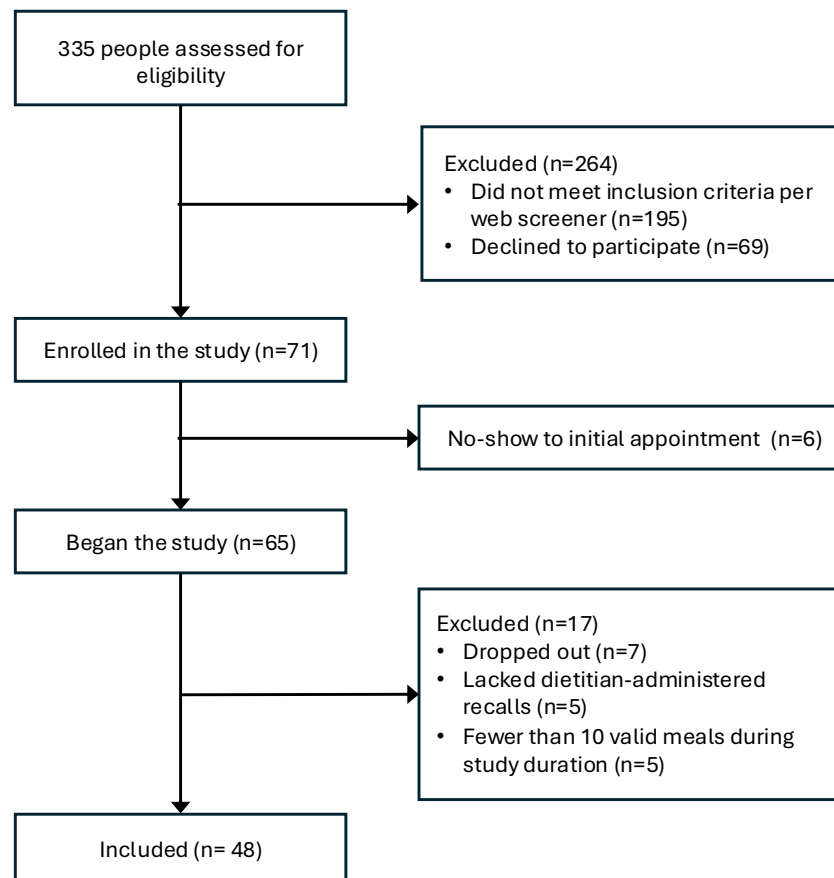


Figure 3.4.1: **Consort diagram.** This figure presents the participant flow through the study, from initial recruitment to final analysis, showing the number of participants at each stage (screened, enrolled, and included in the analysis) and the reasons for exclusions at various points.

Participant characteristics, mean (range)	
Age	41.0 (21–66)
BMI (kg/m ²)	37.2 (29.8–58.7)
Sex, <i>n</i> (%)	
Female	37 (77.1)
Male	11 (22.9)
Race, <i>n</i> (%)	
Black	20 (41.7)
White	19 (39.6)
Unknown	5 (10.4)
Asian	4 (8.3)
Ethnicity, <i>n</i> (%)	
Non-Hispanic	41 (85.4)
Hispanic	7 (14.6)
Objective overeating, <i>n</i> (%)	
Non-overeating	1934 (84.0)
Overeating	369 (16.0)
EMA-based features, <i>n</i> (%)	
<i>Location</i>	
Home	1896 (82.0)
Other	161 (7.0)
Work	157 (7.0)
Restaurant	67 (3.0)
School	21 (1.0)
<i>Food source</i>	
Snacks/cereal	731 (32.0)
Self-cooked from scratch	603 (26.0)
Bought from restaurant—Takeout	287 (12.0)
Self-cooked from frozen/prepackaged meal	272 (12.0)
No food, just beverage	239 (10.4)
Delivery	77 (3.0)
Bought from restaurant—Dine in	73 (3.0)
Self-cooked from ingredient delivery service	20 (1.0)
<i>Activity co-occurrence</i>	
Watching TV	838 (36.0)
Nothing else	478 (21.0)
Socializing	467 (20.0)
Working/studying	443 (19.0)
Driving	76 (3.0)

Table 3.4.1: Baseline Demographics and EMA Response Summary

EMA-based features, mean (range) (Continued)	
Hedonic Eating	3.03 (0.0–4.0)
Desire for Food	2.82 (0.0–4.0)
Pleasure-driven Desire for Food	2.78 (0.0–4.0)
Pre-meal Biological Hunger	2.59 (0.0–4.0)
Pre-meal Emotion (calm)	2.27 (0.0–4.0)
Pre-meal Emotion (upbeat)	1.70 (0.0–4.0)
Loss of Control (LOC)	0.94 (0.0–3.0)
Pre-meal Stress	0.92 (0.0–4.0)
Post-meal Stress	0.90 (0.0–4.0)
Pre-meal Emotion (lonely)	0.72 (0.0–4.0)
Perceived Overeating	0.59 (0.0–4.0)
Cognitive Restraint	0.30 (0.0–3.0)

Table 3.4.1: Baseline Demographics and EMA Response Summary (Cont.)

This table presents a comprehensive overview of the study participants' characteristics and their responses to Ecological Momentary Assessment (EMA) questions. It includes baseline characteristics (age and BMI), demographic information (sex, race, and ethnicity), objective overeating statistics, and summaries of EMA-based features.

3.4.1 Supervised overeating detection

We selected XGBoost as the best-performing model after comparing it with SVM and naïve bayes. While SVM is particularly effective in high-dimensional spaces and naïve bayes is efficient and robust, XGBoost, as a non-linear ensemble method, proved more effective in capturing complex patterns in the data, as indicated by Figure 3.4.2. We also evaluated the model by checking the training and validation errors across iterations to ensure robustness (see Supplementary Figure 3.7.4). For the detection of overeating, we conducted three separate analyses:

3.4.1.1 EMA-only analysis

XGBoost yielded a AUROC with mean (SD) of 0.83(0.02), AUPRC of 0.81(0.02), and a Brier score loss of 0.13(0.01). The top five features identified by SHAP were: light refreshment (negative), pre-meal biological hunger (positive), perceived overeating (positive), evening

eating (positive), and pleasure-driven desire for food (mixed association). In this context, positive association means the feature increases the likelihood of overeating, negative association means it reduces the likelihood, and mixed association indicates varying effects depending on the instance.

3.4.1.2 Passive sensing-only analysis

XGBoost resulted in an AUROC of 0.69(0.04), AUPRC of 0.69(0.05), and a Brier score loss of 0.18(0.02). The top five features were: number of chews (positive), chew interval (negative), chew-bite ratio (negative), number of bites (positive), and chew rate (mixed).

3.4.1.3 Feature-complete dataset

In this combined dataset, XGBoost achieved an AUROC of 0.86(0.04), AUPRC of 0.84(0.04), and a Brier score loss of 0.11(0.02). To improve prediction probability calibration, we applied post-calibration using the sigmoid method (Platt's scaling), which resulted in better alignment between predicted probabilities and observed outcomes, as shown in Supplementary Figure 3.7.5. The top five predictive features were: perceived overeating (positive), number of chews (positive), light refreshment (negative), loss of control (positive), and chew interval (negative). Figure 3.4.2 displays SHAP dot and bar plots, highlighting feature importance and their impact across all study levels.

3.4.2 Semi-supervised overeating phenotype clustering

After removing 56 zero-calorie meals (e.g., non-caloric foods and beverages), we arrived at a final dataset of 2,246 meals, of which 369 (16.4%) were identified as overeating episodes. This adjustment ensured that the analysis focused on meals that accurately represented eating behavior, allowing for more reliable detection of overeating patterns. We em-

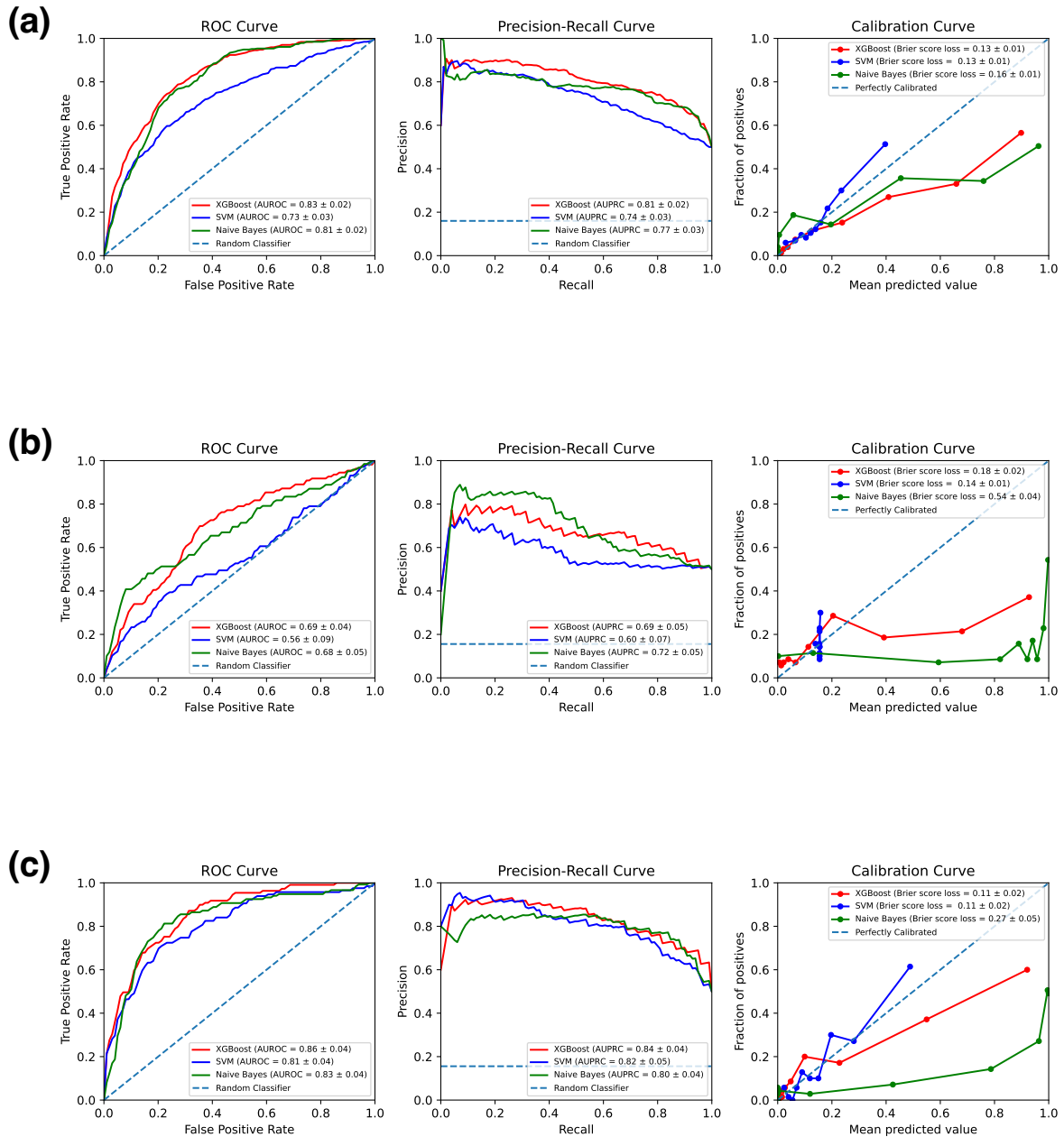


Figure 3.4.2: AUROC, AUPRC, and Brier Score for XGBoost compared to SVM and Naïve Bayes. Performance of XGBoost, SVM, and Naïve Bayes models across three data scenarios: **a** EMA-only data, **b** Passive sensing-only data, and **c** Feature-complete data. Each graph shows the AUROC, AUPRC, and Brier score loss for each model to allow comparison of performance across different data inputs and modeling approaches.

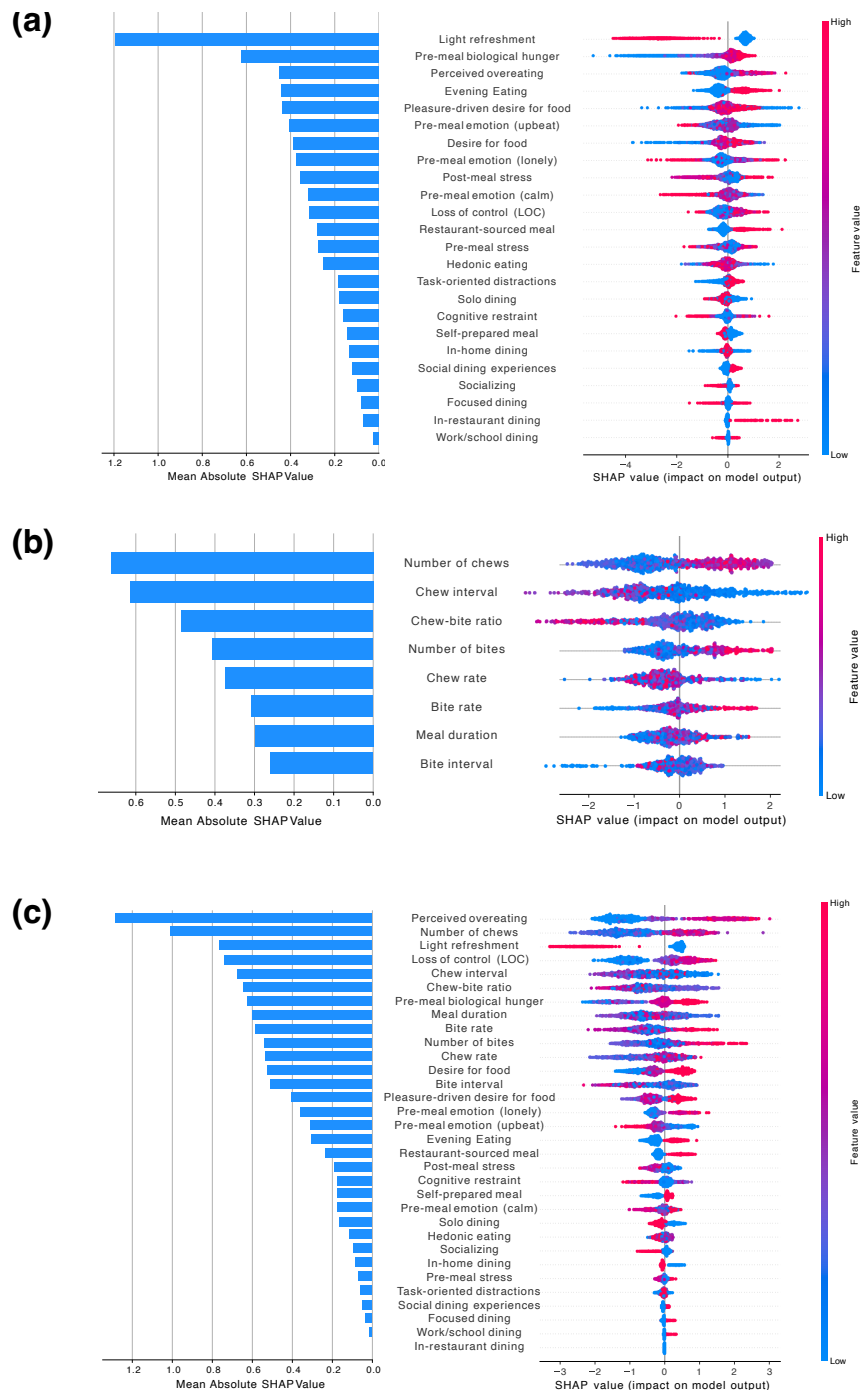


Figure 3.4.3: SHAP plots illustrating the feature importance and impact on XGBoost model output for different datasets. **a** EMA-only, **b** Passive sensing-only, and **c** Feature-complete data combining EMA and passive sensing.

ployed the semi-supervised clustering pipeline and evaluated cluster separability using the silhouette score over a range of 2 to 35 clusters. The separability was visually confirmed through 2D projection following UMAP. The pipeline, applied to the entire dataset of both normal and overeating meals, identified 30 distinct clusters, with the maximum silhouette score of 0.53 and a homogeneity score of 0.66 (Supplementary Figure 3.7.6). To define a cluster as an overeating cluster, we set a threshold of 0.05 for the proportion of the total overeating instances. Visual inspection further confirmed separability within five predominant clusters characterized by a high proportion of overeating instances (Supplementary Figure 3.7.7). The final clustering solution achieved a mean purity of 81.4%, a cumulative proportion of overeating instances of 0.85, and an entropy score of 0.36 across the five overeating clusters (further validated using GMM, as shown in Supplementary Note 3.7.2). Additionally, the silhouette score of 0.59 supported the coherence and distinctiveness of the identified clusters.

3.4.3 Z-score analysis of overeating clusters

We applied Z-score analysis to highlight differences within each cluster while identifying shared patterns and overarching themes across clusters. For contextual and psychological factors within each cluster, we selected features with z-scores exceeding the predefined cut-off ($|z| \geq 1$), allowing us to identify the dominant co-occurring factors within the same clusters and uncover shared patterns across clusters. Figure 3.4.4 presents a polar bar plot illustrating the z-scores for each contextual and psychological factor across all overeating clusters. We characterized each phenotype based on which feature-level z-scores were large in magnitude for each cluster (to highlight the distinctive characteristics of the cluster). Cluster labels (or names) were assigned to concisely reflect the key factors of each cluster, to foster and provide an intuitive understanding of the overeating

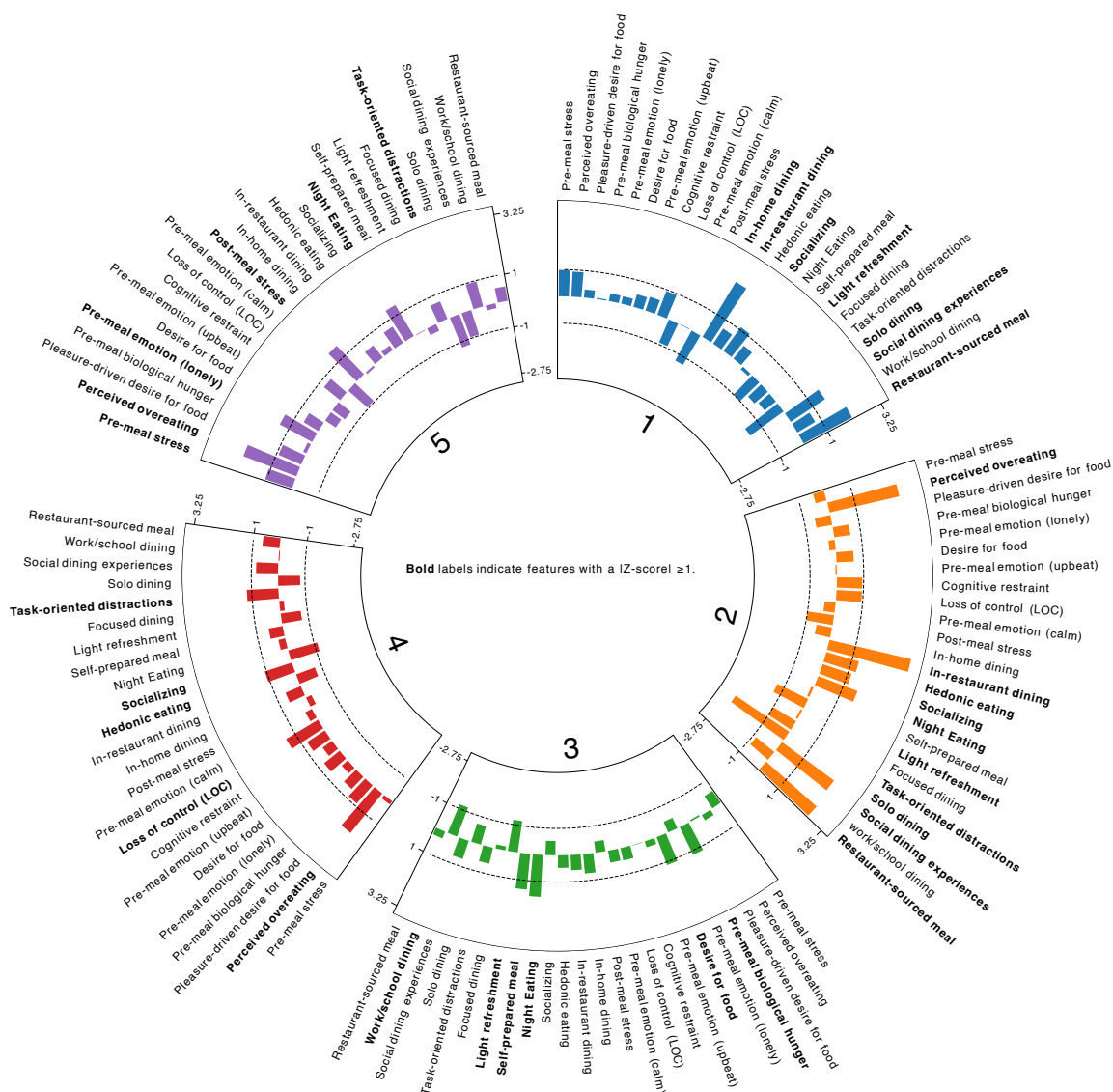


Figure 3.4.4: Polar bar plot illustrating the z-scores of features associated with distinct overeating phenotypes. The radial axis represents the magnitude of the z-scores, with bold labels indicating features with a $|z| \geq 1$. The circumference is divided into five phenotype clusters, each represented by a unique color, highlighting the key features that differentiate these clusters. This visualization provides insights into the characteristic features of each overeating phenotype, such as psychological factors, contextual influences, and behavioral patterns, enabling a deeper understanding of their unique profiles.

patterns observed in the data. A detailed analysis of all contextual factors exceeding the z-score threshold is provided in Table 3.4.2. Based on these results, we characterized the following overeating phenotypes:

1: Take-out Feasting

A preference for indulging in restaurant-sourced meals (take-out), often enjoyed in a social setting, emphasizing the social aspect of shared dining experiences.

2: Evening Restaurant Reveling

Pleasure-driven indulgence in food, with a preference for restaurant-sourced meals (dine-in), typically consumed in the evening as part of social dining experiences.

3: Evening Craving

Eating in the evening, often involving self-prepared meals and characterized by hunger, serving as a way to unwind at the end of day.

4: Uncontrolled Pleasure Eating

A focus on the hedonic aspect of food to indulge in eating for pleasure, often perceived as overeating and experiencing loss of control, engaged in task-oriented distractions.

5: Stress-driven Evening Nibbling

Eating in the evening in response to stress and feelings of loneliness.

Table 3.4.2: Phenotype interpretation

Overeating Cluster	Psychological Factors	Contextual Factors	Phenotype Name	Energy intake mean (SD) (kCal)	Cluster size (purity score (%))	Proportion of overeating instances	Entropy	Silhouette
Cluster 1		Restaurant-sourced meal (+) Social dining experience (+) Socializing (+) In-restaurant dining (+) Light refreshment (-) In-home dining (-) Solo dining (-)	"Take-out Feasting"	1006.53 (378.11)	71 (100%)	0.19	0.00	0.68
Cluster 2	Perceived overeating (+) Hedonic eating (+)	Social dining experience (+) Restaurant-sourced meal (+) In-restaurant dining (+) Light refreshment (-) Socializing (+) Night eating (+) Solo dining (-)	"Evening Restaurant Reveling"	1154.89 (445.87)	64 (95.31%)	0.17	0.19	0.77
Cluster 3	Pre-meal biological hunger (+) Desire for food (+)	Night eating (+) Self-prepared meal (+) Work/school dining (-) Light refreshment (-)	"Evening Craving"	1024.96 (434.18)	132 (85.61%)	0.31	0.41	0.42
Cluster 4	Perceived overeating (+) Loss of control (LOC) (+) Hedonic eating (+)	Task-oriented distractions (+) Socializing (-)	"Uncontrolled Pleasure Eating"	1100.70 (434.11)	50 (78%)	0.11	0.53	0.67
Cluster 5	Perceived overeating (+) Pre-meal stress (+) Post-meal stress (+) Pre-meal emotion (lonely) (+)	Night eating (+)	"Stress-driven Evening Nibbling"	1360.44 (433.44)	56 (48.21%)	0.07	0.69	0.43

This table presents a detailed breakdown of the identified overeating phenotypes. It includes information on psychological and contextual factors associated with each cluster, proposed phenotype names, mean energy intake, cluster size and purity score, proportion of overeating instances, entropy, and silhouette score for each cluster.

3.5 Discussion and Limitations

In the face of a global obesity crisis imposing staggering healthcare costs on societies, the need for innovative and effective interventions has never been more critical. Obesity treatment presents the unique challenge of modifying a core behavior—eating—that is essential for survival yet is influenced by several factors. Research indicates that eating is affected not only by physiological hunger but also by psychological factors such as stress and emotions, social interactions, and environmental cues.[88], [89] The complicated nature of obesity necessitates a nuanced understanding of eating behaviors and surrounding influences. Recognizing this complexity, the health sector is increasingly embracing innovative technologies like wearable sensors and EMA to detect, monitor, and interpret eating behaviors.[90], [91] These tools offer promising avenues for capturing detailed data on dietary intake and contextual factors, enabling more personalized and effective interventions. By leveraging these advancements, we can gain valuable insights into overeating patterns and develop strategies that address the multifaceted nature of eating behaviors, ultimately advancing our efforts to combat obesity.

Findings from our study using EMA data revealed that light refreshments, pre-meal biological hunger, perceived overeating, and pleasure-driven desire for food were key factors influencing overeating. Our results demonstrate that biological hunger and pre-meal pleasure-driven desire (i.e., appetite) for food reflect distinct influences on eating behavior. Biological hunger is a physiological drive based on energy requirements, whereas pleasure-driven desire leads to food intake based on hedonic value. Prior research supports that despite satiation, hedonic factors contribute to excessive intake when highly palatable foods are available.[15], [92] Additionally, our finding of a negative association between light refreshments and overeating suggests that smaller meals may help pre-

vent overeating, aligning with existing literature.[93] Perceived overeating positively contributes to objective overeating, highlighting the role of self-awareness. Since perceived overeating was assessed post-meal, this suggests individuals are able to reliably reflect on their eating behavior post-consumption and may benefit from interventions that enhance this self-awareness, such as by using mobile health (mHealth) apps to track eating behaviors, provide feedback, and promote mindfulness when eating.[94], [95]

Given the significance of eating speed and bite and chewing patterns in influencing overeating,[96], [97] our analysis of bite and chew counts indicates the potential for defining data-driven thresholds beyond which the likelihood of overeating substantially increases, as demonstrated by SHAP values. Specifically, our results show that at thresholds of approximately 500 chews and 75 bites per meal, SHAP values increase rapidly, signifying an elevated risk of overeating (Supplementary Figure 3.7.8). These findings could inform early interventions targeting individuals who exceed these thresholds. Notably, the model's AUROC of 0.69 and AUPRC of 0.69 for passive sensing data alone demonstrate moderate predictive performance, supporting the potential for real-time feedback systems to encourage users to reflect on their eating behavior. Such systems could translate complex patterns—such as elevated chew rate or bite count—into simple, actionable feedback or nudges (e.g., timely reminders to slow down), while also providing an objective measure for evaluating the success of various behavioral nutrition interventions. Adding EMA features further enhanced the model's performance, increasing the AUROC to 0.86 and the AUPRC to 0.84, thus substantially improving predictive accuracy.

We also found that a high number of chews—around 500 or more—might indicate a prolonged meal duration, potentially leading to overeating due to extended exposure to food (see Supplementary Figure 3.7.8). Conversely, when the number of chews is rel-

atively low, approaching approximately 100, the SHAP values also increase, suggesting that a large amount of food has been consumed rapidly (see Supplementary Figure 3.7.8). This rise in SHAP values at lower number of chews corresponds to a higher bite rate, serving as a proxy for eating pace. A higher bite rate in these instances indicates that food is being consumed more quickly, further elevating the likelihood of overeating during these meals. Furthermore, our SHAP analysis of the chew-bite ratio (Supplementary Figure 3.7.8) indicates that a higher average number of chews per bite predicts a lower likelihood of overeating. This finding suggests that thoroughly chewing food might enhance satiety and reduce the risk of overeating while contributing to prior literature that report increased chewing promotes fullness and regulates appetite hormones.[98], [99]

Through clustering analysis, we uniquely identified distinct overeating phenotypes that align with well-documented behaviors in the literature, offering insights into the drivers of overeating and their potential implications for intervention strategies. The "Evening Craving" and "Stress-driven Evening Nibbling" phenotypes reflect circadian-driven and emotional eating behaviors, respectively. The "Evening Craving" phenotype, characterized by eating at night involving self-prepared meals and driven by biological hunger as a way to unwind, aligns with research on circadian rhythm disruptions influencing eating patterns. Studies have indicated that eating later in the day or night hours can lead to increased hunger and a preference for energy-dense foods.[100] For instance, Goel et al.[101] found that individuals with delayed circadian timing exhibited heightened appetite in the evening, which may promote late-night eating. This behavior is also consistent with findings on Night Eating Syndrome (NES), where individuals consume a significant portion of their daily intake during the night due to altered circadian eating rhythms.[102], [103] These findings suggest that interventions targeting meal timing, such as structured eating schedules and circadian-based dietary strategies, may help mitigate

the metabolic consequences of eating later in the day. Our operationalization of evening eating diverges from the traditional definition of NES, which involves nocturnal awakenings to eat due to hunger or emotional distress, accompanied by heightened stress levels before and after the meal. Given the limitations of our data, we categorized evening eating as consumption occurring between 5 PM and 6 AM, without the capacity to objectively assess nocturnal awakenings or associated stress responses.

The "Stress-driven Evening Nibbling" phenotype reflects emotional eating patterns widely reported in the literature. Emotional eating involves consuming food in response to negative emotions rather than physiological hunger, often leading to overeating, particularly of high calorie "comfort" foods. Research indicates that stress can elevate cortisol levels, which increases cravings for energy-dense foods and can trigger overeating later in the day.[104], [105] Researchers[106], [107] demonstrated that individuals under stress are more likely to engage in emotional eating, using food as a coping mechanism to alleviate negative feelings such as loneliness or anxiety. While most studies emphasize the impact of stress on increased eating, some research suggests that stress can also lead to reduced appetite in certain individuals.[105] We observed a group of meals that were not defined as overeating episodes but had some of the highest pre-and post-meal stress levels. These instances often involved snacking rather than consuming large meals, suggesting a pattern related to undereating or altered eating behaviors under stress. This indicates that stress responses can be heterogeneous, and individual differences may influence whether stress leads to overeating or undereating. Strategies like mindfulness-based stress reduction, and structured coping mechanisms have been shown to help regulate emotional eating behaviors and mitigate stress-related disruptions in appetite.[108]

Furthermore, the "Evening Restaurant Reveling" phenotype reflects the concept of

social facilitation of eating, where the presence of others affects food intake. Our results show that this phenotype specifically involves eating with family and friends, indicating that individuals are more likely to overeat in comfortable social settings. Studies have shown that individuals tend to consume more food when eating in groups with friends compared to eating alone or with strangers, due to extended meal duration and the influence of social norms. Prior literature suggests that individuals with obesity are less likely to overeat in groups where they feel less comfortable.[109] These factors contribute to increased portion sizes and caloric intake, as individuals adjust their eating to match that of their companions, a phenomenon known as behavioral mimicry.[110], [111], [112] Similarly, the "Take-out Feasting" aligns with research on convenience eating and the impact of readily accessible, energy-dense foods on consumption patterns. The accessibility of take-out or fast food, combined with social settings, has been shown to lead to overconsumption due to larger portion sizes and the high palatability of foods.[113], [114] Cohen et al.[115] discussed how environmental factors, such as the ubiquity of fast-food outlets and marketing strategies, contribute to automatic eating behaviors that override internal hunger cues, especially in social situations where food is a central component. In our analysis, we identified another distinct overeating phenotype—"Uncontrolled Pleasure Eating"—characterized by overeating for pleasure and a loss of control during tasks such as work or study. The inclusion of loss of control in this phenotype suggests a deeper psychological component, where external cues or emotional states trigger compulsive eating behaviors.[116] Environments associated with work or study may contribute to this behavior, as cognitive load can impair self-regulation, leading to mindless eating and a diminished ability to control food intake.[117] Overall, these phenotypes highlight the multifaceted nature of overeating behaviors, encompassing circadian misalignment, emotional stress, social influences, environmental convenience, and psychological factors such as loss of control. Understanding these distinct patterns is essential for designing

targeted interventions that address the underlying mechanisms of overeating and support more effective treatments to reduce overeating.

The phenotypes were derived at the meal level, as our primary objective was to characterize meal-based overeating patterns. Because individuals may exhibit multiple overeating phenotypes across different meals, a single individual-level classification can obscure the nuanced interplay of factors, which are shown by prior research to overlap rather than remain mutually exclusive.[118], [119] While this study focuses on meal-level classification, future work should explore personalized phenotype trajectories over time, potentially leading to more individualized frameworks for identifying and intervening on shifting overeating behaviors. Another avenue of research involves examining these phenotypes in relation to clinically relevant variables (e.g., BMI). Because all participants in this study met obesity criteria, BMI variability was limited, precluding robust analyses of phenotype-BMI associations. Future work could assess whether specific phenotypes correlate with broader health markers or weight trajectories, offering deeper insights into the underlying drivers and progression of overeating and obesity.

A strength of this research was the use of a personalized, objective measure of overeating. Overeating in the literature is often defined subjectively, relying on an individual's perception of whether someone consumed more than needed,[72] and while this is important, it is long known that subjective overeating is fraught with recall and participant bias, resulting in errors.[120] Moreover, overeating is typically reported over extended periods—days, weeks, or months—with little attention given to overeating at the meal level. Although overeating over long periods contributes to weight gain, with the promise of sensors[121] and EMA timely interventions become feasible, and without a solid meal-level definition of overeating, we may not be able to identify when, where, and how to

intervene effectively. We define overeating episodes as those where an individual's energy intake exceeds their personal meal/snack average by one standard deviation. Fixed thresholds[122] (e.g., meals over 1,000 kCal) do not adjust for differences in BMI, sex, or typical eating habits, potentially misclassifying normal consumption as overeating for some individuals while overlooking it in others. Our method is also consistent with participants' own perceptions of overeating, as people often consider they have overeaten when they consume more than what is typical for them personally.[123] This approach assumes relatively stable meal patterns and energy intake over the 14-day study period. Future studies of longer duration might segment data into shorter intervals for recalculating z-scores or employ dynamic methods (e.g., rolling window averages, adaptive Bayesian models) to continuously update reference distributions and account for evolving eating behaviors over time.

A further strength of this work was employing advanced machine learning techniques using a clustering approach that captures complex data representations, enabling deeper insights into nuanced eating patterns. While clustering algorithms are inherently sensitive to data characteristics, we mitigated potential limitations by leveraging a DNN encoder to extract nonlinearities from the features to learn latent representations. Additionally, UMAP allows for preserving the intrinsic data topology without distorting the overall data structure. Through two-dimensional projection visualization, we confirmed cluster separability with minimal overlap. While we acknowledge the necessity of testing our method on an independent dataset to assess its generalizability, our novel pipeline effectively addresses common clustering pitfalls, enhancing the reliability and validity of our findings. Moreover, we note that addressing class imbalance with SMOTE can introduce synthetic examples that may not fully represent the true data distribution. While critiques of SMOTE are valid, our Supplementary Note 3.7.3 provides further analyses in-

dicating its continued utility in this context.[124]

Our study additionally distinguishes itself by observing individuals with obesity in their natural settings, encouraging them to maintain their usual routines. This approach captures authentic, often problematic eating habits, which we redefine using digital longitudinal data. By characterizing distinct overeating phenotypes via clustering methodologies, we can stratify individuals during a run-in phase, enabling more targeted phenotype-specific interventions that improve both effectiveness and scalability in addressing obesity. Although we prioritized accurate caloric intake estimates via 24-hour recalls collected by dietitians, the actual classification of overeating episodes was data-driven, allowing the process to be automated for broader adoption. For large-scale implementation, mobile applications and self-reporting tools could estimate caloric intake while wearable devices passively track behavioral and physiological signals, reducing user burden. Future work could evaluate the trade-offs between accuracy and feasibility to ensure that these automated methods align with established dietary assessments, enabling timely, phenotype-specific interventions to help individuals reduce their propensity to overeat.

Despite these notable strengths, some additional limitations warrant discussion. Reactivity to continuous measurement is a recognized concern in studies using wearable cameras and EMA methodologies. To address this, participants were instructed to “act naturally” and maintain their usual routines, and the devices were designed to be small, lightweight, and require minimal user input, thus reducing awareness of data collection. The two-week study duration also allowed participants to acclimate to the procedures, further diminishing potential reactivity. Also, prior research suggests that activity-oriented cameras can reduce perceived surveillance and social discomfort, helping to

mitigate reactivity.[125] Nevertheless, some level of reactivity is inevitable, even with these precautions. EMA can reduce biases compared to other self-report methods,[4] yet it remains vulnerable to underreporting or stigma.[126] Integration of wearable sensors and cameras provides an objective means to validate EMA by capturing micromovement patterns, meal timing, dietary composition, and other consumption behaviors. Discrepancies between self-reported EMA data and sensor-derived metrics may indicate inaccuracies arising from stigma or recall bias. Moving forward, triangulating data from EMA, passive sensing, and dietitian-led recalls can yield a more comprehensive and validated dataset that improves our understanding of human behavior. In conclusion, there has been growing interest in examining phenotypes of overeating. Our study supports a promising new direction by utilizing both EMA and sensor data to lay the foundation for testing timely personalized interventions that mitigate overeating.

3.6 Summary

This chapter investigated overeating as a dynamic, context-dependent behavior that contributes to obesity and remains difficult to address using traditional behavioral interventions. Existing approaches often rely heavily on self-report or focus on isolated predictors such as stress, cravings, or loss of control, which limits the ability to understand how psychological, contextual, and behavioral factors interact during real-world eating episodes. To address this gap, this study used longitudinal, multimodal data from the SenseWhy study to characterize and predict overeating in free-living conditions.

The SenseWhy study monitored adults with obesity over a two-week period using a combination of ecological momentary assessment (EMA), an activity-oriented wearable camera, mobile app-based dietary reporting, and dietitian-administered 24-hour recalls. Of

the participants who began the study, 48 adults were included in the final analysis, contributing 2,302 meal-level observations. The study also included detailed passive sensing data derived from 6,343 hours of wearable camera footage across 657 days, from which eating-related micromovements such as bites and chews were manually annotated. Overeating was operationalized at the meal level using a personalized threshold, where an overeating episode was defined relative to each participant's usual intake rather than by a fixed calorie cutoff.

The first analytic objective was to predict objectively defined overeating episodes using supervised machine learning. Three feature sets were evaluated: EMA-only features, passive sensing-only features, and a feature-complete model combining EMA and passive sensing data. XGBoost was selected as the best-performing classifier and achieved strong performance in the feature-complete setting, with a mean AUROC of 0.86 and a mean AUPRC of 0.84. EMA-only features also performed well, while passive sensing-only features showed moderate predictive performance, suggesting that both self-reported psychological/contextual data and objective eating behavior signals provide complementary information for detecting overeating.

Explainability analyses using SHAP identified the most important predictors of overeating across the different feature sets. In the EMA-only model, light refreshments, biological hunger, perceived overeating, evening eating, and pleasure-driven desire for food were among the most influential predictors. In the passive sensing-only model, the number of chews, chew interval, chew-bite ratio, number of bites, and chew rate were most predictive. In the combined model, perceived overeating, number of chews, light refreshments, loss of control, and chew interval were key contributors. These results indicate that overeating is not explained by a single behavioral or psychological factor, but instead

emerges from the interaction of subjective experience, meal context, and objective eating microstructure.

The second analytic objective was to identify distinct overeating phenotypes using a semi-supervised clustering pipeline. The clustering approach used EMA-derived features to generate latent representations, followed by dimensionality reduction and cluster analysis. This analysis identified five clinically interpretable overeating phenotypes: “Take-out Feasting,” “Evening Restaurant Reveling,” “Evening Craving,” “Uncontrolled Pleasure Eating,” and “Stress-driven Evening Nibbling.” These phenotypes captured different combinations of contextual and psychological factors, including restaurant-sourced meals, evening eating, social dining, biological hunger, hedonic eating, stress, loneliness, task-oriented distractions, and loss of control.

This work contributes to the digital health and obesity literature by demonstrating that overeating can be detected and phenotyped using multimodal longitudinal data collected in naturalistic settings. Rather than treating overeating as a uniform behavior, this study showed that overeating occurs through multiple pathways involving distinct combinations of affective, contextual, social, and behavioral factors. The chapter also advances the field by operationalizing overeating at the personalized meal level and by showing that interpretable machine learning can identify actionable predictors of overeating. These findings provide a foundation for future just-in-time adaptive interventions that can deliver phenotype-specific support, such as prompts to slow eating pace, regulate stress, manage evening cravings, or modify high-risk social and restaurant-based eating contexts.

3.7 Supplementary

3.7.1 Supplementary Note 1

Wear time, adherence, and participant feedback for the SenseWhy wearable camera

At the end of the study, participants were asked to rate the burden of the SenseWhy wearable camera by completing the Difficulty of Use, Physical, Time & Social, Mental & Emotional, and Privacy subscales of the User Burden Scale (UBS).[127] Each of the 18 questions was on a 5-point Likert-style scale, with 1 being least burdensome and 5 being most burdensome. Mean participant responses to each subscale fell in the C–D range of the A–F scoring system used in interpreting the UBS. The UBS is typically used to evaluate software, so while the SenseWhy wearable—a hardware product—is rated as substantially more burdensome than the software applications the UBS is validated on, the fact that no composite mean exceeds 2 on the 1–5 burden scale indicates a favorable outcome for user acceptability of the SenseWhy wearable camera.

Participants also answered the open-ended question, “What improvements would you like to see in this device?” These responses were thematically coded against the UBS subscales. 2 independent coders started with identical spreadsheets containing all participant responses, a tally column for each UBS subscale, and a notes column. Coders first extracted words/phrases in participant responses that corresponded to specific burden/design considerations and added them to the notes column with light standardization (e.g., “not as big” and “less bulky” both recorded as “smaller”). Coders then deductively paired these initial codes with UBS subscales, adding a tally for each correspondence (e.g., “smaller” corresponds to “physical” subscale). Coders then met to compare initial

code lists and resolve disagreements, updating the subscale tallies to reflect these collaborative revisions. The coders then independently reviewed the updated codebook and inductively merged closely related codes into more general codes (e.g., “stronger magnet” and “better magnet placement” merged into “better magnet”), producing condensed code lists. Not all codes were merged—those deemed sufficiently distinct with respect to burden and design considerations were not changed during this round.

Coders then met again to compare condensed code lists and rationales for merging or not merging codes, deliberating until full agreement was reached. Incidentally, changes made during inductive code merging did not affect the subscale tallies. Across all participants’ responses, 72 thematically codable statements were extracted, coded, and iteratively refined into a final set of 18 codes grouped by pertinent UBS subscale such that subscale tallies represent the types of burden most commonly experienced by participants and codes within each subscale grouping represent the sources of said burden.

The majority of the open-ended responses concerned physical and difficulty-of-use aspects of the SenseWhy wearable camera design. Certain frequently used words and phrases offer a more specific indication of the types of improvements participants would like to see in future iterations of the device: “smaller”, “lighter”, “prettier”, “more comfortable”, and “better magnet.”

3.7.2 Supplementary Note 2

Sensitivity analysis of clustering methods comparing K-means and Gaussian Mixture Modeling

The sensitivity analysis using Gaussian Mixture Modeling (GMM) produced results that were consistent with the K-means approach. GMM yielded an entropy score of 0.39, a

purity of 80.5%, and a silhouette score of 0.58, indicating a slightly higher level of cluster uncertainty compared to K-means.

K-means produced an entropy score of 0.36, a purity of 81.4%, and a silhouette score of 0.59, showing marginally better-defined cluster boundaries. Despite these small variations, both methods converged on a five-cluster solution for overeating behaviors.

GMM, unlike K-means, models data as a mixture of Gaussian distributions, allowing for probabilistic cluster assignments rather than hard assignments. Despite this methodological difference, the overall clustering structure remained stable, reinforcing the robustness of the identified phenotypes across different clustering approaches.

3.7.3 Supplementary Note 3

Comparison of SMOTE and downsampling for addressing class imbalance

Our dataset exhibits a moderate class imbalance, with the minority class comprising approximately 18% of the total samples. We evaluated two approaches to address this imbalance: downsampling the majority class and applying SMOTE. Downsampling achieved an AUROC of 0.81 (0.03), AUPRC of 0.79 (0.03), and a Brier score of 0.26 (0.04). In comparison, SMOTE substantially improved performance, yielding an AUROC of 0.86 (0.04), AUPRC of 0.84 (0.04), and a Brier score of 0.11 (0.02).

These results demonstrate the utility of SMOTE in preserving critical context from the majority class while effectively addressing imbalance, leading to better overall model performance. Although calibration concerns with oversampling methods have been noted, the substantial improvement in the Brier score highlights the enhanced reliability of our model. Further calibration, if required, can be addressed using post-hoc techniques like

isotonic regression or Platt scaling. Thus, SMOTE remains the preferred approach for our dataset and analysis.

3.7.4 Supplementary Note 4

Annotation protocol and quality control

1. Definitions and Initial Annotation Step

Eating episodes were defined as consecutive feeding gestures with intervals not exceeding 15 minutes. Isolated drinking episodes were annotated separately. Chews were characterized as consecutive small jaw open/close movements, while bites were defined as larger jaw open/close movements coinciding with food entering the mouth. Meal start and end times were initially annotated by our research team in an iterative process. These annotations were cross-referenced and validated using timestamped image records from the application to ensure accuracy.

2. Identifying a Professional Labeling Company

The annotation of micro-movements associated with chews and bites was outsourced to a third-party professional labeling service, which provided a dedicated project manager and a team of annotators. The company implemented strict quality standards and auditing protocols, utilizing a system capable of providing real-time quality reports to ensure high-quality annotations.

3. Rater Training and Team Synchronization

The project manager and select annotators underwent a training program designed by our research team (FS, BW, CR, MP, and NA). The training included reviewing labeled example footage during interactive Zoom sessions, participating in hands-on annotation sessions with detailed feedback provided by our research team, and familiarization with

the annotation protocol and expected standards.

4. Quality Control Measures

Inter-rater agreement was managed internally by the labeling company through their auditing system, ensuring consistent quality across raters. Intra-rater agreement was assessed by having a subset of annotations independently reviewed by our research team at different stages to confirm adherence to the established protocol. Periodic checks were performed on batches of data to maintain high annotation standards. Upon receiving all annotated data, our research team (FS, BW, CR, MP) conducted a final quality assurance process by randomly selecting and reviewing meal footage to ensure alignment with the protocol and the reliability of the annotations.

3.7.5 Supplementary Note 5

Participant-level splitting analysis

To assess whether the meal-based and participant-based splitting strategies produced comparable results, we replicated the analysis using a participant-level splitting approach. In this method, participants were assigned exclusively to the training, validation, or test sets, ensuring that no meals from the same participant appeared in multiple sets. Participants were randomly allocated to each set while maintaining a 60/20/20 distribution of meals and preserving the ratio of overeating to normal meals across all splits.

Using the participant-level splitting strategy, the model achieved an AUROC of 0.83, which closely aligns with the performance observed in the meal-based split (mean AUROC = 0.86, SD = 0.03) and shows no substantial difference between the two approaches. Similarly, the AUPRC score was 0.82, which remained consistent with the results from the meal-based split (AUPRC = 0.85, SD = 0.03). These findings indicate that the model

performs consistently and comparably across both splitting strategies, supporting its reliability and generalizability.

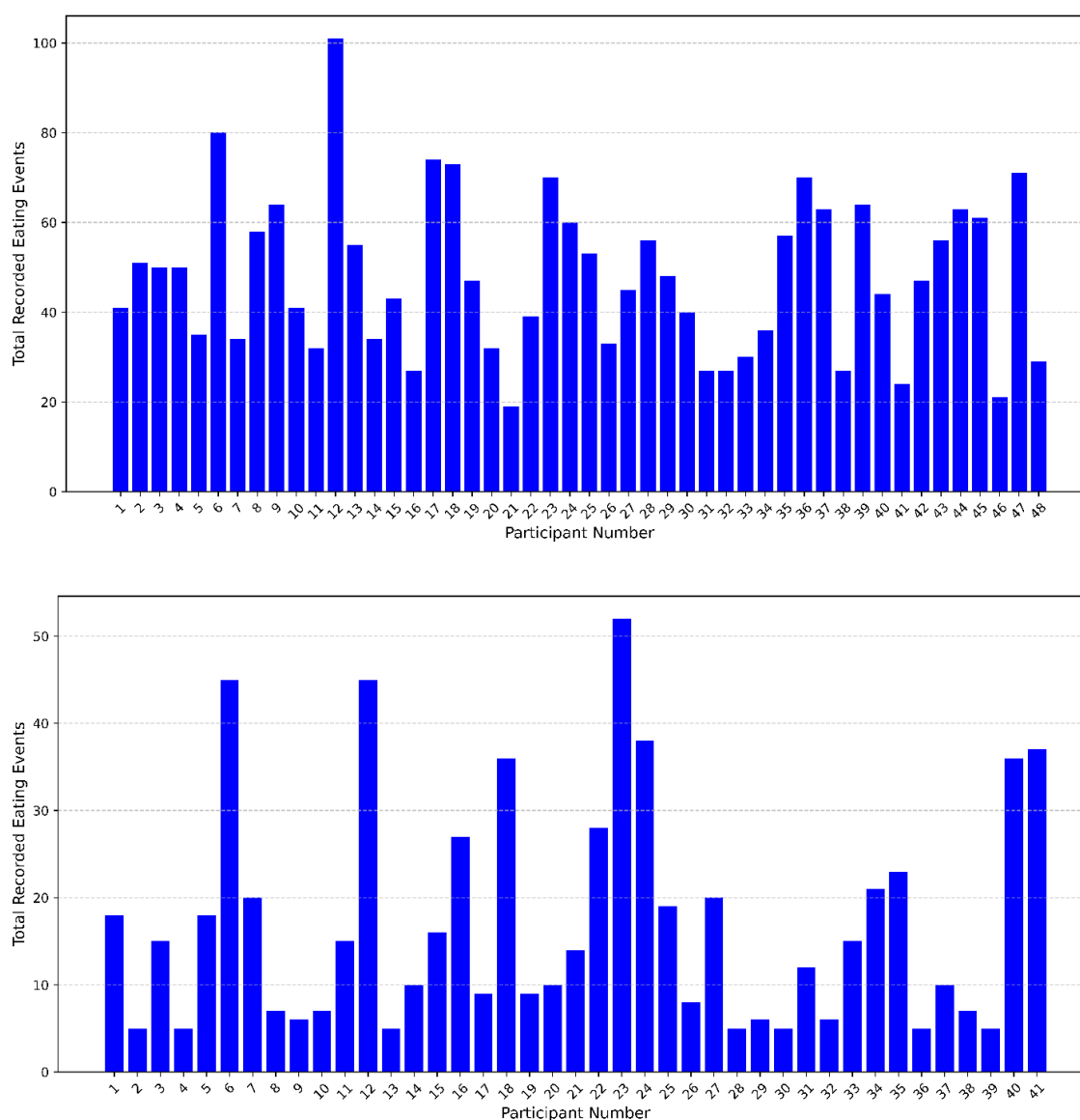


Figure 3.7.1: Distribution of meal counts across participants for EMA-only and feature-complete datasets. (a) Meal count distribution for the EMA-only dataset, with a mean of 48 meals per participant. (b) Meal count distribution for the feature-complete dataset, with a mean of 17 meals per participant.

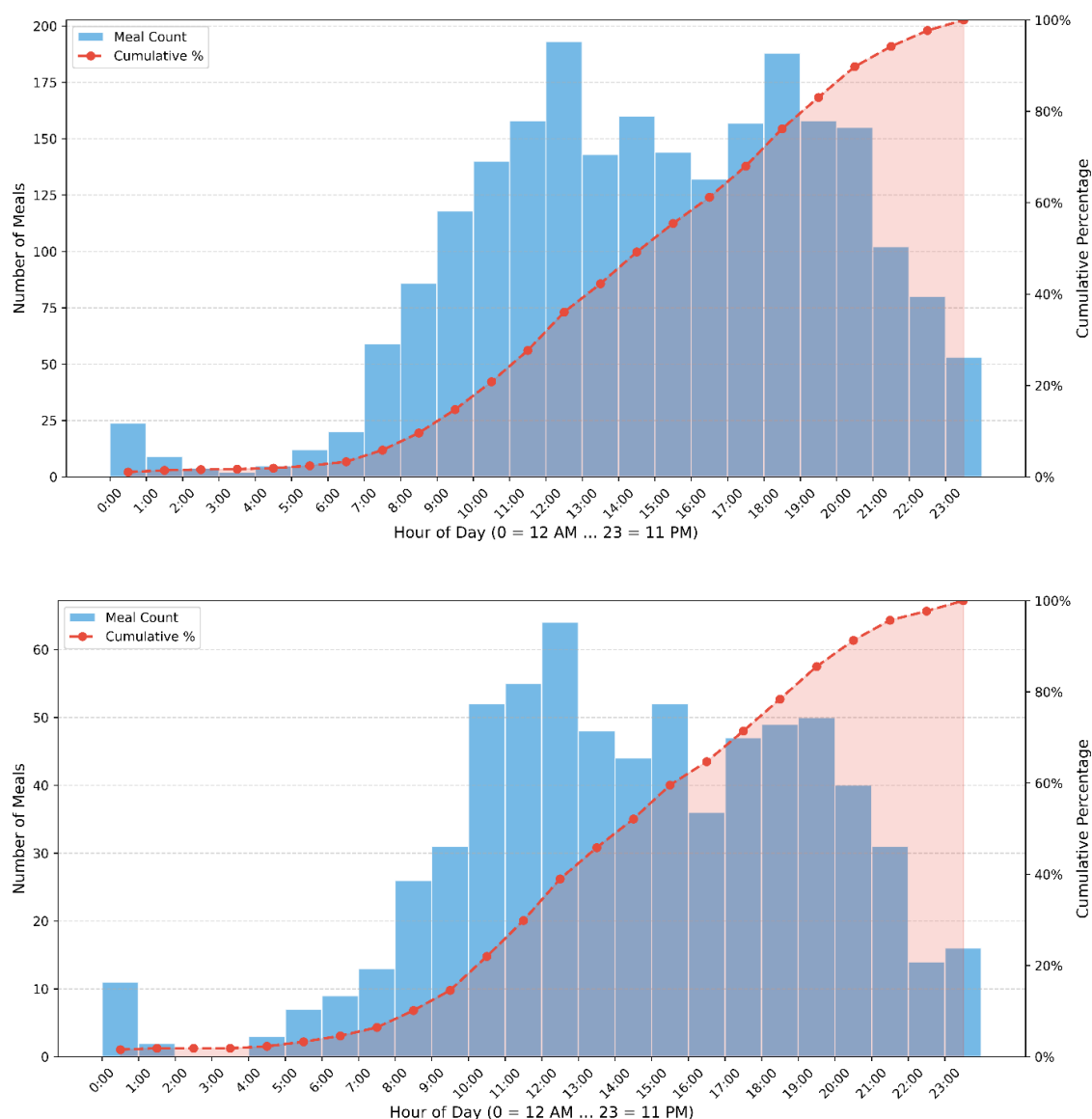


Figure 3.7.2: **Hourly distribution of meals across the day for all participants.** (a) Distribution based on EMA-only data. (b) Distribution for meals with feature-complete set (EMA and passive sensing features). The blue bars represent meal counts per hour, while the red dashed line shows the cumulative percentage.

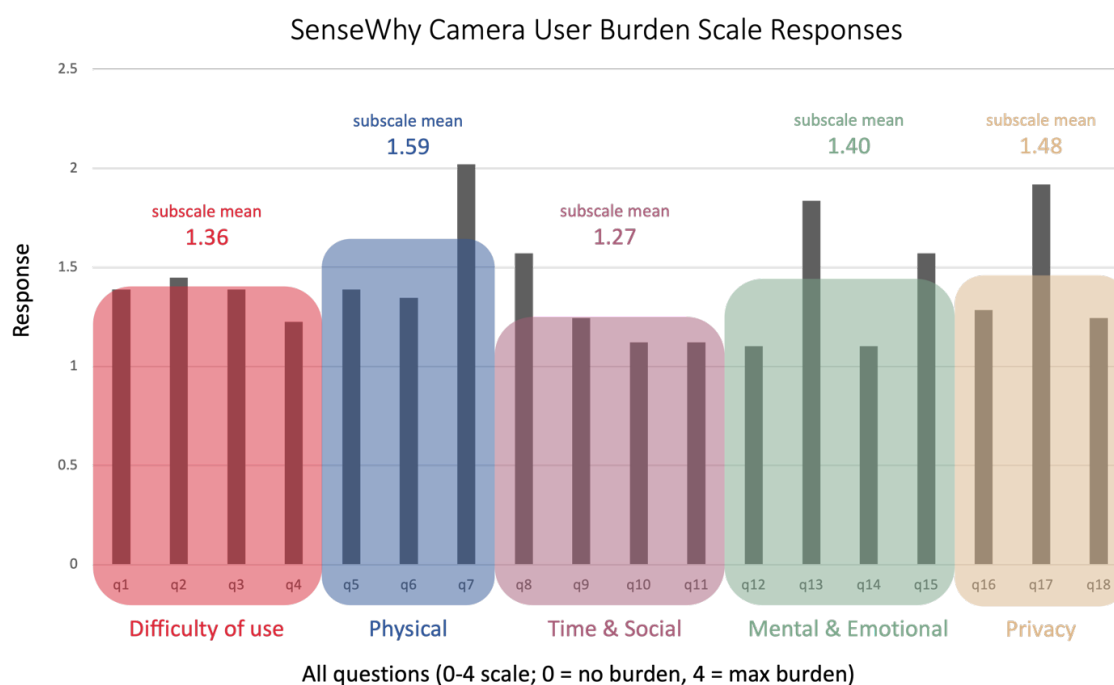


Figure 3.7.3: Mean responses to UBS inventory for the SenseWhy Camera wearable. Subscale composite means are given in color, i.e., the Difficulty of Use subscale (comprising questions 1–4) composite mean is represented by the height of the red box on the *y*-axis and is also given in the red text above the box.

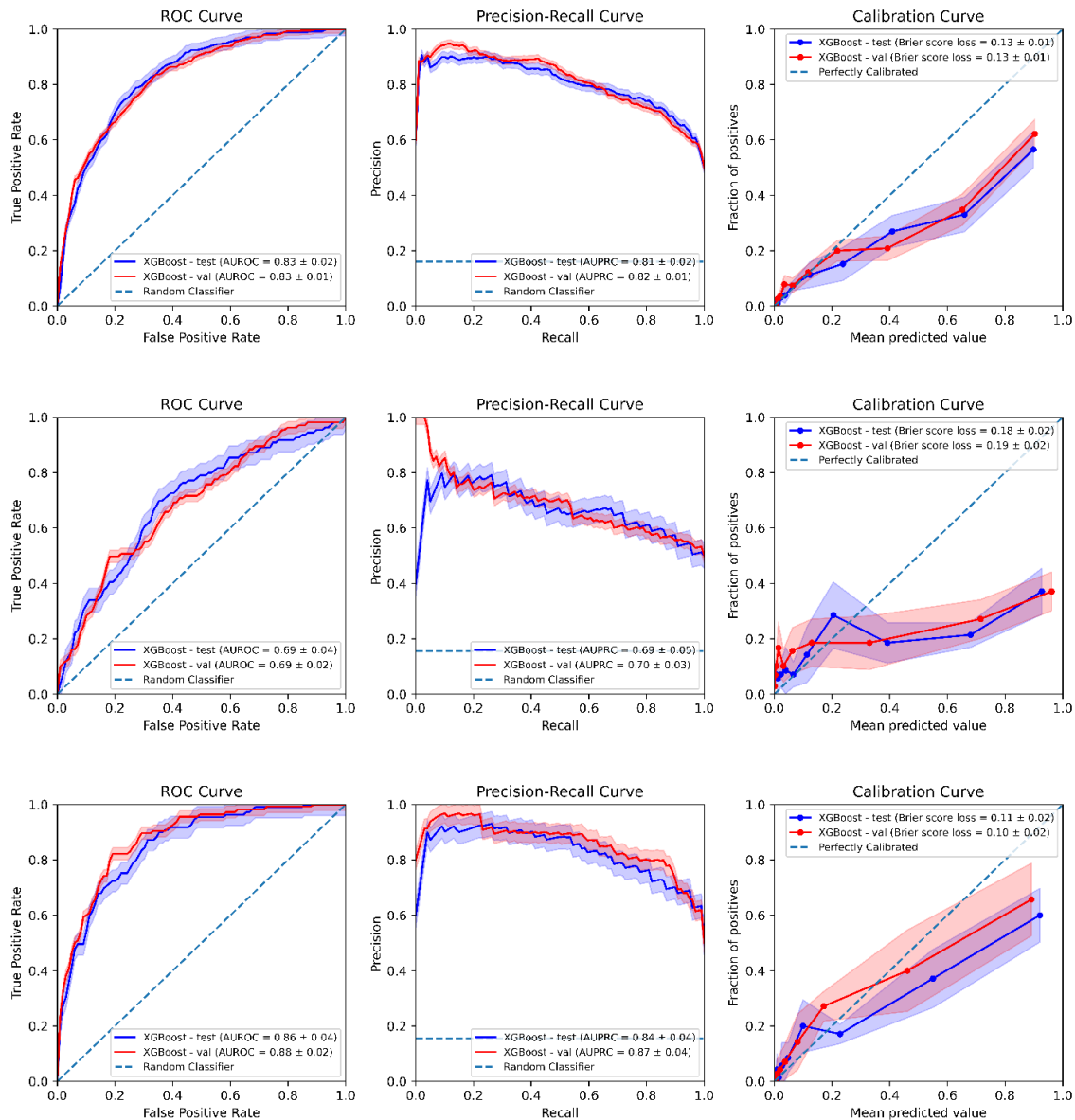


Figure 3.7.4: Performance evaluation of the XGBoost model on validation and test sets using ROC curves (left), Precision–Recall curves (center), and Calibration curves (right) for EMA-only (a), passive sensing-only (b), and feature complete (c) datasets. The similar AUROC, AUPRC, and Brier score loss values between the validation and test datasets indicate generalization and minimal overfitting. Shaded regions represent the standard deviation across iterations.

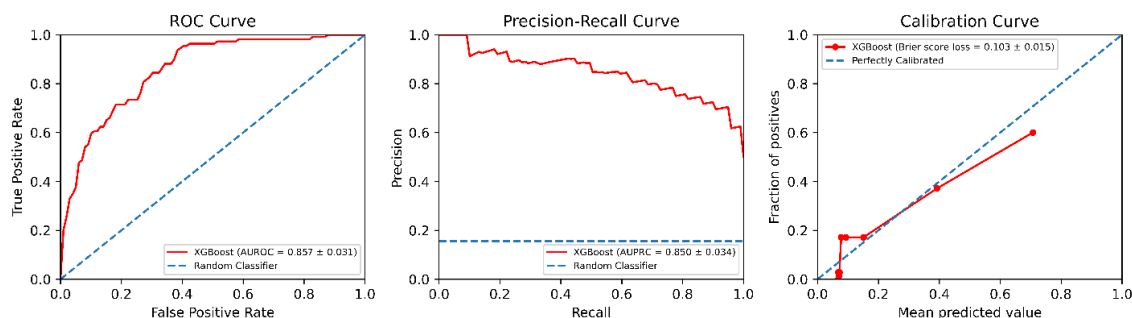


Figure 3.7.5: Calibration curve for the XGBoost classifier after applying Platt's scaling, a post-calibration method using a sigmoid function to refine predicted probabilities. This approach fits a logistic regression model to the classifier's outputs, improving the alignment of predicted probabilities with actual class distributions. The analysis reveals a reduction in the Brier score from 0.114 ± 0.017 (pre-calibration) to 0.103 ± 0.015 (post-calibration), indicating enhanced model calibration and reduced error in probability estimates.

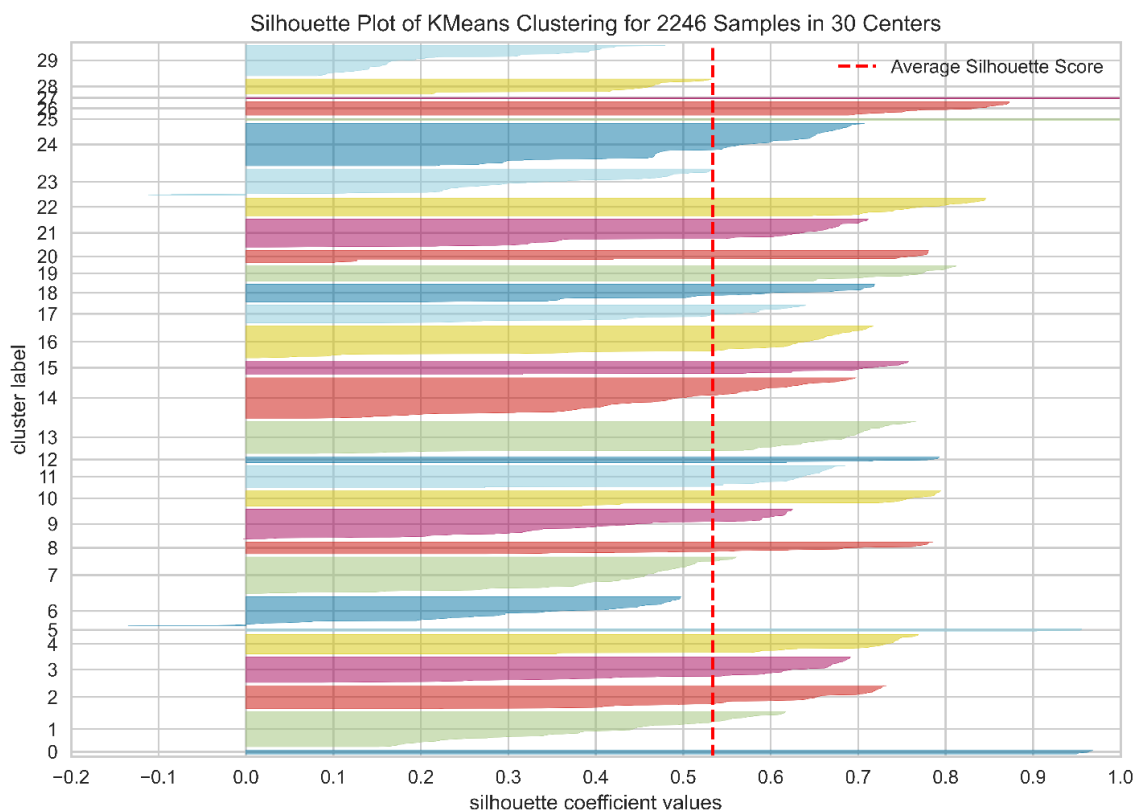


Figure 3.7.6: Silhouette plot illustrating the silhouette coefficient values for all meals across all clusters derived from K-Means clustering. Each bar represents the silhouette score for individual samples within a cluster, sorted by decreasing score. The red dashed line indicates the average silhouette score, providing a measure of overall clustering quality.

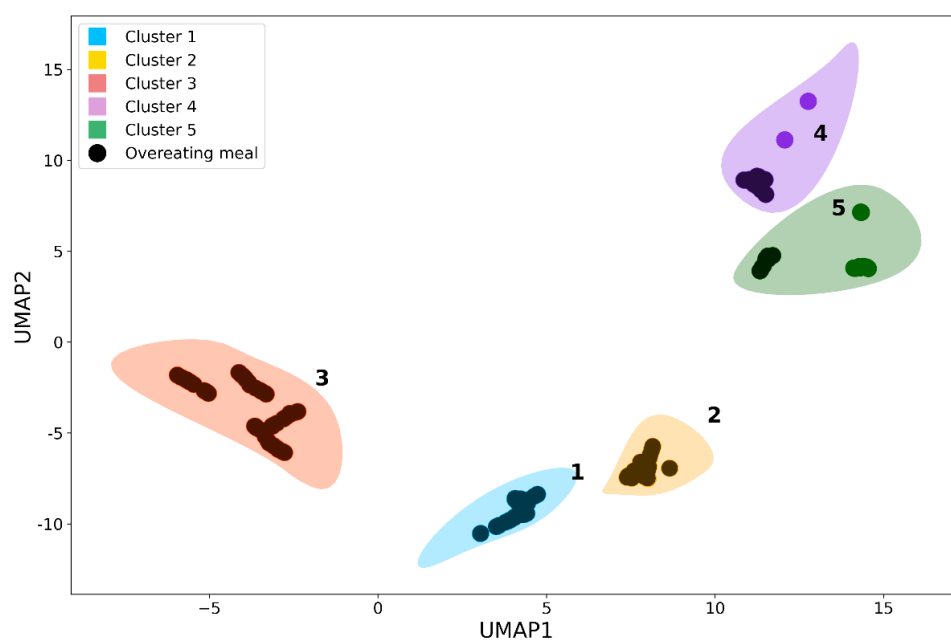


Figure 3.7.7: **Overeating clusters visualization in 2 dimensions using UMAP.** Each cluster is represented by a unique color, with clusters 1 through 5 labeled. Overeating meals are highlighted as black points, showing their distribution across clusters. This visualization highlights the clustering patterns and the association of overeating behaviors within the feature space.

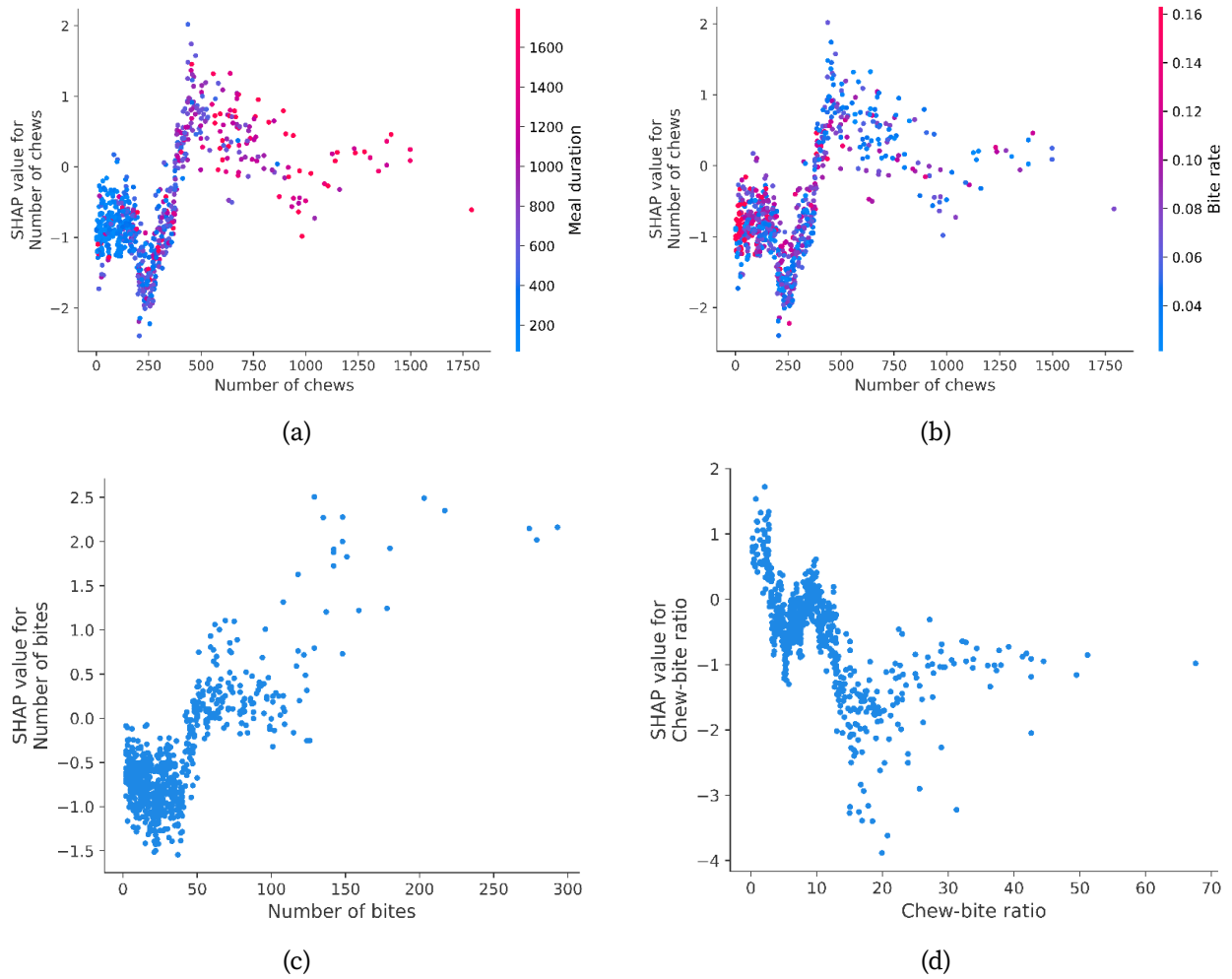


Figure 3.7.8: SHAP value analysis for features associated with overeating behavior (i.e., number of chews, number of bites, and chew-bite ratio). (a) SHAP values for the number of chews, colored by meal duration, indicating the influence of prolonged chewing patterns on overeating. (b) SHAP values for the number of chews, colored by bite rate, highlighting the interaction between chewing intensity and bite rate. (c) SHAP values for the number of bites, demonstrating their contribution to predicting overeating. (d) SHAP values for the chew-to-bite ratio, showing its inverse relationship with overeating.

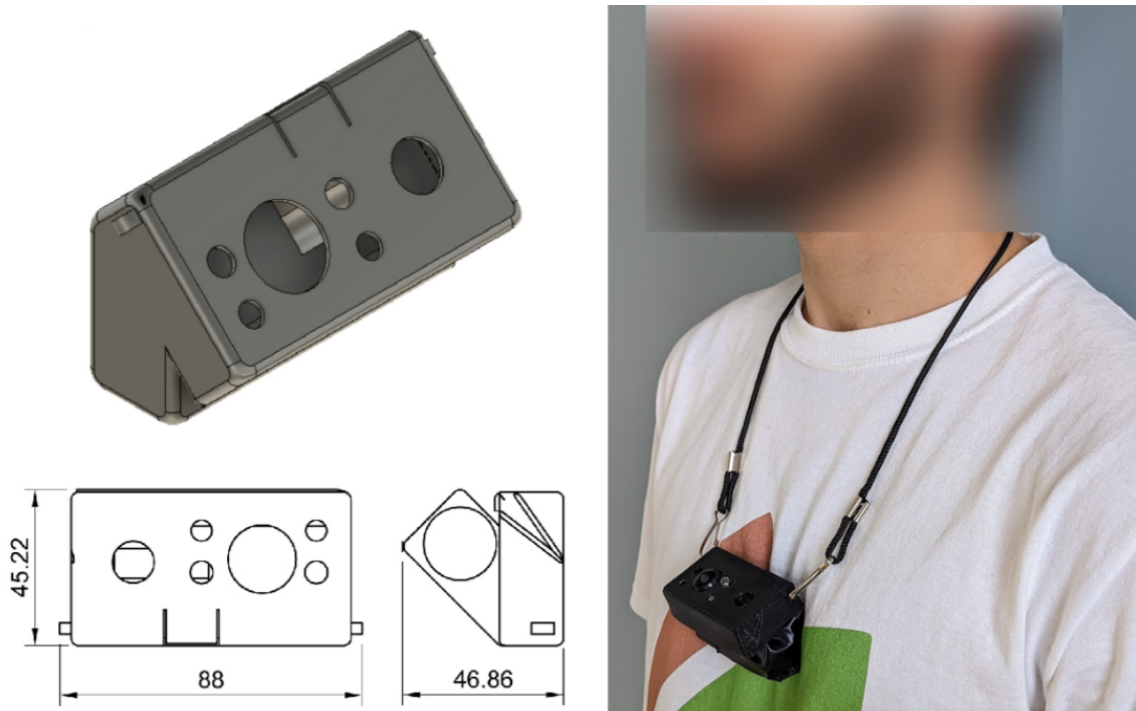


Figure 3.7.9: The design and usage of the sensing suite, optimized for user comfort and functionality. The device has dimensions of 88 mm $\times$ 46.86 mm $\times$ 45.22 mm (length $\times$ width $\times$ height) and a weight of 106 grams, making it compact and lightweight for everyday use. Its total volume is approximately 102.18 cm³. To ensure consistent placement, the sensing suite is equipped with an adjustable neckband, allowing for precise calibration to maintain stable positioning on the user's body. Additionally, the device features a magnetic back plate for secure attachment and flexibility in adjustments. These design elements enable users to fine-tune the device's position, ensuring optimal alignment with the body and minimizing movement artifacts during data collection. This design allows the device to remain stable and functional in various real-world scenarios, contributing to the robustness and reliability of the collected data. The right panel of the figure demonstrates how the device is worn on the body, while the left panel provides detailed dimensions and a technical schematic of the sensing suite.

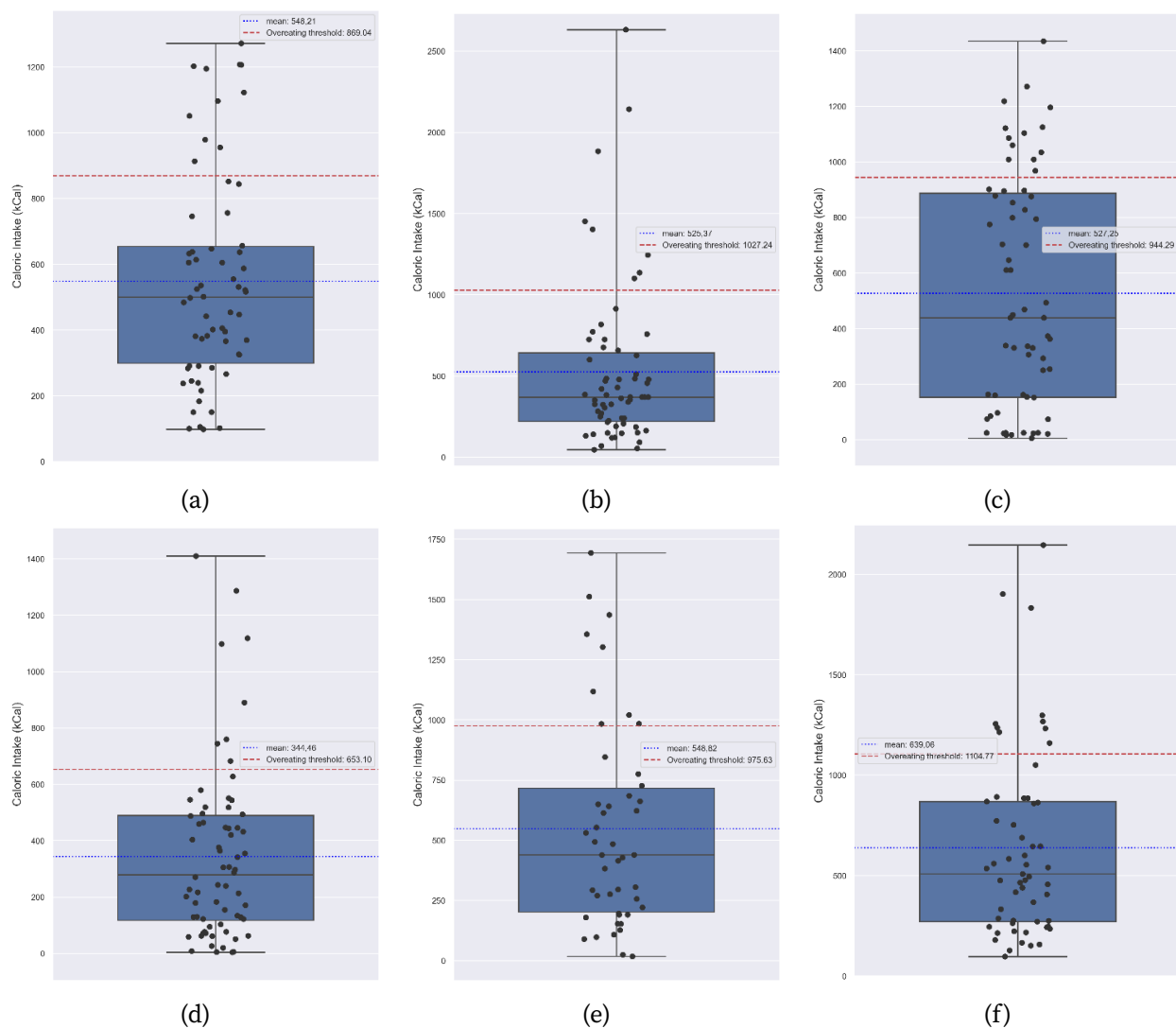


Figure 3.7.10: Eating profile for 6 participants over 14 days with each dot representing a single meal. The blue dashed line indicates the mean caloric intake, while the red dashed line represents the overeating threshold. The plot captures the variability and frequency of meals exceeding the overeating threshold, displaying individual differences in eating patterns. The overeating thresholds for each panel are as follows: (a) 869.04 kCal, (b) 1027.24 kCal, (c) 944.29 kCal, (d) 653.10 kCal, (e) 975.63 kCal, and (f) 1104.77 kCal.

Participant characteristic, mean (range)	
Age	44.0 (21–66)
BMI (kg/m ²)	36.9 (29.8–58.7)
Sex, <i>n</i> (%)	
Female	32 (78.0)
Male	9 (22.0)
Race, <i>n</i> (%)	
White	18 (43.9)
Black	15 (36.6)
Unknown	5 (12.2)
Asian	3 (7.3)
Ethnicity, <i>n</i> (%)	
Non-Hispanic	35 (85.4)
Hispanic	6 (14.6)
Objective overeating, <i>n</i> (%)	
Non-overeating	591 (84.4)
Overeating	109 (15.6)
EMA-based features, <i>n</i> (%)	
<i>Location</i>	
Home	623 (89.0)
Work	32 (5.0)
Other	29 (4.0)
Restaurant	11 (2.0)
School	5 (1.0)
<i>Food source, <i>n</i> (%)</i>	
self-cooked from scratch	221 (32.0)
Snacks/cereal	204 (29.0)
Self-cooked from frozen/prepackaged meal	108 (15.0)
Bought from restaurant—Takeout	91 (13.0)
Delivery	31 (4.0)
No food, just beverage	20 (3.0)
Bought from restaurant—Dine in	14 (2.0)
Self-cooked from ingredient delivery service	10 (1.0)
<i>Activity co-occurrence, <i>n</i> (%)</i>	
Watching TV	308 (44.0)
Working/studying	146 (21.0)
Socializing	128 (18.0)
Nothing else	109 (16.0)
Driving	9 (1.0)

Table 3.7.1: Summary statistics for the analytic subsample

Table 3.7.1 continued	
EMA-based features, mean (range)	
Hedonic Eating	3.12 (0.0–4.0)
Desire for Food	2.93 (0.0–4.0)
Pleasure-driven Desire for Food	2.89 (0.0–4.0)
Pre-meal Biological Hunger	2.73 (0.0–4.0)
Pre-meal Emotion (calm)	2.33 (0.0–4.0)
Pre-meal Emotion (upbeat)	1.82 (0.0–4.0)
Loss of Control (LOC)	0.99 (0.0–3.0)
Pre-meal Stress	0.96 (0.0–4.0)
Post-meal Stress	0.83 (0.0–4.0)
Perceived Overeating	0.58 (0.0–4.0)
Pre-meal Emotion (lonely)	0.57 (0.0–4.0)
Cognitive Restraint	0.31 (0.0–3.0)
Sensor-based features, median (IQR)	
Number of bites	33.0 (15.0–48.8)
Number of chews	293.0 (120.0–413.0)
Meal Duration, sec	604.0 (267.25–871.25)

Table 3.7.1: Summary statistics for the analytic subsample (Cont.)

Summary statistics for the analytic subsample of participants ($n = 41$) who provided both EMA data and corresponding passive sensing features (feature-complete) for their reported meals. This table includes baseline demographic characteristics, EMA response metrics, and descriptive statistics of derived sensor-based features. These participants represent the subset of the full study sample ($n = 48$) for whom complete feature data were available. * Data are n (% of the participant sample), mean (range), or median (interquartile range).

Thematic Analysis: “What improvements would you like to see in this device?”			
Theme (total n*)	Description	Code**	n
Physical (n=40)	Burden/discomfort related to the physical form of the device.	Smaller	20
		Lighter	8
		More secure	5
		Better construction	4
		More comfortable	2
		Less noticeable (to me)	1
Difficulty of Use (n=18)	Mismatch between system and user’s intuitive abilities, general operating difficulty	Better magnet	6
		Better battery	3
		Easier to wear	3
		Less cumbersome	3
		Waterproof	2
		Easier power switch	1
Social (n=10)	Disruption of/influence on social interactions and/or relationships	Less noticeable	6
		More attractive	2
		Color options	2
None (n=8)	User suggested no improvements to system	N/A	8
Mental (n=3)	Burden of user’s attention/concentration	Clearer status LEDs	3
Privacy (n=1)	Risk of system capturing/sharing information user would prefer not to share	More apparent FOV	1
<i>*some participants’ responses included multiple thematic codes, hence why $n[\text{participants}] \neq n[\text{codes}]$</i> <i>**codes represent direct quotations and paraphrases (e.g., “more discreet” was coded as ‘Less noticeable’)</i>			

Table 3.7.2: Thematic analysis of open-ended post-study question asking participants for suggestions to improve the SenseWhy Camera wearable. Codes represent specific improvements proposed for the device. Sets of codes with similar user/design considerations are categorized together under broader themes. Themes are adapted from UBS subscales, helping us interpret quantitative UBS results to inform design priorities. Note on adaptation: UBS subscale scopes are reduced in accordance with participant provided codes, e.g., UBS subscale “Time and Social” became the theme “Social” because no participant response related to time-based device burden.

Table 3.7.3: Feature descriptions

Feature	Description	Data	Source	Question (self-reported)
<i>Behavioral</i>				
Bite	Large jaw open/close when food enters the mouth	Count and rate of bites per meal	Visual (IR-AOD)	NA
Chew	Consecutive small jaw open/close, followed by bite	Count and rate chews per meal	Visual (IR-AOD)	NA
Eating duration	Time between end and start of given eating episode	Length of feeding gesture per meal	Visual (IR-AOD), Self-report (Phone, 24-hour recall)	NA
<i>Contextual</i>				
Location	Type of the location: work, home, school, restaurant, and other	Categorical response	Self-report (Phone)	Where did you eat?
Food Source	Type of food is: cooked from scratch, cooked from ingredient, delivery service, cooked from frozen/prepackaged meal, bought from restaurant—dine in, bought from restaurant—takeout, delivery, snack/cereal, and no food, just beverage.	Categorical response	Self-report (Phone)	The food is?
Social Eating	Presence of other people accompany during the meal	Categorical response	Self-report (Phone)	Did you eat with anyone else?
Activity co-occurrence	Activities performed while eating: Nothing else, working /studying, watching TV, driving, socializing, and other	Categorical response	Self-report (Phone)	What were you doing while eating?
<i>Psychological</i>				
Emotion—down/lonely[128]	Degree of feeling “Down/Lonely”	5-pt Likert scale	Self-report (Phone, pre-meal)	How down or lonely do you feel right now?
Emotion—upbeat/excited[128]	Degree of feeling “Upbeat/Excited”	5-pt Likert scale	Self-report (Phone, pre-meal)	How upbeat or excited do you feel right now?
Emotion—calm/peaceful[128]	Degree of feeling “Calm/Peaceful”	5-pt Likert scale	Self-report (Phone, pre-meal)	How calm or peaceful do you feel right now?
Desire for food[128]	Desire for food before meal	5-pt Likert scale	Self-report (Phone, pre-meal)	How much desire do you have for food at the moment?
Pleasure-driven desire for food[129]	Desire for food motivated by taste/pleasure	5-pt Likert scale	Self-report (Phone, pre-meal)	How much is your desire for food motivated by taste or pleasure?
Biological hunger[129]	Desire for food motivated by hunger	5-pt Likert scale	Self-report (Phone, pre-meal)	How much is your desire for food motivated by hunger?

Feature	Description	Data	Source	Question (self-reported)
<i>Table 3.7.3 continued</i>				
Perceived stress[128]	Degree of feeling “Stressed/Anxious”	5-pt Likert scale	Self-report (Phone, pre- and post-meal)	To what extent do you feel stressed or anxious right now?
Cognitive restraint[130]	Stopped eating to manage calorie intake	4-pt Likert scale	Self-report (Phone, post-meal)	I consciously held back on how much I ate to keep from gaining weight.
Uncontrolled eating (LOC)[130]	Perceived inability to stop oneself from eating	4-pt Likert scale	Self-report (Phone, post-meal)	When I started eating, I just couldn’t seem to stop.
Perceived overeating[130]	Eating more than intended	5-pt Likert scale	Self-report (Phone, post-meal)	Did you eat more than you intended to?
Hedonic eating[128]	Pleasure obtained from meal	5-pt Likert scale	Self-report (Phone, post-meal)	How much pleasure did you get from the meal/snack?

Table 3.7.3: Feature descriptions (Cont.)

Feature table summarizing the unprocessed EMA responses and camera-derived initial features used in the study, categorized into behavioral, contextual, and psychological domains. Each feature is described along with its data type, source (e.g., visual or self-reported), and the corresponding self-report question where applicable. Behavioral features include objective measures such as bites, chews, and eating duration, while contextual and psychological features capture self-reported environmental and emotional factors associated with eating behavior.

Chapter 4

Digital Phenotyping

4.1 Technical Challenge: Digital Phenotyping Under Sparse and Heterogeneous Free-Living Behavioral Data

As of 2023, 40.3% of US adults have obesity[131], a disease caused by energy intake in excess of expenditure[132]. Individual factors shape the formation of overeating habits that sustain energy imbalance over time[133]. Consequently, understanding the patterns underlying individual dietary habits is crucial for designing effective long-term weight loss interventions [134]. Recent advances in sensor analytics enable automated capture of such patterns. This approach supports identifying both known (e.g., emotional eating, impulsive eating in response to cues, hedonic eating) and novel eating phenotypes, enabling personalized, meal-level interventions.

Significant efforts have been made to understand eating behaviors through objective and subjective monitoring methods such as passive sensing and Ecological Momentary Assessment (EMA) [90]. Previous passive sensing approaches include wrist-worn accelerometers to detect eating, inertial sensors to monitor chewing, and multi-sensor systems combining proximity, IMU, and ambient light sensors to track eating gestures

via jaw and hand movements [90]. Compared to traditional self-reports, EMA provides more accurate and nuanced data by capturing eating contexts in real time, which aids in identifying co-occurring psychological or situational exposures over time [4]. Combining image-based EMA with 24-hour dietary recalls can further enhance accuracy by reducing errors in dietary reporting [79]. The wealth of data these tools provide invites us to explore new analytic approaches to overeating.

Prior studies examine overeating through individual factors such as stress, night-time eating, and loss of control (LOC). Emotional states (e.g., anxiety, boredom, depression) are closely linked to emotional eating [135], while behaviors can increase intake (e.g., rapid eating) by overriding satiety cues [136]. Environmental influences, (e.g., access to high-calorie foods, social settings) also contribute to overeating [132]. Most existing methods examining these determinants have primarily relied on linear statistical approaches, which utilize a limited set of predictors to tackle the complex and multifaceted issue of overeating and thus fall short in capturing the interplay of physiological, psychological, behavioral, and environmental factors at work in overeating.

To address limitations in prior work, we introduce a multimodal framework that integrates passive sensing and EMA to explore overeating. Leveraging a dataset of 700 dietitian-verified meals from 41 adults with obesity, we extracted psychological, contextual, and camera-derived features across 14 days. Our contributions include: (1) collecting a high-resolution, multimodal dataset linking passive sensing and EMA in real-world eating contexts; (2) developing a supervised contrastive learning model to derive meaningful meal-level embeddings; and (3) identifying interpretable overeating phenotypes through clustering, revealing novel sensor-EMA behavioral signatures that may guide targeted intervention strategies for nutrition and obesity care.

4.2 Multimodal Meal-Level Data and Behavioral Feature Construction

4.2.1 Study Design and Data Collection

The 14-day observational study involved 65 adults with obesity ($\text{BMI} \geq 30 \text{ kg/m}^2$) in the Chicago area, designed to examine real-world eating behaviors using a combination of passive sensing and EMA. Participants were equipped with a wearable sensing suite that included a neck-mounted, activity-oriented wearable camera capable of capturing thermal and RGB video data to monitor eating activity in naturalistic settings. A custom smartphone app prompted participants to complete event-based EMA surveys before and after meals. Dietary intake was further assessed through 24-hour recalls conducted by trained dietitians using the multiple-pass method, supported by participant-captured food images. Of the 65 initial participants, 41 participants were included in the final analysis after excluding individuals who dropped out ($n=7$), lacked dietitian-administered recalls ($n=5$), had fewer than 10 valid meals ($n=5$), or had insufficient wearable camera data ($n=7$). The resulting dataset comprised 700 meals with complete passive sensing and EMA data. Figure 4.2.1 displays an overview of the analysis pipeline. More details about the study design and data collection can be found in the SenseWhy design paper [76].

4.2.2 Feature Extraction

EMA-derived features captured psychological and contextual factors surrounding each meal. Prior to and following each eating episode, participants completed surveys via a mobile app assessing hunger, affect, stress, and emotional states using Likert-scale items. Contextual features included meal source (e.g., self-prepared, restaurant), setting (e.g.,

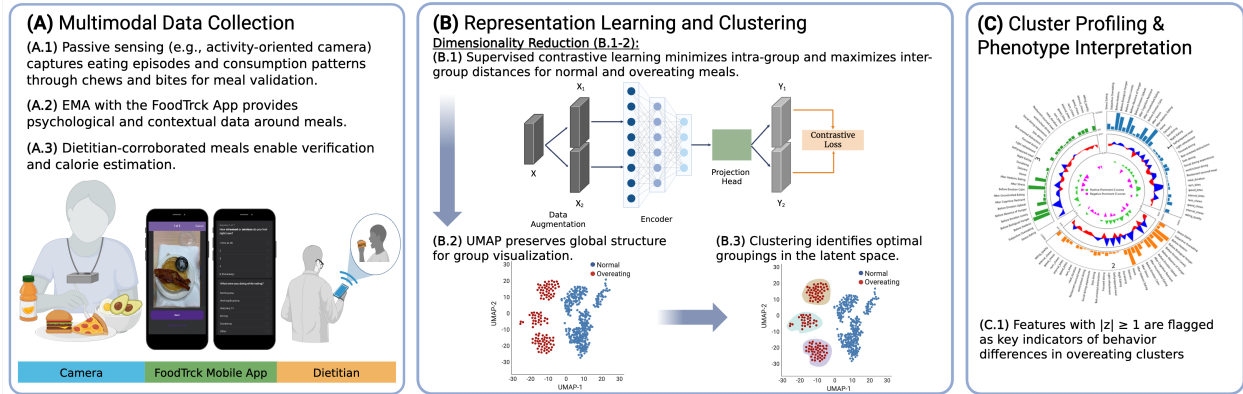


Figure 4.2.1: Multimodal pipeline for overeating phenotype discovery: (A) Passive sensing, EMA, and dietitian-verified meals enable rich data collection. (B) Supervised contrastive learning and UMAP uncover latent clusters. (C) Key features ($|z\text{-score}| \geq 1$) profile behavioral differences.

solo or social dining), and concurrent activities (e.g., task-oriented distraction or focused eating). Evening meals were defined as those consumed between 5 PM and 6 AM. Following a standardized protocol validated by trained raters, we conducted fine-grained annotations of eating activity using wearable camera footage. For each eating episode, we manually labeled the start and end times. Within each episode, we annotated individual bites, defined as large jaw open/close movements when food enters the mouth, and chews, defined as smaller, consecutive jaw movements during mastication. From these annotations, we derived passive sensing features including total number of chews/bites, chew-to-bite ratio, average time interval between chews and bites, meal duration from start/end of eating episodes, eating pace, and percent of meal time spent chewing. The outcome of interest, overeating, was defined as a binary variable indicating whether a meal's energy intake exceeded 1 z -score above the individual's 14-day average intake, based on validated data from 24-hour dietary recalls [76].

4.2.3 Eating Pace

In obesity research, eating speed influences satiety, intake, and weight outcomes, making it a key behavioral marker. Traditional assessments of eating speed, however, often rely on self-report or coarse measurements of meal duration, which may be prone to recall bias or distortion by long inactive periods during eating. To address these limitations, we propose a refined, behaviorally grounded method of defining eating based on chews, allowing for precise quantification of eating dynamics. We applied DBSCAN, a density-based clustering algorithm, to identify natural groupings of chews and remove temporally isolated outliers. DBSCAN parameters (ϵ and minimum samples) were optimized for each participant using elbow-method clustering to accommodate individual differences in chewing frequency. Chews occurring within ϵ seconds of one another were grouped into chewing sequences, while those separated by longer intervals were considered isolated and excluded. This segmentation filters out spurious chewing activity—such as a single chew minutes apart from others—that could otherwise inflate eating time or distort derived metrics. An eating episode was defined as chewing sequences containing at least 15 consecutive chews, in line with prior work, which established this threshold as a minimal criterion for sustained eating behavior[137]. Building on this segmentation, we introduced a novel metric for eating pace, defined as the number of chews per meal divided by the total chewing duration—calculated from the time between the first and last chew in the identified sequence. Unlike conventional definitions based on bite timing or meal start/end timestamps, our chew-based approach is robust to pauses in eating and better reflects actual masticatory engagement. To our knowledge, this is the first study to examine the relationship between overeating and detailed passive sensing features—such as eating pace, chew-to-bite ratio, percent of time spent chewing, and average chew/bite intervals—using validated, fine-grained annotations from wearable sensor data.

4.3 Representation Learning and Phenotype Discovery

4.3.1 Supervised Contrastive Learning Framework

We developed a supervised contrastive learning framework, SupConNet, inspired by the architecture and training strategy proposed in [18]. SupConNet consists of an MLP-based encoder followed by a projection head, trained using a supervised contrastive loss to learn discriminative embeddings of meal-level behavioral features. During training, two augmented views of each sample are generated through stochastic transformations, and the model is optimized to pull together embeddings from the same class (e.g., overeating vs. non-overeating) while pushing apart those from different classes. Formally, let $f_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^p$ denote the full network parameterized by θ , where $\mathbf{x} \in \mathbb{R}^d$ is a high-dimensional input feature vector and $\mathbf{z} = f_\theta(\mathbf{x}) \in \mathbb{R}^p$ is the learned embedding in a latent space. After training, we discard the projection head and extract embeddings from the encoder for downstream tasks. To visualize the learned representations, we apply Uniform Manifold Approximation and Projection (UMAP) to project $\mathbf{z} \in \mathbb{R}^p$ into a two-dimensional space $\mathbb{R}^2$, facilitating qualitative exploration of latent structure and clustering in eating behaviors.

4.3.2 Latent Space Clustering and Z-Score Analysis

We applied k -means clustering to the two-dimensional UMAP projections of the SupConNet embeddings to identify distinct subgroups of participants based on their learned behavioral representations. To determine the optimal number of clusters k , we computed the average silhouette score across a range of candidate values $k \in [2, 10]$, selecting values near local maxima for further evaluation.

We evaluated the resulting clustering using standard metrics including homogeneity, entropy, Davies-Bouldin Index (DBI), purity, and silhouette score. Homogeneity measures whether clusters contain only members of a single class (range: 0–1, higher is better); DBI evaluates cluster compactness and separation (range: 0 to ∞ , lower is better); entropy quantifies class distribution uncertainty within clusters (range: 0–1, lower is better); and the silhouette score captures cohesion vs. separation (range: -1 to 1, higher is better). Purity was computed as:

$$\text{Purity} = \frac{1}{N} \sum_k \max_j |C_k \cap L_j|,$$

where C_k is the set of samples in cluster k , L_j is the set of samples in class j , and N is the total number of samples.

To assess the distinctiveness of clusters, we conducted a z -score analysis across all features included in the study. For each cluster c and feature j , we calculated:

$$z_{c,j} = \frac{\mu_{c,j} - \mu_j}{\sigma_j},$$

where $\mu_{c,j}$ is the mean of feature j within cluster c , μ_j is the overall mean across all clusters, and σ_j is the corresponding standard deviation. This analysis enabled standardized comparisons across clusters to identify behavioral patterns and characterize each subgroup.

4.4 Results and Discussion

4.4.1 SupConNet Specifications

We trained SupConNet using a fully connected encoder with three hidden layers (128–128–64 units, ReLU activations) followed by a two-layer projection head with a final projection dimension of 32. Input features were augmented by adding Gaussian noise and applying random masking before contrastive training. The model was optimized using a supervised contrastive loss with a temperature parameter of 0.07. Training was conducted for 500 epochs using the Adam optimizer with a learning rate of 10^{-3} and a batch size of 4. After training, embeddings from the encoder were extracted and used for UMAP projection and then downstream clustering and phenotype profiling.

4.4.2 Overeating Subgroup Discovery via Clustering in Latent Space

After performing k -means clustering with $k = 8$, selected based on the maximum silhouette score, on the two-dimensional UMAP projections of SupConNet embeddings, we obtained clusters that exhibited strong separation and internal consistency. The resulting clustering achieved an average silhouette score of 0.55, homogeneity = 0.98, DBI = 0.51, average purity = 0.99, and average entropy = 0.05 across all clusters. Among these, three clusters were highly pure with respect to the overeating class and collectively accounted for all of the overeating meals. These clusters were selected for further profiling as candidate overeating phenotypes.

Table 4.4.1: Phenotype Profiles of Overeating Clusters

Phenotype Name	Psychological Features	Contextual Features	Camera Features
“Fast-Fork Social Dining”	Loss of control ^(p) Pre-meal biological hunger ^(p) Desire for food ^(p)	Social dining experience ^(p) Restaurant-sourced meal (dine-in) ^(p) In-restaurant dining ^(p) Socializing ^(p) In-home dining ⁽ⁿ⁾	Eating pace ^(p)
“Serene Screen Suppers”	Pre-meal biological hunger ^(p) Pre-meal calm ^(p) Hedonic eating ^(p) Desire for food ^(p) Perceived overeating ⁽ⁿ⁾ Pre-meal upbeat ⁽ⁿ⁾ Post-meal stress ⁽ⁿ⁾	Evening eating ^(p) Task-oriented distraction (watching TV) ^(p) Self-prepared meal ^(p) Solo dining ^(p)	Chews number ^(p) Bites number ^(p) Meal duration ^(p) Chew-bite ratio ⁽ⁿ⁾ Eating pace ⁽ⁿ⁾ Chew time-meal duration ratio ⁽ⁿ⁾
“Stressed Solo Bites”	Pre-meal stress ^(p) Post-meal stress ^(p) Pre-meal lonely ^(p) Perceived overeating ^(p) Cognitive restraint ^(p) Hedonic eating ⁽ⁿ⁾ Pre-meal upbeat ⁽ⁿ⁾ Pre-meal calm ⁽ⁿ⁾	Restaurant-sourced meal (takeout) ^(p) In-home dining ^(p)	–

The table shows the principal features and characteristics associated with each derived overeating phenotype cluster. *Note.* Superscripts ^(p) and ⁽ⁿ⁾ indicate positive and negative z-scores, respectively.

4.4.3 Phenotype Profiling

Table 4.4.1 summarizes the three overeating phenotypes identified via clustering, highlighting features with $|z\text{-score}| \geq 1$ as the most distinctive. For each cluster, we report cluster size (n), purity score (p), mean energy intake with standard deviation (kCal), entropy (e), and silhouette score (s). Additionally, we assign a concise name to each phenotype, highlight novel behavioral patterns, and contextualize them with existing literature.

“Fast-Fork Social Dining” ($n = 66$, $p = 100\%$, 1190.3 ± 355.2 kCal, $e = 0.00$, $s = 0.83$): reflects elevated loss of control and biological hunger, with a tendency to eat socially or at restaurants and at a faster pace. Our findings suggest that loss of control, as experienced in the context of social dining, may be coupled with an elevated eating pace—providing a behavioral lens through which sensor-derived features reflect internal states. This extends prior associations between eating speed and reduced satiety awareness[138].

“Serene Screen Suppers” ($n = 27$, $p = 92.6\%$, 1206.2 ± 536.2 kCal, $e = 0.381$, $s = 0.93$): are marked by biological hunger, low stress, and calm affect. Individuals in this group often eat self-prepared, evening meals while watching TV alone. Despite frequent chewing and biting, their low chew-to-bite ratio and slow eating pace suggest distracted, extended meals. Although prior work has linked screen-related distractions to overeating [139], we found that such distracted, mindless eating may also be reflected in altered chewing behavior (e.g., a lower chew-to-bite ratio and slower eating pace), highlighting new sensor-derived indicators of disengaged eating.

“Stressed Solo Bites” ($n = 18$, $p = 100\%$, 1143.4 ± 259.5 kCal, $e = 0.00$, $s = 0.94$): characterized by stress-related solitary takeout eating, elevated pre- and post-meal stress, loneliness, and perceived overeating, alongside low positive affect and reduced hedonic

motivation. This pattern reinforces and extends prior research linking emotional distress and social isolation to maladaptive eating behaviors and increased risk of overeating [135].

4.5 Conclusion and Future Work

These findings highlight distinct phenotypes of overeating meals, integrating psychological, contextual, and sensor-derived features. The outlined phenotypes provide a better understanding of prominent overeating behaviors at the meal level, which can then be utilized for tailored interventions deployed in real-time when passive sensing detects eating patterns consistent with specific phenotypes. Future work will focus on validating these phenotypes in larger, diverse populations and exploring their utility for targeted intervention strategies.

Chapter 5

Active Learning

5.1 Technical Challenge: Scalable Wearable Behavioral Sensing Under Limited Annotation Budget

Research Question 3

How can wearable-sensing models be developed under limited annotation budgets to support scalable behavioral AI?

With the advent of portable and wearable devices, such as mobile phones and smart-watches, large amounts of sensory data are created effortlessly by users [140], [141], [142]. As a result, human activity recognition (HAR) systems that detect lifestyle habits like eating[143], brushing[144], [145], and smoking[146], have been actively growing in recent years, and improving their ability to identify and recognize users' actions in a controlled setting [147]. These systems function through activity classifiers that rely on sufficient and representative amounts of accurately labeled data to generalize well in real-world settings.

Data labeling and annotation is a crucial bottleneck to machine learning systems

and can lead to costly, task-intensive, error-prone labor work (e.g., labeling Electronic Health Records [148]). In [149], [150], researchers propose a recall approach for the labeling process, which is fraught with burden and memory recollection error. More recently, researchers are using data from wearable video cameras to generate more reliable annotations [16], [151], [152]. However, a lot of time and cost in watching the video can be wasted by labeling redundant and irrelevant instances containing minimal information for a machine learning (ML) system.

Human-in-the-loop machine learning algorithms, such as "Active Learning" methods, help minimize the data annotation effort by adaptively selecting a subset of samples from the unlabeled set, based on a defined query budget in each querying iteration. These candidate samples are designed to be maximally informative to the classifier being built, which intuitively lead to improvements in classification decision boundaries if accounted for in the classification process [20]. Typically, the classifier is initialized by a "labeled batch," used for active learning model initialization to form the model classifier's initial decision boundaries.

In general, there are three types of selective sampling strategies within Active Learning: 1) membership query synthesis (MQS), 2) stream-based sampling, and 3) pool-based sampling. In MQS, the algorithm generates data instances from a certain underlying distribution and a labeler is then asked to annotate. In the stream-based sampling scenario, samples are generated sequentially in an online/real-time setting, and a decision needs to be made in real-time whether this sample should be labeled. In pool-based sampling, annotators have access to all unlabeled samples, and can assess each before deciding which sample to label [20], [153], [154], [155]. Within each of these methods, uncertainty sampling is one of the most widely adopted query strategies for selecting the most in-

formative sample to label [156]. Under this paradigm, an uncertainty metric needs to be defined to determine data samples or segments for which the model is least certain in its decision and prioritize them for annotation and training. In [157], Thomaz *et al.* build an active learning system that reduces the number of labels needed by applying pool-based sampling strategies and requiring annotation of only 8% of the training data to achieve similar performance. They selected cluster-based query strategies (which select the most informative samples within each cluster) to make sure that sample diversity is accounted for. Huang *et al.* [158] study the cluster structure of unlabeled data and how that effects performance, and shows that clustering methods impact active-learning classification outcomes. In [159], Aldana *et al.* followed an information theoretic approach to create clusters with highest entropy levels from the samples which help the ML models learn the underlying data patterns. In [160], Ashari *et al.* build a memory-aware active learning system, to reduce response time to the query, and use a maximum entropy uncertainty metric with data clustering to measure how certain the model is about its predicted label. Among all the different querying strategies that measure sample uncertainty, maximum entropy is one of the most commonly used approaches [20].

However, defining uncertainty metrics that are able to capture the samples that are necessary for building a generalizable model is still a major challenge within active learning. Most approaches calculate sample uncertainty based on the informativeness determined by model uncertainty, which is often captured by variability of the model's confidence in predicting an instance, without paying too much attention to the model's confidence value or intrinsic uncertainty, i.e., lower quality signals which are representative of difficult to see patterns in the signal itself and are often harder to learn. Simply taking model uncertainty into account is known to introduce bias in sample selection [161], which effects model generalizability. Introducing samples that are both model uncertain

and intrinsically uncertain is likely to enable faster model convergence (i.e., yield good results faster). This paper aims to assess the effect of intrinsic uncertainty on the speed of model convergence.

In [21], Swabha *et al.* introduced a new model-based tool known as Data Map, which assists in diagnosing easy-to-learn and hard-to-learn samples during the training phase based on training dynamics, allowing us to classify uncertainty as a function of correctness, confidence, and variability of a model. While Data Map allows us to visualize variability and confidence of model predictions, Data Map requires assessing a Deep Neural Network in a supervised-learning fashion, across multiple epochs, and needs ground truth labels (the very labels we aim to optimize in obtaining from the annotators). Needed is a method to retrieve intrinsic uncertainty in a way that is unsupervised, prior to obtaining the ground truth.

Inspired by [162], in which Pleiss *et al.* introduce the concept of Area-Under-Margin (AUM) statistics to identify mislabeled samples in an unsupervised machine-learning manner, we repurpose AUM to determine samples that the model is intrinsically uncertain about. Similar to Data Map, we are able to then assess easy-to-learn and hard-to-learn samples based solely on the sample itself (prior to requiring its ground truth label). Due to Data Map limitations, we adopt the power of the AUM statistic to determine samples for which the sample exhibits intrinsic uncertainty. In this paper, we propose and test a novel pool-based active learning framework that combines cluster-based maximum entropy (CME) with AUM (CME+AUM) to determine the effectiveness of the oracle selecting samples that are both model uncertain and intrinsically uncertain. We compare our approach to using the AUM-only method and CME-only. We evaluate this framework on the Necksense free-living dataset[80]. The Necksense free-living dataset is a multi-sensor

dataset obtained from 10 participants in a naturalistic setting, with the goal of detecting eating behavior, a known challenging mobile health area. The sensors used are from a neck-worn device that captures chewing sequences using proximity, ambient light, and inertial measurement unit (IMU) sensors. We make the key following contributions:

- We investigate the AUM statistic’s ability to measure hard-to-learn and easy-to-learn samples and its relationship with uncertainty visualization through Data Map.
- We propose and test a novel pool-based active learning model that attempts to capture intrinsic uncertainty (measured by AUM statistic) as well as model uncertainty (measured by cluster-based uncertainty sampling). In addition, we compare it against sampling strategies that use only model uncertainty or only intrinsic uncertainty. We evaluate our proposed model against the best evaluation model obtained from the Necksense free-living dataset.

5.2 Active Learning Framework for Informative Sample Selection

This section begins with a brief introduction about the Necksense dataset (section 5.2.1) followed by data uncertainty visualization using Data Map (section 5.2.2) with the training dynamics (i.e., confidence, variability, and correctness). Then in section 5.2.3, we measure the logit values (pre-softmax output) obtained from the final layer neurons of the MLP network and utilize them to compute AUM values for data instances in the training phase. In section 5.2.4, we explore how AUM statistic is linked to the Data Map representation of the samples through visual substantiation and show examples of a selected easy and hard chewing signal. In section 5.2.5, we introduce our novel framework combining

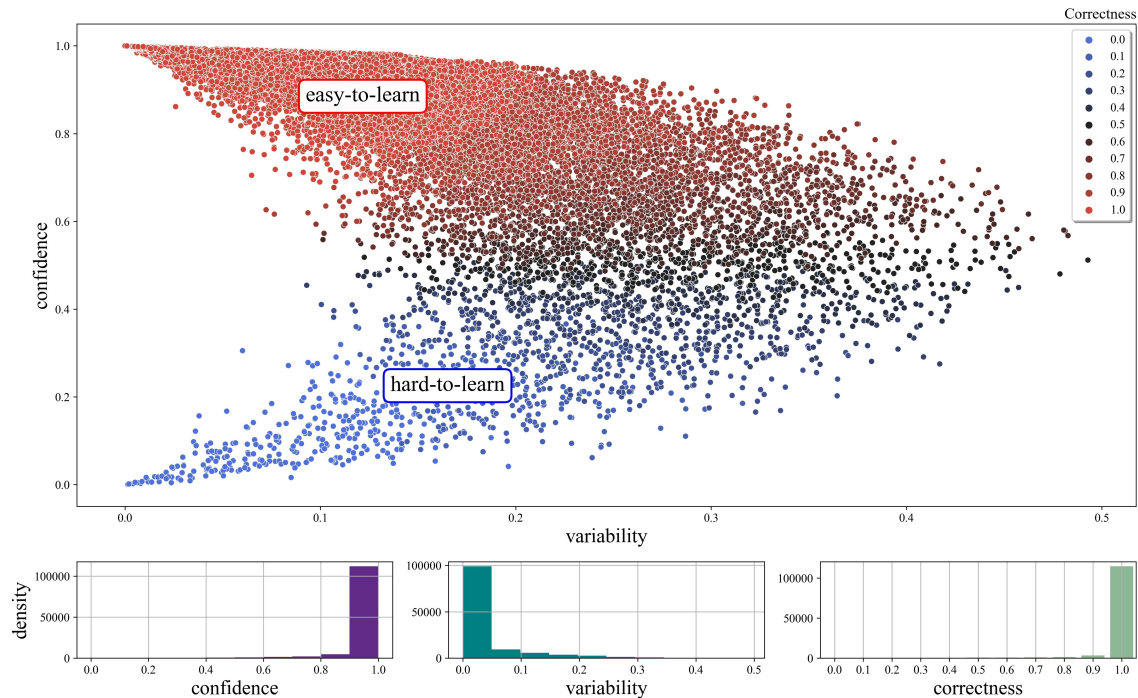


Figure 5.2.1: Uncertainty Visualization through Data Map for the Necksense Dataset. Here we see that the easy-to-learn segments are those with high confidence, low variability, and high correctness in model prediction. Hard-to-learn segments are typically those with lower confidence and correctness in model prediction. As shown in the distribution plots, the majority of the segments are considered easy-to-learn.

cluster-based maximum entropy and AUM.

5.2.1 Necksense Dataset

Necksense free-living dataset was obtained from 10 participants both with and without obesity with varying Body Mass Index (BMI) in a naturalistic setting. The data is collected by a multi-sensor necklace, worn around the neck throughout the day, that comprises an IMU sensor, ambient light, and proximity sensor to capture motion, leaning forward, and chewing actions occurring during an eating episode. The candidate segments from classification are obtained from the proximity sensor (which estimate the distance from the sensor to an object), and when oriented towards the jaw it captures periodicity in the

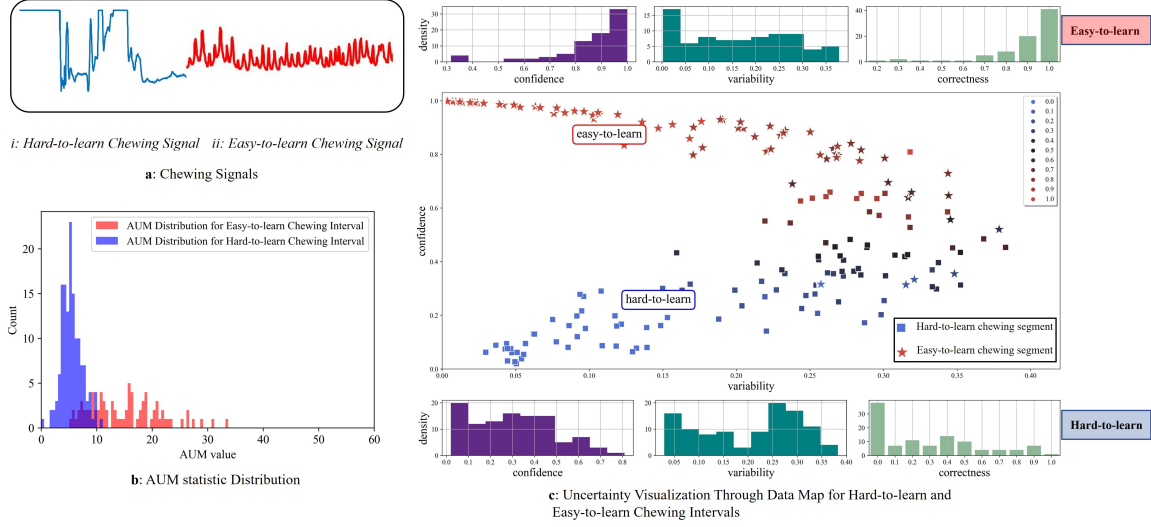


Figure 5.2.2: Uncertainty Visualization by Data Map and AUM statistic distribution for hard and easy chewing intervals. a) Here we see a snippet from long chewing signals, one that is hard-to-learn (blue) and one that's easy-to-learn (red); b) Here the algorithm extracted 144 candidate segments from the hard-to-learn and 80 segments from the easy-to-learn, and we see the distribution of the segments; c) Here we see the Data Map visual for the easy-to-learn signal (represented by stars), and hard-to-learn signal (represented by squares).

chewing signal. More than 137 hours of data was acquired from the participants, which provides information about chewing, feeding gesture and neck motion in order to detect eating episodes. In this study, to evaluate our proposed active learning method, we use the Necksense free-living study, comprising 123,155 candidate chewing segments (9.4% of the segments correspond to actual chews, and 90.6% correspond to non-chews).

5.2.2 Uncertainty Visualization Through Data Map

Data Map visualization generates three different statistics i.e., confidence, variability, and correctness obtained from the training phase of a Deep Neural Network across multiple epochs. Consider a training dataset D of size N where $D = \{(\mathbf{x}, \mathbf{y}^*)\}_{i=1}^N$, $\mathbf{x}$ is the candidate segment (or data instance), and $\mathbf{y}^*$ is the ground truth class label for the data instance. We assume the model predicts a probability distribution across all classes and data in-

stances over E epochs. Confidence is the first training dynamic which is defined as the average model's confidence of a data instance across different epochs which is estimated by:

$$\mu_i = \frac{1}{E} \sum_{e=1}^E p_{\theta}^e(y_i^* | x_i) \quad (5.1)$$

where p_{θ}^e is the model's probability distribution over the instance i at the end of epoch number e . Intuitively a higher confidence value for an instance means the sample, in terms of learnability, is easier to learn from the model's perspective. Variability is another statistic defining model's confidence variation in predicting one instance with respect to ground truth over epochs estimated by the following:

$$\sigma_i = \sqrt{\frac{\sum_{e=1}^E (p_{\theta}^e(y_i^* | x_i) - \mu_i)^2}{E}} \quad (5.2)$$

Correctness is another statistic measured by the number of times the model has predicted a sample correctly across different epochs. Figure 5.2.1 demonstrates how Necksense instances are scattered based on the three defined training statistics. It can be observed that the trainset is decomposed into easy-to-learn and hard-to-learn regions. A vast majority of the data instances are easy-to-learn with high confidence and low variability in the top-left of Figure 5.2.1. In the right side of Figure 5.2.1, there are instances with relatively high variability. The bottom-left of Figure 5.2.1 corresponds to the hard-to-learn region, which contains instances with low confidence and low variability. Easy-to-learn samples play an important role in model convergence and learning, while hard-to-learn instances are needed for the model's robustness and out-of-distribution generalizability[21]. It should be noted that uncertainty sampling strategies utilizing the metrics (e.g., entropy or margin in confidence) choose instances with highest variability in confidence from the data during the training phase since it implies model's indecisiveness about these samples.

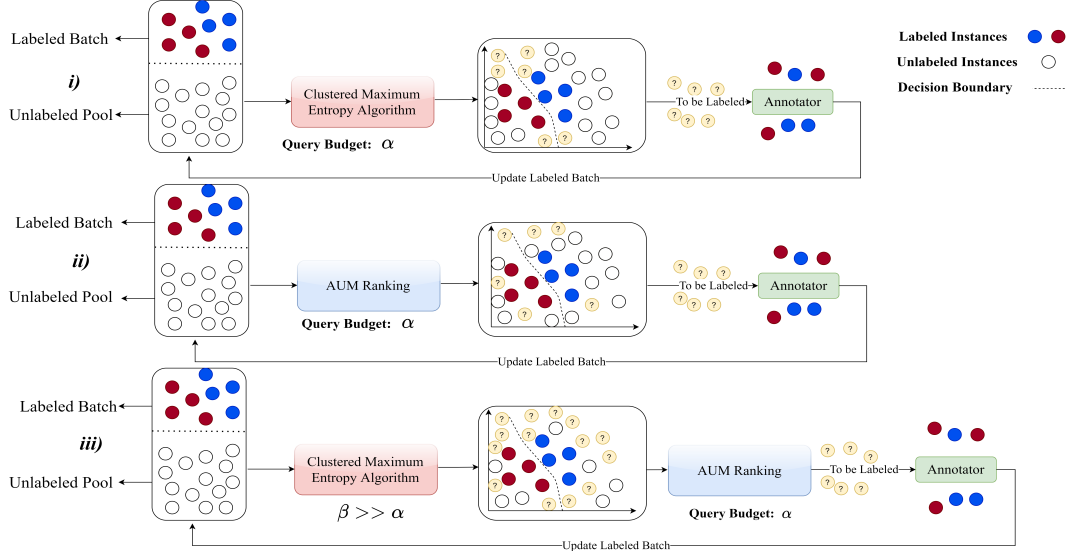


Figure 5.2.3: Part *i* is the traditional active learning pipeline with uncertainty sampling strategy (CME). Part *ii* is the AUM-only method which selects α instances with AUM values. Part *iii* shows the proposed active learning framework (CME+AUM)

5.2.3 Area-under-margin (AUM) for Necksense

To obtain AUM values, we used a four-layer feed forward deep neural network and trained it on the Necksense candidate segments. During the training phase, AUM statistic captures the mean differences between the logit values obtained from the neurons in the final layer of the deep learning model across different epochs. Originally used for identifying mislabeled instances, we repurpose this statistic and apply it to find the candidate segments that should be labeled based on intrinsic uncertainty. We show its link to Data Map to be able to generate a measure of difficulty and improve our understanding of the segments the oracle selects for labeling. AUM statistic is formulated as follows:

$$\text{AUM}(\mathbf{x}, \mathbf{y}) = \frac{1}{E} \sum_{e=1}^E M^{(e)}(\mathbf{x}, \mathbf{y}) \quad (5.3)$$

$$\mathbf{M}^{(e)}(\mathbf{x}, \mathbf{y}) = \mathbf{z}_i^{(e)}(\mathbf{x}) - \max_{i \neq j} \mathbf{z}_j^{(e)}(\mathbf{x}) \quad (5.4)$$

To elaborate more on the previous formulation, $\mathbf{M}^{(e)}(\mathbf{x}, \mathbf{y})$ is the margin at the end of epoch e which is equal to the difference between logit $\mathbf{z}_i$ of the ground truth subtracted by the largest other assigned logit $\max_{i \neq j} \mathbf{z}_j$. Logit value is defined as the pre-softmax output of the final layer of the DNN (Deep Neural Network). This representation allows us to have another measure to identify hard, ambiguous, and easy samples, which can be translated as another metric for training loss (e.g., easy samples have low training loss). The AUM statistic is modified to suit the unsupervised setting formulated by: $\text{AUM}(\mathbf{x}, \mathbf{y}) = \frac{1}{E} \sum_{e=1}^E |\mathbf{M}^{(e)}(\mathbf{x}, \mathbf{y})|$ where $\mathbf{M}^{(e)}(\mathbf{x}, \mathbf{y})$ is defined as in equation (4). Since the ground truth for logit value $\mathbf{z}_i$ is unknown prior to querying, we take the absolute value of the $\mathbf{M}^{(e)}(\mathbf{x}, \mathbf{y})$ in the binary classification setting and add them up across epochs.

5.2.4 Data Map and AUM statistic for Hard and Easy Chewing Intervals

This section explores the relationship between Data Map training dynamics and AUM statistic. Therefore, we can analyze further how visually high/low-quality chewing intervals are translated in terms of AUM statistic distribution for the candidate segments and compare it against uncertainty visualization using Data Map. When a model fails to predict one instance correctly, innately, this error originates from two sources. Model uncertainty is referred to as the model's limitation to classify and predict the instance correctly[163]. It is observed that model uncertainty has a strong correlation with variability statistic. However, intrinsic uncertainty is inherent ambiguity that the sample has on its own, which human annotators can notice. Intrinsic uncertainty has a strong relationship with model confidence, highlighting that when the model is highly confident

about an instance, there is a strong agreement between the annotator and the model. To the authors' best knowledge, no uncertainty metric in the literature accounts for both model uncertainty (model variability) and intrinsic uncertainty (model confidence) for instances. AUM statistic provides the desired filtering approach combined with uncertainty sampling using cluster-based maximum entropy to represent and account for the model and intrinsic uncertainty.

We can visually evaluate the quality of the chewing intervals by looking at their raw signals in Figure 5.2.2.a part *i* shows that the signal has reached proximity sensor saturation at some periods and does not follow a normal chewing pattern where peaks are noticeable and well-separated. Figure 5.2.2.a part *ii* indicates a visually higher quality chewing interval where peaks in the proximity signal are distinct and exhibit periodicity in jaw motion. Figure 5.2.2.b displays AUM values distribution for the segments of two hard-to-learn and easy-to-learn chewing intervals. AUM model could determine that the harder-to-learn chewing segments follow a distribution with a lower mean of 5.30 compared with the easy-to-learn interval with a mean of 25.23. Figure 5.2.2.c displays the uncertainty visualization for the segments of easy-to-learn and hard-to-learn chewing intervals. It can be inferred from Figure 5.2.2.b and Figure 5.2.2.c that lower values in AUM distribution are translated to lower model confidence and higher sample ambiguity which we visually confirmed earlier.

5.2.5 AUM-based Active Learning Framework

Figure 5.2.3 shows the high-level pipeline of the proposed active learning framework. The top diagram *i* shows how a traditional active learning method e.g., cluster-based maximum entropy would work. A cluster-based active learning algorithm would first perform efficient clustering on the unlabelled batch of data to account for sampling diversity

from the data clusters which arise from potentially different distributions. It should be noted that the data clusters might not be the same number of class labels that exist in the dataset. The active learning algorithm, in each querying iteration, picks a subset of samples, called the query budget α from the unlabelled batch of data as labelling candidates nominated to be labelled and identified with respect to the existing class labels by the annotator. The labeled batch then is updated by the newly labeled instances. Samples selected by active learning algorithms e.g., maximum entropy tend to account for model uncertainty due to significant changes in decision boundaries in each querying iteration. This is one of the primary reasons that model uncertainty is best translated into the Data Map variability statistic. The performance of the active learning method is evaluated on an unseen labelled test set in each querying iteration. The middle diagram *ii* depicts the AUM-only approach where the algorithm picks α least AUM values from the unlabeled pool as candidates for annotation. This approach only accounts for sample ambiguity in sample selection where the model's confidence for the selected samples is relatively lower. The bottom diagram *iii* displays the framework for the proposed active learning approach. Similar active learning approaches e.g., cluster-based maximum entropy is utilized to consider model uncertainty. However, to cast a wider confidence net, we increase the querying budget to $\beta \gg \alpha$. Since Data Map is unable to be employed in an unsupervised setting, we harness AUM statistic, which is originally proposed to identify mislabeled data, and modified it to suit the unsupervised setting. Therefore, an AUM filter is added to select the α samples with least AUM values (considered harder-to-learn with lower confidence).

5.3 Evaluation and Results

5.3.1 Evaluation

We evaluate our active learning framework on the Necksense free-living study. We randomly select 85% of the data as the unlabeled batch and 15% of the data as the unseen labeled set in a participant independent fashion. We assess the performance of the proposed model against random sampling, AUM-only, and cluster-based maximum entropy methods. We empirically select a labeled subset of 0.1% of the unlabeled instances for model initialization and a query budget of six for all methods. For cluster-based maximum entropy, we cluster the unlabeled batch using a k-means clustering approach. Based on the Silhouette metric, which captures cluster separation quality, we set $k=6$ clusters. During each querying iteration, one sample with highest entropy is selected from each data cluster resulting in six selected samples for annotation. In our proposed model, we set $\beta = 120$ which selects 20 samples with highest entropy from each cluster. The selected samples are filtered and ranked with respect to their AUM values. At the end, we select 6 samples with the least AUM value. In order to get AUM values, we train a 4 layer feed-forward multi-layer perceptron which is trained on the labeled batch and evaluated on 120 selected samples chosen by maximum entropy method for 100 epochs in each iteration. The pre-softmax values are retrieved at the end of each epoch from the final layer. For the AUM-only approach, we train the deep feed-forward network on the labeled batch and retrieve and rank the AUM values for the unlabeled pool in each querying iteration. The random sampling method, also named passive learning, randomly selects unlabeled instances in each querying iteration and updates the labeled batch. The best evaluation shows the results which utilize the whole training set. We use the XGBoost [164], the optimal classifier used in the Necksense dataset, and is available as an open-source software

package for the classification assessment. The results are aggregated and shown at per-second level, where overlapping candidate segments are combined to determine whether each second is a chew or not.

5.3.2 Results

Figure 5.3.1 shows the results for CME-only, AUM-only, CME+AUM, random sampling, and compares them with the best evaluation used in Necksense. We can see that our proposed method outperforms other approaches and reaches the best evaluation performance after 12 query iterations which is equivalent to 0.18% of the unlabeled data. However, cluster-based maximum entropy (CME) reaches the performance of best evaluation after 30 iterations equivalent to 0.29%, performing better than random sampling and AUM-only methods after 10 iterations. The AUM-only approach shows weak performance when it comes to model convergence due to selecting and training primarily on hard-to-learn samples. Also, since the sample selection process is not diverse in the AUM-only method, the algorithm might become biased and keep sampling from limited regions of the data subspaces. It is noted that at some points, active learning approaches outperform the best evaluation. Therefore, it highlights that the right composition of model uncertainty and intrinsic uncertainty needs to be considered in order to ensure proper representation of the underlying data distribution to assist with model generalizability. This signifies the fact that adding trivial or redundant samples might actually lower the model's performance. Further analysis and quantification of this is needed on other datasets to confirm.

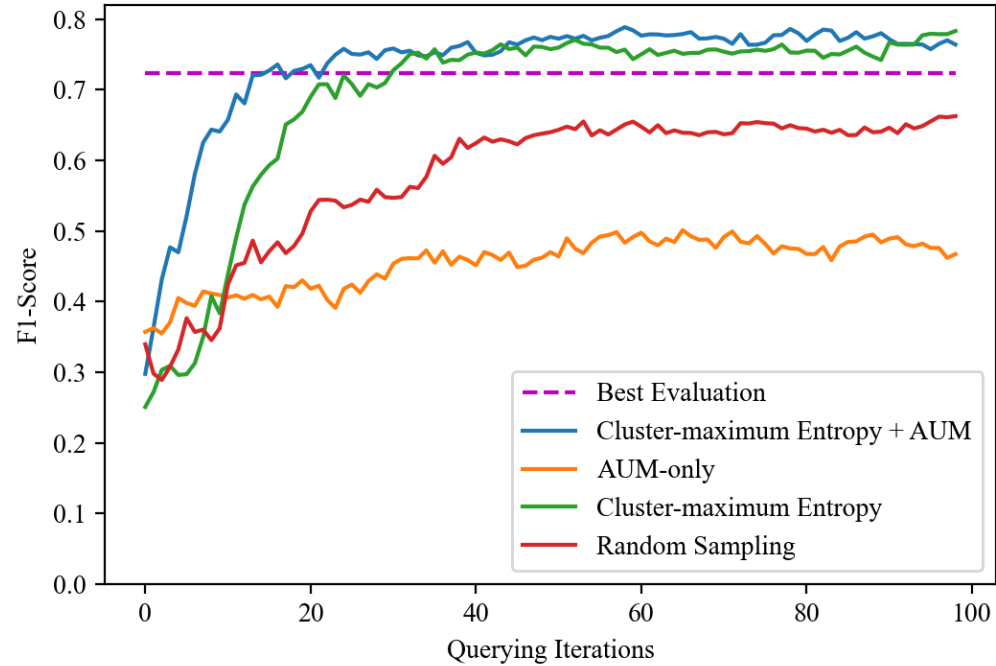


Figure 5.3.1: ActiveSense Result on Necksense dataset

5.4 Conclusion

In this paper, we propose a new active learning framework which harnesses intrinsic uncertainty as well as model uncertainty and tested it on Necksense free-living dataset. We showed the proposed method reaches best evaluation performance by using 0.18% of the data compared to 85% of the data in the best evaluation method. We also show how the AUM statistic, originally designed to detect mislabeled data, can be repurposed to capture segments that are easy and hard to learn without the need for ground truth labels. We visualized the uncertainty of Necksense data instances utilizing training dynamics (e.g., confidence, variability, and correctness) through Data Map. Future research should validate this method on other mobile health datasets and further assess the benefits of our proposed framework in reducing annotation time and cost.

Chapter 6

Discussion

6.1 Overview

This dissertation developed a computational framework for using continuous monitoring, multimodal sensing, digital phenotyping, and label-efficient learning to support precision behavioral health. Across the preceding chapters, the central argument is that health-risk behaviors cannot be fully understood through isolated clinical measurements or retrospective self-report alone. They unfold dynamically across time, within everyday contexts, and through interactions among psychological state, behavioral patterning, environmental exposure, and treatment engagement. Therefore, the next generation of behavioral health systems must move beyond static assessment toward longitudinal, multimodal, interpretable, and scalable AI-enabled methods.

The work in this dissertation follows a sensing-to-insight progression. *Chapter 2* examined whether early longitudinal self-monitoring data can identify individuals at risk for suboptimal weight loss. *Chapter 3* then moved from distal treatment outcomes to proximal behavior by using multimodal sensing and ecological momentary assessment to detect overeating episodes in free-living conditions. *Chapter 4* extended detection into

digital phenotyping by identifying interpretable meal-level overeating patterns. Finally, *Chapter 5* addressed the annotation bottleneck that limits the scalability of wearable sensing systems. Together, these chapters show how behavioral data can be transformed from raw longitudinal signals into actionable risk estimates, interpretable behavioral phenotypes, and more scalable model-development pipelines.

6.2 From Episodic Assessment to Continuous Behavioral Risk Modeling

A primary contribution of this dissertation is the demonstration that early, continuously collected behavioral signals can support risk identification before conventional clinical decision points. In *Chapter 2*, early weight trajectories collected during the first two weeks of behavioral weight management were used to predict longer-term treatment response. Rather than relying only on a single early weight-loss threshold, the model incorporated trajectory-based features such as weight change, slope, distributional statistics, variability, and missingness-adjusted patterns of self-monitoring.

This finding has two broader implications. First, early treatment response is not only an outcome marker but also a dynamic behavioral signal. The pattern of self-monitoring and weight change during the first weeks of treatment can encode information about adherence, physiological response, behavioral consistency, and early intervention engagement. Second, explainable machine learning can make these signals more useful for clinical decision-making by identifying which features contribute to predictions and how they influence model output. In this way, continuous monitoring can support adaptive treatment models in which care intensity is adjusted based on emerging risk rather than delayed assessment.

However, early risk prediction alone does not explain the behavioral mechanisms that produce risk. A model may identify that an individual is unlikely to achieve sufficient weight loss, but intervention still requires understanding which proximal behaviors and contexts should be targeted. This limitation motivates the transition from *Chapter 2* to *Chapter 3*: from predicting risk to measuring the daily behaviors that may drive that risk.

6.3 From Risk Prediction to Proximal Overeating Detection

The SenseWhy study addressed the need to characterize proximal eating behaviors in naturalistic settings. This chapter integrated wearable camera sensing, ecological momentary assessment, mobile dietary reporting, and dietitian-administered recalls to detect and interpret overeating at the meal level. By defining overeating relative to each participant's usual intake, this work avoided treating overeating as a fixed universal calorie threshold and instead aligned the outcome with individual behavioral baselines.

The multimodal design of SenseWhy is important because overeating is not captured fully by any single data source. EMA provides access to psychological and contextual factors, including hunger, stress, loss of control, perceived overeating, food source, and concurrent activity. Passive sensing provides objective markers of eating microstructure, including chews, bites, chewing intervals, bite rate, eating duration, and related behavioral dynamics. The results show that these data streams are complementary: subjective context and objective eating behavior together provide a richer representation of overeating than either modality alone.

This chapter also contributes methodologically by linking interpretable machine learning with behavioral theory. Features such as perceived overeating, loss of control, biological hunger, light refreshments, number of chews, chew-bite ratio, and chewing

interval are not merely predictive variables; they are behavioral constructs that can be connected to intervention targets. For example, a high number of bites or rapid eating pace suggests a different intervention pathway than stress-related or loss-of-control eating. Thus, multimodal overeating detection provides a bridge between sensor data and actionable behavioral mechanisms.

6.4 From Detection to Interpretable Digital Phenotypes

While *Chapter 3* focused on detecting overeating episodes, *Chapter 4* addressed a deeper question: whether overeating represents a single behavioral category or a set of heterogeneous patterns. The findings support the latter interpretation. Overeating can occur through multiple pathways, including evening eating, restaurant-based eating, stress-related eating, social dining, distracted eating, hedonic eating, and altered eating pace. Treating all overeating episodes as equivalent risks obscuring this heterogeneity.

The digital phenotyping framework used multimodal meal-level data to learn latent representations and identify interpretable clusters of overeating behavior. This approach moves beyond binary classification by characterizing the structure within overeating episodes. Instead of asking only whether overeating occurred, the phenotype framework asks what type of overeating occurred and which psychological, contextual, and sensor-derived features define that pattern.

This shift is central to precision behavioral health. A binary overeating detector may support event recognition, but a phenotype model can inform intervention selection. Stress-related overeating may require emotion regulation or coping strategies. Social or restaurant-based overeating may require planning, portion control, or environmental restructuring. Distracted eating may call for attention-based or mindful eating strategies.

Eating patterns defined by rapid pace or low chew-bite ratio may support real-time feedback to slow eating. Thus, digital phenotyping transforms sensing data into interpretable behavioral profiles that can guide more targeted interventions.

6.5 Scalability and the Annotation Bottleneck

The final technical contribution of this dissertation, presented in *Chapter 5*, addressed a practical barrier that cuts across all sensing-based behavioral health systems: the cost of labeled data. Wearable sensing systems can generate large volumes of behavioral data, but supervised learning requires annotations that are often expensive, slow, and difficult to obtain. This problem is especially acute for eating behavior research, where labels may require human review of long sensor streams to identify meal boundaries, chewing segments, bites, or other fine-grained behaviors.

The active learning framework developed in this dissertation reduces this burden by identifying informative samples for annotation. Rather than labeling data randomly or exhaustively, the method combines model uncertainty with intrinsic uncertainty derived from training dynamics. This allows the system to prioritize samples that are likely to improve model performance or clarify difficult decision regions. The framework therefore supports a more efficient human-in-the-loop pipeline for wearable sensing research.

This contribution is not separate from the rest of the dissertation; it is what makes the broader sensing-to-insight pipeline more scalable. Early risk prediction, overeating detection, and digital phenotyping all depend on reliable models trained on sufficiently representative data. Without efficient annotation strategies, these systems remain difficult to expand across larger populations, longer deployments, new sensors, and more diverse behavioral contexts. Active learning provides a mechanism for scaling behavioral

AI while reducing manual labeling effort.

6.6 Scientific Implications

Collectively, this dissertation advances three scientific claims. First, behavioral risk is dynamic and can be modeled earlier than traditional clinical endpoints. The early risk prediction work shows that short-term longitudinal trajectories can provide meaningful signals about longer-term outcomes. This supports a shift from delayed treatment evaluation to earlier, adaptive intervention logic.

Second, proximal health-risk behaviors are multimodal. Overeating cannot be fully explained by self-report, sensor data, or dietary recall alone. Its detection and interpretation require combining subjective experience, objective behavioral microstructure, and contextual information. This supports the broader view that behavioral health AI should integrate multiple data streams rather than optimize isolated modalities.

Third, behavioral phenotypes are more informative than behavior labels alone. A single label such as “overeating” hides the different pathways that produce the behavior. By identifying interpretable phenotypes, this dissertation shows how machine learning can move from prediction toward behavioral explanation. This is necessary for building interventions that are not only timely but also mechanistically targeted.

6.7 Practical Implications for Precision Behavioral Health

The findings have practical implications for the design of future behavioral health systems. First, early monitoring systems can be used to identify individuals who may require additional support before treatment failure becomes apparent. This is particularly

relevant for stepped-care models, where intervention intensity can be adapted based on emerging response patterns.

Second, wearable and mobile sensing systems should be designed to capture both behavioral events and the context surrounding those events. For overeating, this means measuring not only whether eating occurred but also how the person was eating, what psychological state preceded the meal, what context surrounded the meal, and whether the meal exceeded the individual's usual intake.

Third, digital phenotypes can support more personalized intervention logic. Instead of deploying the same feedback to all individuals who overeat, future systems could tailor support based on phenotype. For example, a stress-driven phenotype may trigger coping support, an evening-craving phenotype may trigger planning or meal-timing strategies, and a rapid-eating phenotype may trigger pacing feedback.

Fourth, scalable deployment requires human-in-the-loop learning systems that minimize annotation burden. Active learning can help researchers and developers focus labeling resources on the most informative segments, reducing cost while improving model development. This is essential for moving wearable sensing systems from small research studies toward larger and more diverse real-world deployments.

6.8 Limitations

Several limitations should be considered when interpreting the findings. First, the studies were conducted within specific behavioral health contexts, primarily weight management and overeating. Although the computational framework is relevant to other health-risk behaviors, additional validation is needed before extending these methods to other

domains such as smoking, physical inactivity, medication adherence, or sleep-related behaviors.

Second, the datasets were limited in size and duration relative to the complexity of real-world behavior. The SenseWhy analyses were based on a two-week observational period, and the digital phenotyping models were developed from participants with obesity in a specific geographic and study context. Larger and more diverse datasets are needed to determine whether the identified overeating phenotypes generalize across populations, cultures, treatment settings, and longer observation periods.

Third, although passive sensing provides objective behavioral information, it introduces practical challenges related to wearability, privacy, missingness, and user burden. Wearable camera systems are especially useful for validating eating behavior and extracting microstructure, but they may not be acceptable or feasible for all populations or long-term deployments. Future systems may need to combine camera-based validation with lower-burden sensors that can operate continuously in everyday life.

Fourth, the models developed in this dissertation are predictive and descriptive rather than causal. The identified features and phenotypes reveal associations with overeating and weight-loss outcomes, but they do not establish that changing a specific feature will necessarily change the outcome. Prospective and experimental studies are needed to test whether phenotype-specific interventions improve behavior change and treatment response.

Finally, active learning reduces annotation burden but does not eliminate the need for high-quality labels. Model-selected samples may still be ambiguous, difficult to annotate, or sensitive to labeler disagreement. Future work should pair active learning with

rigorous annotation protocols, uncertainty-aware label aggregation, and quality-control processes.

6.9 Future Directions

Future work should extend this dissertation in four directions. First, early risk prediction models should be evaluated prospectively in adaptive intervention studies. A key next step is to test whether predictions made in the first weeks of treatment can trigger treatment adaptations that improve long-term outcomes compared with standard care or fixed decision rules.

Second, the SenseWhy sensing framework should be validated in larger and more diverse populations. This includes examining whether personalized overeating thresholds remain stable over longer periods, whether passive sensing features generalize across device configurations, and whether the combination of EMA and passive sensing can be reduced to a lower-burden measurement protocol without sacrificing interpretability.

Third, digital phenotypes should be connected directly to intervention strategies. The phenotypes identified in this dissertation suggest different behavioral mechanisms, but their clinical utility depends on whether phenotype-specific feedback or treatment adaptation improves outcomes. Future studies should test whether phenotype-guided interventions are more effective than generic dietary or behavioral feedback.

Fourth, active learning should be integrated into longitudinal sensing deployments. Instead of treating annotation as a one-time offline process, future systems could use active learning continuously to identify uncertain or underrepresented behavioral segments during deployment. This would allow models to adapt to new participants, new

contexts, and shifting behavioral patterns over time.

Chapter 7

Conclusion

Health-risk behaviors are central to the prevention and management of chronic disease, yet they remain difficult to measure, interpret, and intervene on because they unfold continuously in daily life. Traditional clinical assessments and retrospective self-report provide valuable information, but they often miss the temporal, contextual, and behavioral details needed to understand why risk emerges and how intervention should be adapted. This dissertation addressed this gap by developing computational approaches that use continuous monitoring, multimodal sensing, digital phenotyping, and label-efficient learning to support precision behavioral health.

The work presented in this dissertation demonstrates that behavioral risk can be identified earlier and with greater nuance when longitudinal data are modeled as dynamic trajectories rather than isolated measurements. Early patterns of self-monitoring and weight change contain meaningful information about later treatment outcomes, supporting the use of adaptive models that can guide intervention before poor response becomes entrenched. This shifts behavioral assessment from delayed evaluation toward earlier and more proactive risk stratification.

This dissertation also shows that proximal health-risk behaviors, such as overeat-

ing, require multimodal measurement. Overeating is not only a matter of caloric intake; it is shaped by hunger, affect, stress, loss of control, food context, social setting, eating pace, and objective eating microstructure. By integrating passive sensing with self-reported psychological and contextual data, this work demonstrates how AI-enabled systems can capture both the observable and subjective dimensions of behavior in everyday life.

A central finding across this dissertation is that behaviorally meaningful outcomes are often heterogeneous. A single label such as overeating can represent multiple underlying pathways, including stress-driven eating, socially situated eating, restaurant-based eating, distracted eating, and changes in eating pace. Digital phenotyping provides a way to move beyond binary detection toward interpretable behavioral subtypes that can better inform intervention design. This approach supports a more precise form of behavioral health, where support can be tailored not only to whether risk is present, but to the mechanisms through which that risk emerges.

The dissertation further addresses the practical challenge of scaling sensing-based behavioral AI. Large volumes of wearable and mobile sensing data are difficult to annotate exhaustively, and manual labeling remains a major bottleneck for model development. By incorporating active learning and training-dynamics-based uncertainty, this work advances a more efficient human-in-the-loop approach for identifying informative data to label. This contribution is essential for moving sensing systems beyond small, curated datasets toward larger, more diverse, and more deployable behavioral health applications.

Collectively, this dissertation contributes to a broader vision of precision behavioral health in which continuous data streams are transformed into actionable insight. The proposed framework connects early risk identification, objective behavior detection, in-

interpretable phenotype discovery, and scalable model development. Rather than treating sensing as a tool for passive data collection alone, this work positions sensing and AI as part of an adaptive system for understanding behavior as it naturally occurs and for informing timely, personalized intervention.

Future behavioral health systems will need to be accurate, interpretable, ethical, and scalable. They must respect the complexity of human behavior while remaining practical enough for real-world use. The methods developed in this dissertation provide a foundation for such systems by showing how multimodal data and machine learning can be used to detect risk, explain behavior, characterize heterogeneity, and reduce annotation burden. In doing so, this dissertation moves behavioral health research closer to systems that can support individuals in the moments and contexts where behavior change is most needed.

References

- [1] A. H. Mokdad, J. S. Marks, D. F. Stroup, and J. L. Gerberding, “Actual causes of death in the united states, 2000,” *Jama*, vol. 291, no. 10, pp. 1238–1245, 2004.
- [2] E. Kvaavik, G. D. Batty, G. Ursin, R. Huxley, and C. R. Gale, “Influence of individual and combined health behaviors on total and cause-specific mortality in men and women: The united kingdom health and lifestyle survey,” *Archives of internal medicine*, vol. 170, no. 8, pp. 711–718, 2010.
- [3] B. Spring, A. C. Moller, and M. J. Coons, “Multiple health behaviours: Overview and implications,” *Journal of public health*, vol. 34, no. suppl_1, pp. i3–i10, 2012.
- [4] S. Shiffman, A. A. Stone, and M. R. Hufford, “Ecological momentary assessment,” *Annu. Rev. Clin. Psychol.*, vol. 4, no. 1, pp. 1–32, 2008.
- [5] J.-P. Onnela and S. L. Rauch, “Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health,” *Neuropsychopharmacology*, vol. 41, no. 7, pp. 1691–1696, 2016.
- [6] D. C. Mohr, M. Zhang, and S. M. Schueller, “Personal sensing: Understanding mental health using ubiquitous sensors and machine learning,” *Annual review of clinical psychology*, vol. 13, pp. 23–47, 2017.
- [7] A. R. Doherty et al., “Wearable cameras in health: The state of the art and future possibilities,” *American journal of preventive medicine*, vol. 44, no. 3, pp. 320–323, 2013.
- [8] P. Kelly et al., “An ethical framework for automated, wearable cameras in health behavior research,” *American journal of preventive medicine*, vol. 44, no. 3, pp. 314–319, 2013.

- [9] L. Gemming, J. Utter, and C. N. Mhurchu, "Image-assisted dietary assessment: A systematic review of the evidence," *Journal of the Academy of Nutrition and Dietetics*, vol. 115, no. 1, pp. 64–77, 2015.
- [10] J. M. Jakicic et al., "Effect of a stepped-care intervention approach on weight loss in adults: A randomized clinical trial," *Jama*, vol. 307, no. 24, pp. 2617–2626, 2012.
- [11] M. E. Waring et al., "Early-treatment weight loss predicts 6-month weight loss in women with obesity and depression: Implications for stepped care," *Journal of psychosomatic research*, vol. 76, no. 5, pp. 394–399, 2014.
- [12] M. L. Patel, C. M. Hopkins, and G. Bennett, "Early weight loss in a standalone mhealth intervention predicting treatment success," *Obesity science & practice*, vol. 5, no. 3, pp. 231–237, 2019.
- [13] A. B. Goldschmidt et al., "Ecological momentary assessment of eating episodes in obese adults," *Biopsychosocial Science and Medicine*, vol. 76, no. 9, pp. 747–752, 2014.
- [14] J. Reichenberger, R. Schnepfer, A.-K. Arend, and J. Blechert, "Emotional eating in healthy individuals and patients with an eating disorder: Evidence from psychometric, experimental and naturalistic studies," *Proceedings of the Nutrition Society*, vol. 79, no. 3, pp. 290–299, 2020.
- [15] M. R. Lowe and M. L. Butryn, "Hedonic hunger: A new dimension of appetite?" *Physiology & behavior*, vol. 91, no. 4, pp. 432–439, 2007.
- [16] E. Thomaz, I. Essa, and G. D. Abowd, "A practical approach for recognizing eating moments with wrist-mounted inertial sensing," in *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, 2015, pp. 1029–1040.
- [17] K. Kyritsis, C. Diou, and A. Delopoulos, "Modeling wrist micromovements to measure in-meal eating behavior from inertial sensor data," *IEEE journal of biomedical and health informatics*, vol. 23, no. 6, pp. 2325–2334, 2019.
- [18] P. Khosla et al., "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.
- [19] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.
- [20] B. Settles, "Active learning literature survey," 2009.

- [21] S. Swayamdipta et al., “Dataset cartography: Mapping and diagnosing datasets with training dynamics,” *arXiv preprint arXiv:2009.10795*, 2020.
- [22] World Health Organization, *Obesity and overweight*, <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight>, Accessed: 2026-01-25, 2021.
- [23] S. J. Curry et al., “Behavioral weight loss interventions to prevent obesity-related morbidity and mortality in adults: Us preventive services task force recommendation statement,” *Jama*, vol. 320, no. 11, pp. 1163–1171, 2018.
- [24] B. Spring, “Sound health care economics: Provide the treatment needed (not less, not more).,” *Health Psychology*, vol. 38, no. 8, p. 701, 2019.
- [25] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, vol. 30, 2017.
- [26] K. Fujihara et al., “Machine learning approach to predict body weight in adults,” *Frontiers in public health*, vol. 11, p. 1090146, 2023.
- [27] O. Babajide et al., “A machine learning approach to short-term body weight prediction in a dietary intervention program,” in *International Conference on Computational Science*, Springer, 2020, pp. 441–455.
- [28] A. F. Pfammatter et al., “Smart: Study protocol for a sequential multiple assignment randomized controlled trial to optimize weight loss management,” *Contemporary clinical trials*, vol. 82, pp. 36–45, 2019.
- [29] 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,” *Contemporary clinical trials*, vol. 38, no. 2, pp. 251–259, 2014.
- [30] B. Spring et al., “Effects of an abbreviated obesity intervention supported by mobile technology: The engaged randomized clinical trial,” *Obesity*, vol. 25, no. 7, pp. 1191–1198, 2017.
- [31] B. Spring et al., “An adaptive behavioral intervention for weight loss management: A randomized clinical trial,” *Jama*, vol. 332, no. 1, pp. 21–30, 2024.
- [32] S. Chen, W. A. Banks, M. Sheffrin, W. Bryson, M. Black, and S. M. Thielke, “Identifying and categorizing spurious weight data in electronic medical records,” *The American journal of clinical nutrition*, vol. 107, no. 3, pp. 420–426, 2018.

- [33] H.-Y. Kim, “Statistical notes for clinical researchers: Simple linear regression 3-residual analysis,” *Restorative dentistry & endodontics*, vol. 44, no. 1, e11, 2019.
- [34] G. E. Nam et al., “Impact of body mass index and body weight variabilities on mortality: A nationwide cohort study,” *International journal of obesity*, vol. 43, no. 2, pp. 412–423, 2019.
- [35] S. Bangalore, R. Fayyad, R. Laskey, D. A. DeMicco, F. H. Messerli, and D. D. Waters, “Body-weight fluctuations and outcomes in coronary disease,” *New England Journal of Medicine*, vol. 376, no. 14, pp. 1332–1340, 2017.
- [36] J. Turicchi et al., “Body weight variability is not associated with changes in risk factors for cardiometabolic disease,” *International Journal of Cardiology Hypertension*, vol. 6, p. 100 045, 2020.
- [37] J. Cologne et al., “Association of weight fluctuation with mortality in japanese adults,” *JAMA network open*, vol. 2, no. 3, e190731–e190731, 2019.
- [38] L. Benson, F. Zhang, H. Espel-Huynh, L. Wilkinson, and M. R. Lowe, “Weight variability during self-monitored weight loss predicts future weight loss outcome,” *International Journal of Obesity*, vol. 44, no. 6, pp. 1360–1367, 2020.
- [39] J. Snoek, H. Larochelle, and R. P. Adams, “Practical bayesian optimization of machine learning algorithms,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25, Curran Associates, Inc., 2012.
- [40] I. Kolyshkina and S. Simoff, “Interpretability of machine learning solutions in public healthcare: The crisp-ml approach,” *Frontiers in big data*, vol. 4, p. 660 206, 2021.
- [41] P. A. Moreno-Sánchez, “Improvement of a prediction model for heart failure survival through explainable artificial intelligence,” *Frontiers in cardiovascular medicine*, vol. 10, p. 1 219 586, 2023.
- [42] R. Elshaw, M. H. Al-Mallah, and S. Sakr, “On the interpretability of machine learning-based model for predicting hypertension,” *BMC medical informatics and decision making*, vol. 19, no. 1, p. 146, 2019.
- [43] L. M. Nackers, K. M. Ross, and M. G. Perri, “The association between rate of initial weight loss and long-term success in obesity treatment: Does slow and steady win the race?” *International journal of behavioral medicine*, vol. 17, no. 3, pp. 161–167, 2010.

- [44] C. Iwendi et al., "Covid-19 patient health prediction using boosted random forest algorithm," *Frontiers in public health*, vol. 8, p. 357, 2020.
- [45] L. Zhu, D. Qiu, D. Ergu, C. Ying, and K. Liu, "A study on predicting loan default based on the random forest algorithm," *Procedia computer science*, vol. 162, pp. 503–513, 2019.
- [46] M. Elliott, F. Gillison, and J. Barnett, "Exploring the influences on men's engagement with weight loss services: A qualitative study," *BMC Public Health*, vol. 20, no. 1, p. 249, 2020.
- [47] L. P. Svetkey et al., "Greater weight loss with increasing age in the weight loss maintenance trial," *Obesity*, vol. 22, no. 1, pp. 39–44, 2014.
- [48] L. Lissner et al., "Variability of body weight and health outcomes in the framingham population," *New England Journal of Medicine*, vol. 324, no. 26, pp. 1839–1844, 1991.
- [49] M. J. Park et al., "High body weight variability is associated with increased risk of depression: A nationwide cohort study in south korea," *Psychological Medicine*, vol. 53, no. 8, pp. 3719–3727, 2023.
- [50] I. Jung et al., "Increased risk of nonalcoholic fatty liver disease in individuals with high weight variability," *Endocrinology and Metabolism*, vol. 36, no. 4, pp. 845–854, 2021.
- [51] A. D. Kaze, P. Santhanam, S. Erqou, R. S. Ahima, A. G. Bertoni, and J. B. Echouffo-Tcheugui, "Body weight variability and risk of cardiovascular outcomes and death in the context of weight loss intervention among patients with type 2 diabetes," *JAMA Network Open*, vol. 5, no. 2, e220055–e220055, 2022.
- [52] E. H. Feig and M. R. Lowe, "Variability in weight change early in behavioral weight loss treatment: Theoretical and clinical implications," *Obesity*, vol. 25, no. 9, pp. 1509–1515, 2017.
- [53] A.-L. Orsama, E. Mattila, M. Ermes, M. Van Gils, B. Wansink, and I. Korhonen, "Weight rhythms: Weight increases during weekends and decreases during weekdays," *Obesity facts*, vol. 7, no. 1, pp. 36–47, 2014.
- [54] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American statistical association*, vol. 74, no. 368, pp. 829–836, 1979.

- [55] J. Turicchi et al., “Data imputation and body weight variability calculation using linear and nonlinear methods in data collected from digital smart scales: Simulation and validation study,” *JMIR mHealth and uHealth*, vol. 8, no. 9, e17977, 2020.
- [56] B. Binsaeed et al., “Barriers and motivators to weight loss in people with obesity,” *Cureus*, vol. 15, no. 11, 2023.
- [57] J. G. LaRose, A. Lanoye, D. Ferrell, J. Lu, and M. Mosavel, “Translating evidence-based behavioral weight loss into a multi-level, community intervention within a community-based participatory research framework: The wellness engagement (we) project,” *Translational Behavioral Medicine*, vol. 11, no. 6, pp. 1235–1243, 2021.
- [58] A. A. Gorin, T. A. Powers, R. Koestner, R. R. Wing, and H. A. Raynor, “Autonomy support, self-regulation, and weight loss,” *Health Psychology*, vol. 33, no. 4, p. 332, 2014.
- [59] J. Petch, S. Di, and W. Nelson, “Opening the black box: The promise and limitations of explainable machine learning in cardiology,” *Canadian Journal of Cardiology*, vol. 38, no. 2, pp. 204–213, 2022.
- [60] H. Okada et al., “Association between variability in body mass index and development of type 2 diabetes: Panasonic cohort study,” *BMJ Open Diabetes Research & Care*, vol. 9, no. 1, 2021.
- [61] E.-J. Rhee et al., “Increased risk of diabetes development in individuals with weight cycling over 4 years: The kangbuk samsung health study,” *Diabetes research and clinical practice*, vol. 139, pp. 230–238, 2018.
- [62] Y. Li et al., “Association of body-weight fluctuation with outcomes in heart failure with preserved ejection fraction,” *Frontiers in cardiovascular medicine*, vol. 8, p. 689 591, 2021.
- [63] A. R. Dyer, J. Stamler, and P. Greenland, “Associations of weight change and weight variability with cardiovascular and all-cause mortality in the chicago western electric company study,” *American journal of epidemiology*, vol. 152, no. 4, pp. 324–333, 2000.
- [64] G. Flore et al., “Weight maintenance after dietary weight loss: Systematic review and meta-analysis on the effectiveness of behavioural intensive intervention,” *Nutrients*, vol. 14, no. 6, p. 1259, 2022.

- [65] K. D. Hall and S. Kahan, "Maintenance of lost weight and long-term management of obesity," *The Medical Clinics of North America*, vol. 102, no. 1, p. 183, 2018.
- [66] M. Blüher, "Obesity: Global epidemiology and pathogenesis," *Nature reviews endocrinology*, vol. 15, no. 5, pp. 288–298, 2019.
- [67] C. J. Ruhm, "Understanding overeating and obesity," *Journal of Health economics*, vol. 31, no. 6, pp. 781–796, 2012.
- [68] S. Gonul, T. Namli, S. Huisman, G. B. Laleci Erturkmen, I. H. Toroslu, and A. Cosar, "An expandable approach for design and personalization of digital, just-in-time adaptive interventions," *Journal of the american medical informatics association*, vol. 26, no. 3, pp. 198–210, 2019.
- [69] Y. H. Yau and M. N. Potenza, "Stress and eating behaviors," *Minerva endocrinologica*, vol. 38, no. 3, p. 255, 2013.
- [70] H. M. Abdella, H. O. El Farssi, D. R. Broom, D. A. Hadden, and C. F. Dalton, "Eating behaviours and food cravings; influence of age, sex, bmi and fto genotype," *Nutrients*, vol. 11, no. 2, p. 377, 2019.
- [71] M. Tanofsky-Kraff, N. A. Schvey, and C. M. Grilo, "A developmental framework of binge-eating disorder based on pediatric loss of control eating," *American Psychologist*, vol. 75, no. 2, p. 189, 2020.
- [72] A. Godet, A. Fortier, E. Bannier, N. Coquery, and D. Val-Laillet, "Interactions between emotions and eating behaviors: Main issues, neuroimaging contributions, and innovative preventive or corrective strategies," *Reviews in Endocrine and Metabolic Disorders*, vol. 23, no. 4, pp. 807–831, 2022.
- [73] E. Ahlich and D. Rancourt, "Boredom proneness, interoception, and emotional eating," *Appetite*, vol. 178, p. 106167, 2022.
- [74] A. Modrzejewska, K. Czepczor-Bernat, J. Modrzejewska, and P. Matusik, "Eating motives and other factors predicting emotional overeating during covid-19 in a sample of polish adults," *Nutrients*, vol. 13, no. 5, p. 1658, 2021.
- [75] D. E. Wilfley, M. B. Schwartz, E. B. Spurrell, and C. G. Fairburn, "Using the eating disorder examination to identify the specific psychopathology of binge eating disorder," *International Journal of Eating Disorders*, vol. 27, no. 3, pp. 259–269, 2000.

- [76] N. I. Alshurafa et al., “Rationale and design of the sensewhy project: A passive sensing and ecological momentary assessment study on characteristics of overeating episodes,” *Digital health*, vol. 9, p. 20 552 076 231 158 314, 2023.
- [77] S. F. Schakel, *Maintaining a nutrient database in a changing marketplace: Keeping pace with changing food products—a research perspective*, 2001.
- [78] T. Baranowski, “24-hour recall and diet record methods,” *Nutritional epidemiology*, vol. 40, pp. 49–69, 2013.
- [79] C. Boushey, M. Spoden, F. Zhu, E. Delp, and D. Kerr, “New mobile methods for dietary assessment: Review of image-assisted and image-based dietary assessment methods,” *Proceedings of the Nutrition Society*, vol. 76, no. 3, pp. 283–294, 2017.
- [80] 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.
- [81] N. Alshurafa, R. Alharbi, A. F. Pfammatter, and B. Spring, “Detecting real-time episodic overeating for just-in-time interventions,” *Annals of Behavioral Medicine*, vol. 52, S726–S726, 2018. DOI: [10 . 1093 / abm / kay013](https://doi.org/10.1093/abm/kay013). [Online]. Available: [https : //doi . org / 10 . 1093 / abm / kay013](https://doi.org/10.1093/abm/kay013).
- [82] J. G. Thomas, S. Doshi, R. D. Crosby, and M. R. Lowe, “Ecological momentary assessment of obesogenic eating behavior: Combining person-specific and environmental predictors,” *Obesity*, vol. 19, no. 8, pp. 1574–1579, 2011.
- [83] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: Synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [84] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.
- [85] C. F. Eick, N. Zeidat, and Z. Zhao, “Supervised clustering-algorithms and benefits,” in *16Th IEEE international conference on tools with artificial intelligence*, IEEE, 2004, pp. 774–776.
- [86] S. Deng, L. Sheng, J. Nie, and F. Deng, “A clustering method based on information entropy payload,” *arXiv preprint arXiv:2209.06582*, 2022.

- [87] A. Rosenberg and J. Hirschberg, “V-measure: A conditional entropy-based external cluster evaluation measure,” in *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, 2007, pp. 410–420.
- [88] B. Wansink, “Environmental factors that increase the food intake and consumption volume of unknowing consumers,” *Annu. Rev. Nutr.*, vol. 24, no. 1, pp. 455–479, 2004.
- [89] J. O. Hill, H. R. Wyatt, and J. C. Peters, “Energy balance and obesity,” *Circulation*, vol. 126, no. 1, pp. 126–132, 2012.
- [90] 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.
- [91] 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.
- [92] A. A. Witt and M. R. Lowe, “Hedonic hunger and binge eating among women with eating disorders,” *International Journal of Eating Disorders*, vol. 47, no. 3, pp. 273–280, 2014.
- [93] B. J. Rolls, L. S. Roe, and J. S. Meengs, “Reductions in portion size and energy density of foods are additive and lead to sustained decreases in energy intake,” *The American journal of clinical nutrition*, vol. 83, no. 1, pp. 11–17, 2006.
- [94] M. Frayn and B. Knäuper, “Emotional eating and weight in adults: A review,” *Current Psychology*, vol. 37, no. 4, pp. 924–933, 2018.
- [95] F. Favieri, A. Marini, and M. Casagrande, “Emotional regulation and overeating behaviors in children and adolescents: A systematic review,” *Behavioral Sciences*, vol. 11, no. 1, p. 11, 2021.
- [96] J. N. Salley, A. W. Hoover, M. L. Wilson, and E. R. Muth, “Comparison between human and bite-based methods of estimating caloric intake,” *Journal of the Academy of Nutrition and Dietetics*, vol. 116, no. 10, pp. 1568–1577, 2016.
- [97] D. Hossain, T. Ghosh, and E. Sazonov, “Automatic count of bites and chews from videos of eating episodes,” *Ieee Access*, vol. 8, pp. 101 934–101 945, 2020.

- [98] Y. Zhu and J. H. Hollis, "Increasing the number of chews before swallowing reduces meal size in normal-weight, overweight, and obese adults," *Journal of the Academy of Nutrition and Dietetics*, vol. 114, no. 6, pp. 926–931, 2014.
- [99] J. Li et al., "Improvement in chewing activity reduces energy intake in one meal and modulates plasma gut hormone concentrations in obese and lean young chinese men-," *The American journal of clinical nutrition*, vol. 94, no. 3, pp. 709–716, 2011.
- [100] G. P. Teixeira et al., "Role of chronotype in dietary intake, meal timing, and obesity: A systematic review," *Nutrition Reviews*, vol. 81, no. 1, pp. 75–90, 2023.
- [101] N. Goel et al., "Circadian rhythm profiles in women with night eating syndrome," *Journal of biological rhythms*, vol. 24, no. 1, pp. 85–94, 2009.
- [102] K. C. Allison et al., "The night eating questionnaire (neq): Psychometric properties of a measure of severity of the night eating syndrome," *Eating behaviors*, vol. 9, no. 1, pp. 62–72, 2008.
- [103] A. J. Stunkard, K. Allison, J. Lundgren, and J. O'Reardon, "A biobehavioural model of the night eating syndrome," *Obesity Reviews*, vol. 10, pp. 69–77, 2009.
- [104] T. C. Adam and E. S. Epel, "Stress, eating and the reward system," *Physiology & behavior*, vol. 91, no. 4, pp. 449–458, 2007.
- [105] S. J. Torres and C. A. Nowson, "Relationship between stress, eating behavior, and obesity," *Nutrition*, vol. 23, no. 11-12, pp. 887–894, 2007.
- [106] C. G. Greeno and R. R. Wing, "Stress-induced eating.," *Psychological bulletin*, vol. 115, no. 3, p. 444, 1994.
- [107] H. Sultson, K. Kuk, and K. Akkermann, "Positive and negative emotional eating have different associations with overeating and binge eating: Construction and validation of the positive-negative emotional eating scale," *Appetite*, vol. 116, pp. 423–430, 2017.
- [108] A. Torske, B. Bremer, B. K. Hölzel, A. Maczka, and K. Koch, "Mindfulness meditation modulates stress-eating and its neural correlates," *Scientific Reports*, vol. 14, no. 1, p. 7294, 2024.

- [109] C. P. Herman, D. A. Roth, and J. Polivy, "Effects of the presence of others on food intake: A normative interpretation.," *Psychological bulletin*, vol. 129, no. 6, p. 873, 2003.
- [110] J. M. De Castro, "Family and friends produce greater social facilitation of food intake than other companions," *Physiology & behavior*, vol. 56, no. 3, pp. 445–455, 1994.
- [111] C. P. Herman, "The social facilitation of eating. a review," *Appetite*, vol. 86, pp. 61–73, 2015.
- [112] E. Robinson, H. Benwell, and S. Higgs, "Food intake norms increase and decrease snack food intake in a remote confederate study," *Appetite*, vol. 65, pp. 20–24, 2013.
- [113] S. A. Rydell, L. J. Harnack, J. M. Oakes, M. Story, R. W. Jeffery, and S. A. French, "Why eat at fast-food restaurants: Reported reasons among frequent consumers," *Journal of the American Dietetic Association*, vol. 108, no. 12, pp. 2066–2070, 2008.
- [114] R. Rosenheck, "Fast food consumption and increased caloric intake: A systematic review of a trajectory towards weight gain and obesity risk," *Obesity reviews*, vol. 9, no. 6, pp. 535–547, 2008.
- [115] D. Cohen and T. A. Farley, "Eating as an automatic behavior," *Preventing chronic disease*, vol. 5, no. 1, A23, 2007.
- [116] M. Tanofsky-Kraff et al., "A prospective study of pediatric loss of control eating and psychological outcomes.," *Journal of abnormal psychology*, vol. 120, no. 1, p. 108, 2011.
- [117] B. Wansink and J. Sobal, "Mindless eating: The 200 daily food decisions we overlook," *Environment and Behavior*, vol. 39, no. 1, pp. 106–123, 2007.
- [118] A. Acosta et al., "Selection of antiobesity medications based on phenotypes enhances weight loss: A pragmatic trial in an obesity clinic," *Obesity*, vol. 29, no. 4, pp. 662–671, 2021.
- [119] E. Koning, J. Vorstman, R. S. McIntyre, and E. Brietzke, "Characterizing eating behavioral phenotypes in mood disorders: A narrative review," *Psychological Medicine*, vol. 52, no. 14, pp. 2885–2898, 2022.
- [120] D. A. Schoeller, "Limitations in the assessment of dietary energy intake by self-report," *Metabolism*, vol. 44, pp. 18–22, 1995.

- [121] Y. Dong, A. Hoover, J. Scisco, and E. Muth, "A new method for measuring meal intake in humans via automated wrist motion tracking," *Applied psychophysiology and biofeedback*, vol. 37, no. 3, pp. 205–215, 2012.
- [122] C. E. F. d. Moraes, M. M. L. Antunes, C. Mourilhe, R. Sichieri, P. Hay, and J. C. Apolinario, "Food consumption during binge eating episodes in binge eating spectrum conditions from a representative sample of a brazilian metropolitan city," *Nutrients*, vol. 15, no. 7, p. 1573, 2023.
- [123] T. Van Strien, C. P. Herman, and M. W. Verheijden, "Eating style, overeating, and overweight in a representative dutch sample. does external eating play a role?" *Appetite*, vol. 52, no. 2, pp. 380–387, 2009.
- [124] A. S. Tarawneh, A. B. Hassanat, G. A. Altarawneh, and A. Almuhaimeed, "Stop oversampling for class imbalance learning: A review," *IEEE Access*, vol. 10, pp. 47 643–47 660, 2022.
- [125] 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.
- [126] J. Macdiarmid and J. Blundell, "Assessing dietary intake: Who, what and why of under-reporting," *Nutrition research reviews*, vol. 11, no. 2, pp. 231–253, 1998.
- [127] H. Suh, N. Shahriaree, E. B. Hekler, and J. A. Kientz, "Developing and validating the user burden scale: A tool for assessing user burden in computing systems," in *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, pp. 3988–3999.
- [128] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and psychopathology*, vol. 17, no. 3, pp. 715–734, 2005.
- [129] E. M. Forman et al., "Ecological momentary assessment of dietary lapses across behavioral weight loss treatment: Characteristics, predictors, and relationships with weight change," *Annals of Behavioral Medicine*, vol. 51, no. 5, pp. 741–753, 2017.
- [130] S. M. Manasse et al., "Are individuals with loss-of-control eating more prone to dietary lapse in behavioural weight loss treatment? an ecological momentary assessment study," *European Eating Disorders Review*, vol. 26, no. 3, pp. 259–264, 2018.

- [131] S. D. Emmerich, C. D. Fryar, B. Stierman, and C. L. Ogden, "Obesity and severe obesity prevalence in adults: United states, august 2021–august 2023," 2024.
- [132] I. Romieu et al., "Energy balance and obesity: What are the main drivers?" *Cancer causes & control*, vol. 28, pp. 247–258, 2017.
- [133] G. A. Bray et al., "The science of obesity management: An endocrine society scientific statement," *Endocrine reviews*, vol. 39, no. 2, pp. 79–132, 2018.
- [134] R. R. Wing and S. Phelan, "Long-term weight loss maintenance," *The American journal of clinical nutrition*, vol. 82, no. 1, 222S–225S, 2005.
- [135] O.-R. Ha and S.-L. Lim, "The role of emotion in eating behavior and decisions," *Frontiers in Psychology*, vol. 14, p. 1265074, 2023.
- [136] G. Privitera, K. Cooper, and A. Cosco, "The influence of eating rate on satiety and intake among participants exhibiting high dietary restraint," *Food & nutrition research*, vol. 56, no. 1, p. 10202, 2012.
- [137] I. Ioakimidis, M. Zandian, L. Eriksson-Marklund, C. Bergh, A. Grigoriadis, and P. Södersten, "Description of chewing and food intake over the course of a meal," *Physiology & behavior*, vol. 104, no. 5, pp. 761–769, 2011.
- [138] S.-q. Yuan et al., "Association between eating speed and metabolic syndrome: A systematic review and meta-analysis," *Frontiers in nutrition*, vol. 8, p. 700936, 2021.
- [139] D. Garg, E. Smith, and T. Attuquayefio, "Watching television while eating increases food intake: A systematic review and meta-analysis of experimental studies," *Nutrients*, vol. 17, no. 1, p. 166, 2025.
- [140] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of things (iot): A vision, architectural elements, and future directions," *Future generation computer systems*, vol. 29, no. 7, pp. 1645–1660, 2013.
- [141] W. Yu et al., "A survey on the edge computing for the internet of things," *IEEE access*, vol. 6, pp. 6900–6919, 2017.
- [142] S. Zhang et al., "Deep learning in human activity recognition with wearable sensors: A review on advances," *arXiv preprint arXiv:2111.00418*, 2021.
- [143] S. Zhang, D. Nguyen, Z. King, J. Pradeep, and N. Alshurafa, "Habits necklace: A neck-worn sensor that captures eating related behavior and more," in *Proceedings*

of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, 2018, pp. 484–487.

- [144] S. Akther, N. Saleheen, S. A. Samiei, V. Shetty, E. Ertin, and S. Kumar, “Moral: An mhealth model for inferring oral hygiene behaviors in-the-wild using wrist-worn inertial sensors,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 1, pp. 1–25, 2019.
- [145] S. Akther, N. Saleheen, M. Saha, V. Shetty, and S. Kumar, “Mteeth: Identifying brushing teeth surfaces using wrist-worn inertial sensors,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 2, pp. 1–25, 2021.
- [146] R. Alharbi, B. Spring, and N. Alshurafa, “Measuring smoking topography in natural settings using non-contact passive wearable sensors,” in *ANNALS OF BEHAVIORAL MEDICINE*, OXFORD UNIV PRESS INC JOURNALS DEPT, 2001 EVANS RD, CARY, NC 27513 USA, vol. 54, 2020, S580–S580.
- [147] E. Kim, S. Helal, and D. Cook, “Human activity recognition and pattern discovery,” *IEEE pervasive computing*, vol. 9, no. 1, pp. 48–53, 2009.
- [148] Z. Ji, Q. Wei, A. Franklin, T. Cohen, and H. Xu, “Cost-sensitive active learning for phenotyping of electronic health records,” *AMIA Summits on Translational Science Proceedings*, vol. 2019, p. 829, 2019.
- [149] E. M. Tapia, S. S. Intille, and K. Larson, “Activity recognition in the home using simple and ubiquitous sensors,” in *International conference on pervasive computing*, Springer, 2004, pp. 158–175.
- [150] Y. Vaizman, K. Ellis, G. Lanckriet, and N. Weibel, “Extrasensory app: Data collection in-the-wild with rich user interface to self-report behavior,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–12.
- [151] C. A. Merck, C. Maher, M. Mirtchouk, M. Zheng, Y. Huang, and S. Kleinberg, “Multimodality sensing for eating recognition,” in *PervasiveHealth*, 2016, pp. 130–137.
- [152] J. Jiang, R. Pozza, K. Gunnarsdóttir, N. Gilbert, and K. Moessner, “Using sensors to study home activities,” *Journal of Sensor and Actuator Networks*, vol. 6, no. 4, p. 32, 2017.

- [153] B. Settles, “From theories to queries: Active learning in practice,” in *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, JMLR Workshop and Conference Proceedings, 2011, pp. 1–18.
- [154] H. Hino, “Active learning: Problem settings and recent developments,” *arXiv preprint arXiv:2012.04225*, 2020.
- [155] D. Angluin, “Queries and concept learning,” *Machine learning*, vol. 2, no. 4, pp. 319–342, 1988.
- [156] D. D. Lewis and W. A. Gale, “A sequential algorithm for training text classifiers,” in *SIGIR’94*, Springer, 1994, pp. 3–12.
- [157] R. Adaimi and E. Thomaz, “Leveraging active learning and conditional mutual information to minimize data annotation in human activity recognition,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–23, 2019.
- [158] S.-J. Huang, R. Jin, and Z.-H. Zhou, “Active learning by querying informative and representative examples,” *Advances in neural information processing systems*, vol. 23, pp. 892–900, 2010.
- [159] E. Aldana-Bobadilla and A. Kuri-Morales, “A clustering method based on the maximum entropy principle,” *Entropy*, vol. 17, no. 1, pp. 151–180, 2015.
- [160] Z. E. Ashari, N. S. Chaytor, D. J. Cook, and H. Ghasemzadeh, “Memory-aware active learning in mobile sensing systems,” *IEEE Transactions on Mobile Computing*, 2020.
- [161] D. Kottke, J. Schellinger, D. Huseljic, and B. Sick, “Limitations of assessing active learning performance at runtime,” *arXiv preprint arXiv:1901.10338*, 2019.
- [162] G. Pleiss, T. Zhang, E. R. Elenberg, and K. Q. Weinberger, “Identifying mislabeled data using the area under the margin ranking,” *arXiv preprint arXiv:2001.10528*, 2020.
- [163] Y. Gal, “Uncertainty in deep learning,” 2016.
- [164] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.