

Generalization and Personalization of Machine Learning for Multimodal Mobile Sensing in Everyday Life

Thèse n. 1234 2023
présentée le 22 novembre 2023
à la Faculté des sciences et techniques de l'ingénieur
Laboratoire de l'IDIAP
Programme doctoral en génie électrique
École polytechnique fédérale de Lausanne
pour l'obtention du grade de Docteur ès Sciences
par

Lakmal Buddika Meegahapola

acceptée sur proposition du jury :

Prof Touradj Ebrahimi, président du jury
Prof Daniel Gatica-Perez, directeur de thèse
Prof Silvia Santini, rapporteur
Prof Mirco Musolesi, rapporteur
Prof Marcel Salathé, rapporteur

Lausanne, EPFL, 2023



To the guiding light in my life, my mother, whose unwavering love, sacrifices, and boundless strength have shaped me into who I am today...

To my father, whose wisdom, patience, steady presence, and unspoken care have been my steadfast pillars...

To my loving wife, who stood by my side throughout the challenging journey of my PhD, turning every obstacle into a shared triumph. Your persistent support, belief, and the way you weathered every storm with me made every success sweeter and every hurdle conquerable...



Acknowledgements

First of all, I would like to thank my amazing thesis advisor, Daniel Gatica-Perez, without whom this thesis would not have been possible. From the very first day I met him at ACM UbiComp 2018, till today, I have come to admire his kindness, patience, thoughtfulness, and clarity of mind. His valuable advice and belief in my work allowed me to grow as an independent researcher and a person. I could not have asked for a better advisor. Thank you so much Daniel, for giving me this wonderful opportunity to do a Ph.D. with you. Special thanks to the Ph.D. committee members, Silvia, Mirco, Marcel, and Touradj, for taking their valuable time to review my thesis, as well as for their insightful comments and feedback.

I would like to express my gratitude to my research collaborators, starting with all the consortium members of the EU H2020 WeNet project. While I won't list all the names from over 15 institutes in more than 10 countries, I want to acknowledge that this thesis would not have been possible without their support and extensive collaborative effort. It was a pleasure working with all of you. Furthermore, I extend my sincere appreciation to Cecilia for her kindness and invaluable advice during and after my research visit to the University of Cambridge. I'm deeply thankful for the internship opportunities provided by Venky, Kai, and Vidhya at Google, and by Marios, Zoran, Hongwei, Sanja, Michael, and Daniele at Nokia Bell Labs. I will always be grateful for these opportunities and the immense learning I gained from these research experiences. My back-to-back visits to Cambridge and the Bay Area were among the highlights of my Ph.D. Moreover, I'm deeply thankful to my amazing undergraduate research advisors from the University of Moratuwa, Dulani and Indika, who ignited my passion for research and helped me immensely. I also extend my thanks to Sampath from Old Dominion University, my external research advisor during my undergraduate studies, who provided us with invaluable assistance. My heartfelt gratitude goes to Archan, who played a pivotal role in my early research career at Singapore Management University, offering so much guidance and support. I also thank the European Commission and the Swiss National Science Foundation for funding my research through the WeNet and Dusk2Dawn projects, respectively.

Spending time in the middle of the Alps, amidst all the majestic mountains, was like a dream, especially as a hiking enthusiast. Even though I wasn't able to do as many hikes as I would have liked, I thoroughly enjoyed my time here. The last 4.5 years were fun, thanks to the wonderful group of friends and colleagues at Idiap. Cheers to all the hikes, meals, meetups, cricket matches, drinks, and good times. Special thanks to Sylvie, Laura, and Nardine for their immense effort in keeping Idiap running smoothly and making our lives easier in many ways. I wish to express my heartfelt gratitude to my parents for their unwavering love, sacrifices, and support. Being over 4,000 miles apart, I cannot emphasize enough how much I miss you both every day and how deeply I wish I could return to stay with you. Lastly, I want to express my profound appreciation to my loving wife, Wathsala. I feel truly blessed to have you by my side, and I am grateful for all the sacrifices you have made to be here with me. You are so much more than I deserve from you as a person, a friend, and a wife.

Martigny, November 22, 2023

Lakmal Meegahapola

Abstract

A range of behavioral and contextual factors, including eating and drinking behavior, mood, social context, and other daily activities, can significantly impact an individual's quality of life and overall well-being. Therefore, inferring everyday life aspects with the use of smartphone and wearable sensors, also broadly known as mobile sensing, is gaining traction across both clinical and non-clinical populations due to the widespread use of smartphones around the world. Such inferences are of use in mobile health apps, mobile food diaries, and generic mobile apps. However, despite the long-standing promise in the domain, realizing the full potential of models, in the wild, is still far from reality due to two primary deployment challenges: the generalization and personalization of models. In addition, there are understudied domains, such as eating and drinking behavior modeling with multimodal mobile sensing and machine learning. Hence, this thesis delves into the realm of multimodal mobile sensing with an eye for the generalization and personalization of models, exploring a range of novel inferences at the intersection of eating and drinking behavior, mood, daily activities, and context.

After introducing the topic in the first chapter and discussing data collection in the second, we expand on passive sensing for drink behavior modeling using multimodal sensor data in the third chapter. The fourth chapter demonstrates how smartphone sensors can infer self-perceived food consumption levels with personalized models. The fifth chapter showcases how phone sensors could be used to infer eating events with personalized models. The sixth chapter highlights the challenge of generic mood inference models struggling to adapt to specific contexts like eating. To tackle this, we propose a personalization technique to enhance model performance even with limited data. In the next three chapters, we delve further into the realm of model generalization within the context of multimodal mobile sensing. We also investigate the impact of personalization on generalization performance. Specifically, we investigate model generalization across countries—a problem that has been scarcely addressed in prior research. To this end, in the seventh chapter, we examine the generalization capabilities of mood inference models, while the eighth chapter focuses on the generalization of models for complex daily activity recognition. Upon highlighting the limitations of model generalization in the aforementioned chapters, we introduce a novel technique to enhance model generalization in the context of multimodal sensor data in the ninth chapter.

In summary, this thesis offers an extensive exploration of novel inferences and deployment challenges in multimodal mobile sensing. First, the thesis explores eating and drinking behavior and its interplay with mood, social context, and daily activities, viewed through the lens of both model personalization and generalization. Additionally, the thesis delves into the challenge of cross-country generalization for mobile sensing-based models and presents a novel deep learning architecture for unsupervised domain adaptation, yielding enhanced performance in unfamiliar domains. As a result, this thesis contributes both empirically and methodologically to the fields of ubiquitous and mobile computing and digital health.

Abstract

Keywords: Mobile Sensing, Passive Sensing, Multimodal, Smartphone Sensing, Sensors, Digital Health, Behavior Modeling, Context-Awareness, Human-Centered Artificial Intelligence, Machine Learning, Deep Learning, Eating, Drinking, Mood, Activity Recognition

Résumé

Un ensemble de facteurs comportementaux et contextuels, notamment le comportement en matière d'alimentation et de consommation d'alcool, l'humeur, le contexte social et d'autres activités quotidiennes, peuvent avoir un impact significatif sur la qualité de vie et le bien-être général d'un individu. Par conséquent, l'inférence des aspects de la vie quotidienne à l'aide de smartphones et de capteurs portables, également connue sous le nom de détection mobile, gagne du terrain auprès des populations cliniques et non cliniques. Cependant, malgré des promesses de longue date dans le domaine, accomplir le plein potentiel des modèles, dans la nature, est encore loin d'être une réalité en raison de deux principaux défis liés au déploiement : la généralisation et la personnalisation des modèles. Par conséquent, cette thèse explore le domaine de la détection mobile multimodale avec un intérêt pour la généralisation et la personnalisation des modèles, explorant une gamme de nouvelles inférences à l'intersection du comportement alimentaire, du comportement de consommation d'alcool, de l'humeur, des activités quotidiennes et du contexte.

Nous nous appuyons, dans le troisième chapitre, sur les recherches antérieures en détection passive dans le domaine de la modélisation du comportement de consommation d'alcool et montrons comment les données de capteurs multimodaux peuvent améliorer la compréhension du contexte social lors d'événements de consommation d'alcool. Dans le quatrième chapitre, illustrez comment les capteurs des smartphones peuvent indiquer les niveaux de consommation alimentaire perçus par les individus. Le cinquième chapitre démontre comment les capteurs des smartphones peuvent fournir des indices liés aux événements alimentaires, facilitant ainsi l'inférence de tels événements grâce à des modèles personnalisés. Dans le sixième chapitre, nous révélons que les modèles génériques d'inférence d'humeur, entraînés avec des données collectées dans divers contextes, pourraient avoir du mal à bien généraliser à des situations spécifiques comme dans le cas d'événements alimentaires, ce qui entraînerait une diminution des performances. Nous introduisons une nouvelle technique de personnalisation qui améliore les performances individuelles même avec des données limitées.

Dans la partie suivante de la thèse, nous étudions également l'impact de la personnalisation sur les performances de généralisation. Plus précisément, nous nous penchons sur le défi de la généralisation des modèles à travers les pays, un problème qui a à peine été abordé dans les recherches antérieures. Dans le septième chapitre, nous examinons les capacités de généralisation des modèles d'inférence d'humeur, tandis que le huitième chapitre se concentre sur la généralisation de modèles de reconnaissance d'activités quotidiennes complexes. Après avoir souligné les limites de la généralisation, nous introduisons, dans le neuvième chapitre, une nouvelle technique pour améliorer la généralisation des modèles dans le contexte de données provenant de capteurs multimodaux.

En résumé, cette thèse propose une exploration approfondie de nouvelles inférences et des défis liés au déploiement dans le domaine de la détection mobile multimodale.

Résumé

Mots clés : détection mobile, détection passive, multimodal, détection par smartphone, capteurs, santé numérique, modélisation du comportement, conscience du contexte, intelligence artificielle centrée sur l'humain, apprentissage automatique, apprentissage profond, manger, boire, humeur, reconnaissance d'activité

Contents

Acknowledgements	i
Abstract (English/Français)	iii
List of Figures	xiii
List of Tables	xvii
1 Introduction	1
1.1 Research Context	1
1.1.1 “In the Wild” vs. “In Lab” Studies	2
1.1.2 Passive Sensing and Self-Reports in “In the Wild” Studies	2
1.2 Motivation	4
1.2.1 Understanding the Interplay between Eating and Drinking Behavior, Mood, Daily Activities, and Context	4
1.2.2 Generalization and Personalization of Models	5
1.3 Summary of Contributions and Dissertation Outline	6
1.4 Publications	8
2 Datasets	11
2.1 Mexico Dataset (MEX)	11
2.1.1 Mobile Application	11
2.1.2 Data Collection and Pre-Processing	13
2.2 Multi-Country Dataset (MUL)	15
2.2.1 LimeSurvey Questionnaires	15
2.2.2 Mobile Application	16
2.2.3 Data Collection and Pre-Processing	17
3 Examining the Social Context of Alcohol Drinking of Young Adults with Smartphone Sensing	21
3.1 Introduction	21
3.2 Background and Related Work	23
3.2.1 The Social Context of Drinking Alcohol	23
3.2.2 Alcohol Consumption and Mobile Phones	25
3.3 Data, Features, and Tasks	26
3.3.1 Mobile Application, Self-Reports, and Passive Sensing	26
3.3.2 Aggregation and Matching of Self-Reports and Passive Sensing Data	27
3.3.3 Deriving Two-Class and Three-Class Social Context Features	29
3.4 Descriptive Analysis (RQ1)	30
3.5 Statistical Analysis (RQ1)	32
	vii

Contents

3.5.1	Pearson and Point-Biserial Correlation for Social Contexts and Passive Sensing Features	32
3.5.2	Statistical Analysis of Dataset Features	33
3.6	Social Context Inference	35
3.6.1	Two-Class and Three-Class Social Context Inference (RQ2)	35
3.6.2	Social Context Inference for Different Sensors (RQ2)	36
3.6.3	Feature Importance for Social Context Inferences (RQ2)	37
3.6.4	Effect of Varying Group Sizes (RQ2)	39
3.7	Discussion	41
3.8	Conclusion	43
4	Inferring the Food Consumption Level of College Students with Smartphone Sensing	45
4.1	Introduction	45
4.2	Defining Food Consumption Level	47
4.3	Related Work	48
4.3.1	Internal and Contextual Factors Affecting Overeating	48
4.3.2	Mobile Health Apps to Analyze Eating Behavior	49
4.3.3	Mobile Sensing to Analyze Eating Behavior	49
4.3.4	Study Objective, Hypothesis, and Dataset	50
4.4	Descriptive Data Analysis (RQ1)	51
4.4.1	Food Types and Food Consumption Level	51
4.4.2	Sociability and Food Consumption Level	53
4.4.3	Mood, Stress, and Food Consumption Level	53
4.4.4	Concurrent Activities and Food Consumption Level	54
4.5	Statistical Analysis (RQ1)	54
4.5.1	Pearson and Point-Biserial Correlation for Self-Report Features	54
4.5.2	Statistical Analysis of Dataset Features	54
4.6	Food Consumption Level Inference (RQ2)	56
4.6.1	Three-class Food Consumption Level Inference	56
4.6.2	Model Personalization	60
4.7	Discussion	61
4.7.1	Feedback from Participants	61
4.7.2	Passive Smartphone Sensing for Characterizing Food Consumption Levels	62
4.7.3	Further Informative Features Regarding Food Consumption Levels.	62
4.7.4	Accounting for Diversity	62
4.7.5	Limitations and Future Directions	63
4.8	Conclusion	64
5	Sensing Eating Events in Context: A Smartphone-Only Approach	65
5.1	Introduction	65
5.2	Background and Related Work	68
5.2.1	Nutrition Science Perspective	68
5.2.2	Mobile Food Diaries	69
5.2.3	Mobile Sensing for Eating Behavior Monitoring	69
5.2.4	Smartphone Sensing for Eating Behavior Monitoring	70
5.3	Data, Smartphone Features, and Definition of Eating/non-eating Episodes	71
5.3.1	Dataset	71
5.3.2	Ground Truth and Passive Sensing Data	72

5.3.3	What is an Eating Event?	72
5.4	Descriptive and Statistical Analysis of Sensor Data and Eating Events (RQ1)	73
5.5	Detecting Eating Events and Important Features (RQ2)	75
5.5.1	Two-Class Eating Event Inference	75
5.5.2	Feature Importance for Eating event Detection (RQ2)	77
5.5.3	Effect of Personalization on Individuals	78
5.6	Discussion and Limitations	79
5.7	Conclusion	82
6	Inferring the Mood-While-Eating with Community Based Personalization	83
6.1	Introduction	83
6.2	Definitions, Background and Related Work	87
6.2.1	Defining Mood-While-Eating	87
6.2.2	Mood and Food Consumption	87
6.2.3	Mood and Smartphone Technologies	88
6.2.4	Eating as a Holistic Event	89
6.2.5	Smartphone Sensing Inference Personalization	89
6.3	Dataset Description	91
6.3.1	Mexico Dataset (MEX)	91
6.3.2	Multi-Country Dataset (MUL)	92
6.4	Methods	92
6.4.1	Generic Models and Context-specific Models Analysis (RQ1)	93
6.4.2	Mood-While-Eating Inference (RQ2)	94
6.4.3	Community-Based Model Personalization Approach (RQ3)	94
6.5	Results	96
6.5.1	Generic and Context-specific Model Performance (RQ1)	96
6.5.2	Mood-While-Eating Inference (RQ2)	97
6.5.3	Mood-While-Eating Inference Using Community Based Personalization Approach (RQ3)	98
6.6	Discussion	101
6.6.1	Summary of Results	101
6.6.2	Implications	102
6.6.3	Limitations and Future Work	102
6.7	Conclusion	103
7	Generalization and Personalization of Mobile Sensing-Based Mood Inference Models	105
7.1	Introduction	105
7.2	Background and Related Work	107
7.2.1	Definitions and Terminology	107
7.2.2	Considerations for Research in Mobile Sensing Involving Geographic Diversity	109
7.2.3	Mood and Smartphone Technologies	111
7.3	Behavioral and Contextual Characteristics Around Mood Reports Extracted from Sensor Data and Self-Reports (RQ1)	112
7.3.1	Descriptive Analysis	112
7.3.2	Statistical Analysis	114
7.4	Mood Inference (RQ2 & RQ3)	116
7.4.1	Experimental Setup	116
7.4.2	Results	117

7.5	Discussion	122
7.5.1	What do the Results Suggest?	123
7.5.2	Comparison of Results to Previous Studies	123
7.5.3	Diversity-Aware Research in Mobile Sensing	124
7.5.4	Diversity-Awareness: Countries or Cultures?	124
7.5.5	Ethical Considerations	124
7.5.6	The Effect of the Pandemic and Weather on Mood Inference Models	125
7.5.7	Domain Adaptation for Multi-Modal Mobile Sensing	125
7.5.8	Other Limitations and Future Work	126
7.6	Conclusion	126
8	Complex Daily Activities, Country-Level Diversity, and Smartphone Sensing	127
8.1	Introduction	127
8.2	Background and Related Work	130
8.2.1	Mobile Sensing	130
8.2.2	Activity Recognition	131
8.2.3	Activity Types	132
8.2.4	Diversity-Awareness in Smartphone Sensing	133
8.2.5	Human-Centered Aspects in Smartphone Usage	134
8.3	Data, Features, and Target Classes	134
8.3.1	Dataset Information	134
8.3.2	Determining Target Classes	135
8.4	How are activities expressed in different countries, and what smartphone features are most discriminant? (RQ1)	135
8.4.1	Hourly Distribution of Activities	136
8.4.2	Statistical Analysis of Features	137
8.5	Machine Learning-based Inference: Experimental Setup, Models, and Performance Measures	139
8.5.1	Data Imputation	140
8.5.2	Models and Performance Measures	140
8.6	Inference Results	142
8.6.1	Country-Specific and Multi-Country Approaches (RQ2)	142
8.6.2	Generalization Issues with Country-Agnostic and Country-Specific Models (RQ3)	144
8.7	Discussion	146
8.7.1	Summary of Results	146
8.7.2	Implications	147
8.7.3	Limitations	149
8.7.4	Future Work	150
8.8	Conclusion	151
9	Unsupervised Domain Adaptation for Multimodal Mobile Sensing with Multi-Branch Adversarial Training	153
9.1	Introduction	153
9.2	Background and Related Work	156
9.2.1	Distribution Shift and Unsupervised Domain Adaptation	156
9.2.2	Mobile Sensing for Inferences Regarding Health and Well-Being, Behavior, and Context	157
9.2.3	Generalization and Distribution Shift in Mobile Sensing	158

9.3 Datasets	159
9.3.1 MUL: Multimodal Smartphone Sensing Dataset from 8 Countries	159
9.3.2 WEEE: Multimodal Wearable Sensing Dataset for Energy Expenditure Estimation	159
9.3.3 Domains and Inferences	160
9.4 M3BAT Architecture	161
9.4.1 Unsupervised Domain Adaptation with Domain Adversarial Training	161
9.4.2 Multiple Branches to Process Multimodal Data	162
9.4.3 Training Process with Multimodal Domain Adversarial Training	163
9.5 Using Statistical Tests to Quantify Distribution Shift of Sensors (RQ1)	164
9.5.1 Methodology	164
9.5.2 Results	166
9.6 Domain Adversarial Training with Multimodal Sensing Features (RQ2)	167
9.6.1 Methodology	167
9.6.2 Results	168
9.7 Multi-Branch Domain Adversarial Training (RQ3)	170
9.7.1 Methodology	170
9.7.2 Results	171
9.8 Discussion	173
9.8.1 Implications	173
9.8.2 Limitations and Future Work	174
9.9 Conclusion	175
10 Conclusion	177
10.1 Summary of Contributions	177
10.2 Limitations and Future Work	179
10.2.1 Data Imbalance, Diversity, and Sampling Biases	179
10.2.2 Privacy and Ethical Considerations	180
10.2.3 Generalization, Distribution Shift, and Cross-Cultural Analysis	180
10.2.4 Feature Extraction and Interpretability	180
10.2.5 Clinical Validity of Prediction Outcomes	181
10.2.6 Transfer Learning and Personalization	181
10.2.7 Scalability and User-Friendliness	181
A Appendix	183
A.1 Feature Groups Used in Different Inference Models (Chapter 5)	183
A.2 WEEE Dataset Features from [18] (Chapter 9)	183
A.3 Selecting an Initial Threshold for Target Users with MEX dataset (Chapter 6)	184
A.4 Algorithm for Data Aggregation (Chapter 6)	185
A.5 Algorithm for Community Detection (Chapter 6)	186
A.6 MEX Dataset Features (Chapter 2)	187
Bibliography	227
Curriculum Vitae	229

List of Figures

1.1	High-level overview of application areas covered in the thesis. The three main topics are eating or drinking behavior, activity or content, and mood. Most chapters focus on either two or three of the above application areas.	5
1.2	High-level overview of machine learning models covered in the thesis. The three main categorizations are chapters examining generic models, generalized models, and personalized models. Most chapters focus on either two or three of the above model types. . .	6
2.1	Block Diagram of Data Collection	12
2.2	High-level overview of the study.	16
3.1	A schematic diagram representing the summary of the study	29
3.2	Diagram summarizing the two-phase technique for combining self-reports and sensing data	29
3.3	Distribution of original self-report features (family, friends) and a derived feature (people)	30
3.4	Self-Reports in terms of Age	31
3.5	Distribution of two-class social contexts	31
3.6	Distribution of three-class social contexts	31
3.7	Self-Reports in terms of Sex and Two-Class Social Context	32
3.8	Self-Reports in terms of Sex and Three-Class Social Context	32
3.9	Feature importance values from random forest classifiers with all features, for different social contexts.	38
3.10	Feature importance value distributions for different sensing modalities in different inferences when using all features	39
3.11	Three-Class Social Context Inference Accuracies for Different Grouping Thresholds . .	40
4.1	Objective of the Study	51
4.2	Bar charts for food intake reports with hour of the day in the x-axis and number of eating events in the y-axis. "how_many" is the self-reported value from the food intake questionnaire regarding how many eating episodes have occurred in the past four hours.	51
4.3	Bar charts for ten different food categories. In each subplot, the x-axis indicates the three classes regarding food consumption level, and the y-axis indicates the total number of reported cases.	52
4.4	Heatmap for hourly food consumption episodes considering snack intake. "hours elapsed" denotes the # of hours from the start of the day.	52
4.5	Heatmap for hourly food consumption episodes considering meal intake. "hours elapsed" denotes the # of hours from the start of the day.	52
4.6	Bar chart representing the number of eating events related to different social contexts and associated food consumption levels.	52

List of Figures

4.7	Bar chart representing reported concurrent activities done while eating.	53
4.8	Mood while eating and count of reported instances.	53
4.9	Stress while eating and count of reported instances.	53
4.10	Feature Importance generated using the RF for G5 Inference	59
4.11	Accuracies for different feature group combinations among GEN and PERS.	59
5.1	Objective of the Study	70
5.2	Time Window for Sensor Data Aggregation	72
5.3	Violin plots of six selected accelerometer features for all the users and for two randomly selected users	73
5.4	Distribution of eating and non-eating events for different radius of gyration values for all the users and for two randomly selected users	74
5.5	Distribution of app usage for eating and non-eating events for all users and for two randomly selected users	74
5.6	Three Phases of the Study	75
5.7	Feature importance values from RFs for (a) BASE and (b)–(d) PERS2 models from three randomly selected users.	78
6.1	Original and Three-Class Mood Distributions of datasets	91
6.2	High-Level Architectural View of the Community-based Personalization Approach	94
6.3	(a) This shows the context-specific accuracies of the generic mood inference model trained for MUL for approaches PLM and HM. (b) This shows the overall mood inference accuracy (All) and accuracy for eating events (Mood-While-Eating) when the number of mood-while-eating reports in the training set changes, with HMs for MUL.	97
6.4	Mean accuracy values(column values in the graphs) calculated using the random forest for CBM for multiple threshold values [No. of Users] and averaged community sizes (line values in the graphs)	98
6.5	MEX dataset: Distribution change in no. of users with the increase of no. of data points in the negative class.	99
6.6	Mean accuracies calculated using the random forest for the PLM and HM of mood detection task for all the users and Maximum mean accuracies for CBM calculated for all the users.	100
6.7	Distributions of CBM using MEX dataset: (a) How the community size of each user changes with threshold th ; (b) How the mean accuracy of the Random Forest model changes with threshold th ; (c) Cumulative Distribution Function (CDF) of Accuracy.	101
7.1	High-level overview of the study.	112
7.2	Summary of self-reported mood distributions.	113
7.3	Distribution of self-reported moods for 24 hours of the day.	113
7.4	Location and social context distributions for negative and positive mood.	114
7.5	Cohen's-d (effect Size) distribution of features for negative and positive classes, grouped by countries and modalities.	114
7.6	Country-Specific HM: Gini feature importance values from RF models for two-class inference.	121
7.7	Country-Specific HM: Gini feature importance values from RF models for three-class inference.	121

8.1	High-level overview of the study. The study uses continuous and interaction sensing modalities and different approaches (country-specific, country-agnostic, and multi-country) to infer complex daily activities.	135
8.2	The original distribution of target classes before any filtering or merging was done. . . .	136
8.3	Distribution of target classes after removing classes that are semantically broad or lack data.	137
8.4	Density functions of target classes as a function of the hour of day in each country. . . .	138
8.5	Proportion of missing data per sensor type.	140
8.6	Mean AUC score comparison for country-specific and multi-country approaches with population-level and hybrid models. MC: Multi-Country; w/o DS: without downsampling; w/ DS: with downsampling.	143
8.7	Mean AUC scores obtained in the country-agnostic approach with population-level models.	144
8.8	Mean AUC scores obtained in the country-agnostic approach with hybrid models. . . .	144
9.1	Base architecture for UDA with features from multimodal sensors, encoder, domain and target classifier/regressor, and gradient reversal layer.	162
9.2	Modification to the base architecture to have multiple branches that concatenate to create a feature embedding.	162
9.3	Using different λ for branches depending on the average distribution shift of features in the branch. When there is little to no shift, $\lambda \approx 0$ (green).	162
9.4	Average Cohen's-d Values for Modalities. Italy is the Source Domain.	165
9.5	Average Cohen's-d Values for Modalities. Mongolia is the Source Domain.	165
9.6	Cohen's-d Values Italy and India, Sorted in the Descending Order. Modalities Marked in Different Colors.	165
9.7	Average Cohen's-d Values for Modalities. EarBuds is the Source Domain	166
9.8	Cohen's-d Values for EarBuds and Empatica, Sorted in the Descending Order. Modalities Marked in Different Colors.	166
9.9	Cohen's-d Values for EarBuds and Muse, Sorted in the Descending Order. Modalities Marked in Different Colors.	166
9.10	Inference results for various α values.	172
A.1	Averaged threshold distribution difference between target users and three threshold value distributions of MEX dataset	184

List of Tables

2.1	Summary of Phases Including Workshop Participation and Number of Recruited Volunteers	14
2.2	Summary of the 56 features used in the analysis. Feature Group describes the type of features, and 'Examples' are some feature names. The number in parenthesis next to categorical indicates the number of categories in categorical features.	15
2.3	Participants of the mobile sensing data collection (countries named in alphabetical order).	18
2.4	Summary of 105 features extracted from sensing data, aggregated around activity self-reports using a time window. A detailed description of sensing modalities is provided in Appendix A.	19
3.1	Summary of features extracted from mobile sensors (134). Sensor data are aggregated for every 10-minute time slot from 8 pm to 4 am. For all the given features, average, minimum, and maximum were calculated during the matching phase, resulting in 402 sensing features for each alcohol consumption event.	28
3.2	Summary of social contexts in the final dataset.	30
3.3	Pearson Correlation Co-efficient (PCC) and Point-Biserial Correlation Co-efficient (PBCC) for Sensor Features and Social Contexts (two-class and three-class). With the top 5 features for each Social Context are included in the table. p-values are denoted with the following notation: p-value $\leq 10^{-4}$:****; p-value $\leq 10^{-3}$:***	33
3.4	t-statistic (T) (p-value $\leq 10^{-4}$:****; p-value $\leq 10^{-3}$:***; p-value $\leq 10^{-2}$:**), and Cohen's-d (C) with 95% confidence intervals (* if confidence interval include zero). Top five features are shown in decreasing order.	34
3.5	Mean (\bar{A}) and Standard Deviation (A_σ) of inference accuracies and the mean area under the curve of the receiver operator characteristic curve (AUC), calculated from 10 iterations, using five different models, for two-class and three-class tasks, with attributes such as family, friends/colleagues, spouse/partner, and alone. Results are presented as: \bar{A} (A_σ), AUC	35
3.6	Social Context Inference accuracy breakdown for sensor type based feature groups and feature group combinations using Random Forest classifiers. Both the mean (\bar{A}) and standard deviation (A_σ) of accuracies from cross-validation are reported in addition to the mean area under the curve (AUC) from receiver operating characteristics graph (ROC)	37
4.1	Pearson and Point-Biserial correlation analysis for some self-report features and food consumption level.	55

List of Tables

4.2 Comparative statistics of 20 features across classes "overeating" and "undereating" (OverUnder) and "overeating" and "as usual" (OverUsual): t-statistic, p-value (* if $p > 0.01$, ** if $p > 0.1$, *** if $p > 0.5$), and Cohen's-d with 95% confidence intervals. Features are sorted based on the decreasing order of t-statistics. 55

4.3 Three-class food consumption inference (overeating, undereating, as usual) accuracy, precision, and recall obtained with a random forest classifier (RF) and a neural network (NN) using different feature group combinations. 58

4.4 Summary of the Qualitative Analysis. 61

5.1 Summary of Eating Detection Approaches in Mobile Sensing. LB: Lab-Based and IW: In-The-Wild experiment. 71

5.2 Averaged F1-score (F1) and AUC of 58 users, calculated using four different models, for the BASE of eating event detection task. For ALL-PCA, the number in square brackets (e.g. [c=6], etc.) indicates the number of principal components used for the inference. For ALL-FS, the notation in square brackets (e.g., [F3], [F7], etc.) indicates the name of the feature group. More details about feature groups, including the list of features in each group, are given in Appendix (Table A.1). 77

5.3 Averaged F1-score (F1) and AUC calculated using random forest classifiers, for BASE, PERS1, and PERS2 of the eating event detection task of 58 users. For ALL-FS, the notation in square brackets (i.e. [F1]) indicates the name of the feature group. More details about the feature group, including the list of features in the group are given in the Appendix (Table A.1). 77

5.4 Personalized eating event detection accuracy breakdown for random five users in PERS2 with ALL-FS. F1-score (F1) and AUC are shown. The top performing feature group, and modalities included in the feature group are shown with \checkmark . F1-score Bump indicates the F1-score of BASE (Blue), increase in F1-score from BASE to PERS1 (Green), from PERS1 to PERS2 (Orange) for each user. 78

6.1 Terminology and description regarding different model types and Mood-while-Eating. The degree of Personalization increases when going from Population-Level to User-Level. 86

6.2 Mean (\bar{A}) and Standard Deviation (A_σ) of inference accuracies, calculated using five models for the PLM and HM of mood inference task: \bar{A} (A_σ), F1. 97

7.1 Terminology and description regarding different model types and approaches. 108

7.2 t-statistic (TS) (p -value > 0.05 : *) and Cohen's-d (CD) (all features reported here had 95% confidence intervals not overlapping with zero) for positive, neutral, and negative moods for each country. 115

7.3 Country-Specific and Multi-Country results with PLM and HM: Mean (\bar{S}) and Standard Deviation (S_σ) AUC scores computed from ten iterations. Results are presented as $\bar{S}(S_\sigma)$, where S is AUC. 118

7.4 Country-Agnostic I PLM & HM: Two-Class Inference – Mean (\bar{S}) and Standard Deviation (S_σ) of AUC scores obtained by testing each Country-Specific model (rows) on a new country. Results are presented as $\bar{S}(S_\sigma)$, where S is AUC score. Aggregate of the reported population-level results and results from hybrid models indicated under 'Aggregate'. . . 119

7.5 Country-Agnostic I PLM & HM: Three-Class Inference – Mean (\bar{S}) and Standard Deviation (S_σ) of AUC scores obtained by testing each Country-Specific model (rows) on a new country. Results are presented as $\bar{S}(S_\sigma)$, where S is AUC score. Aggregate of the reported population-level results and results from hybrid models indicated under 'Aggregate'. . . 119

7.6	Country-Agnostic II PLM: Mean (\bar{S}) and Standard Deviation (S_σ) of AUC scores obtained by testing each a seven-country model on data from a new country. Results are presented as $\bar{S}(S_\sigma)$, where S is the AUC.	120
7.7	Multi-Country and Continent-Specific with PLM and HM: Mean (\bar{S}) and Standard Deviation (S_σ) of F1-scores and AUC scores obtained by testing the "worldwide" model. Results are presented as $\bar{S}(S_\sigma)$, where S is any of the two metrics.	122
8.1	A summary of participants of the data collection. Countries are sorted based on the number of participants.	136
8.2	ANOVA F-values (F) with p-value < 0.05 for each target activity and each country. The best feature is the first in the list. Comparing F-values are only valid locally within the same activity and country.	138
8.3	Mean (\bar{S}) and Standard Deviation (S_σ) of inference F1-scores, and AUC scores computed from ten iterations using three different models (and two baselines) for each country separately. Results are presented as $\bar{S}(S_\sigma)$, where S is any of the two metrics.	142
8.4	Mean (\bar{S}) and Standard Deviation (S_σ) of F1-scores and AUC scores were obtained by testing each Country-Agnostic model (trained in four countries) on data from a new country. Results are presented as $\bar{S}(S_\sigma)$, where S is any of the two metrics.	146
9.1	Summary of datasets, source and target domains, modalities used, and the performed inferences. C stands for classification and R stands for regression.	160
9.2	Results for classification tasks. Results are presented as average AUC scores (higher the better). TL refers to transfer learning where labelled target domain data are available.	169
9.3	Results for regression tasks. Results are presented as mean absolute errors (MAE) (the lower the better).	170
A.1	Feature Groups Used in Different Inference Models	183
A.2	Summary of the features used in the analysis.	183
A.3	Summary of Mobile Sensing Features Extracted from Smartphone Sensors	187

1 Introduction

1.1 Research Context

Over the past decade, smartphones have undergone rapid evolution, becoming pervasive devices due to advancements in various fields, including hardware (such as CPU and GPU) [128], deep learning [541, 219], telecommunication [541], and human-computer interaction [419, 103]. These technological progressions, coupled with their associated benefits, have firmly entrenched smartphones as integral components in people's lives. A study indicates that smartphone adoption among young adults aged 18 to 29 years in the United States is 96% [344], while further analyses shed light on how smartphones influence human behavior, underscoring their profound integration into the lives of millions [479]. Smartphones offer enhanced user-friendliness and interactivity while possessing the capability to gather and process substantial volumes of passive contextual data in real-time [415]. Furthermore, the existence of app distribution platforms such as the Google Play Store [186] and Apple App Store [21] has facilitated the seamless distribution of smartphone applications by both developers and researchers, reaching millions of users worldwide.

Mobile sensing refers to the utilization of the various built-in sensors in mobile devices such as smartphones and wearables, including accelerometers, compasses, gyroscopes, GPS, microphones, and cameras, to collect data for a wide range of applications [268, 325]. The use of mobile sensing for behavior and health monitoring gained prominence around two decades ago, as researchers began exploring the use of wearable sensors to monitor behavioral patterns, health conditions, and lifestyles [151, 387, 398]. However, the translation of this research into real-world settings remained limited due to various factors, including the high costs associated with creating wearable devices, prevailing attitudes towards wearable technology, and challenges in distributing such devices to broader populations. As a result, much of the research efforts were confined to controlled laboratory environments. The landscape shifted with the widespread adoption of mobile phones during the 2000s, mitigating several of these impediments. The increasing embrace of mobile phones, particularly among the youth, altered the dynamics and prompted an increase in literature on the potential and applicability of mobile phone sensing for large-scale applications. A seminal contribution in this trajectory is Reality Mining [143], which demonstrated the capacity of smartphone sensing to passively collect contextual data (such as GPS traces, Bluetooth interactions, app usage, and charging events) in real-world contexts involving a substantial number of individuals (100+) over an extended time frame (1 year). Initiatives like UbiFit Garden [105], Nokia-Idiap Mobile Data Challenge [274], and MyExperience [159] further illustrated the capabilities of mobile phones in processing data acquired from both external and internal sensors,

combined with self-reports, to facilitate everyday life behavior analysis.

1.1.1 “In the Wild” vs. “In Lab” Studies

In the realm of behavioral research, studies can be broadly categorized into controlled ‘in lab’ settings and real-world ‘in the wild’ environments [437]. ‘In lab’ studies take place in controlled laboratory settings, providing researchers with precise control over variables and conditions, ensuring experimental rigor and reproducibility [300]. These studies are valuable for investigating specific hypotheses and isolating particular factors. On the other hand, ‘in the wild’ studies occur in real-world, uncontrolled settings, allowing researchers to explore behaviors, interactions, and phenomena as they naturally occur [406, 143]. While ‘in-lab’ studies offer controlled environments, they might not fully capture the complexities and nuances of real-world behaviors. In contrast, ‘in the wild’ studies provide a more ecologically valid understanding of human behaviors but present challenges like variability, unpredictability, and limited control over extraneous factors. This thesis specifically focuses on ‘in the wild’ studies, emphasizing the exploration of behaviors and interactions in real-world settings. While ‘in the wild’ studies provide a more holistic and authentic understanding of behaviors, they also introduce significant challenges due to the uncontrolled and dynamic nature of the environment. Factors such as diverse contexts, varying user behaviors, and technological limitations can impact data collection and analysis in complex ways.

1.1.2 Passive Sensing and Self-Reports in “In the Wild” Studies

The main sources of data considered in the thesis are passive sensing and self-report data collected using mobile applications. These two data types are discussed below.

Passive Sensing

Passive sensing, within the realm of mobile and wearable computing, refers to the unobtrusive and continuous collection of data from users’ devices without requiring active user input or explicit interactions [268]. It involves the extraction of information from various sensors embedded in smartphones, smartwatches, and other devices, capturing users’ behaviors, activities, and environmental context as they go about their daily routines [325]. This approach offers several advantages, including minimizing user burden and avoiding disruptions to natural behaviors, thereby providing a more holistic and accurate representation of individuals’ experiences and interactions with their surroundings. Passive sensing leverages data from sensors in devices to infer various aspects of users’ everyday lives, such as physical activity, location, social interactions, sleep patterns, and even emotional states [268, 325, 492]. This rich stream of data enables researchers and developers to gain insights into human behaviors and well-being, paving the way for the design of personalized applications, interventions, and services that cater to users’ needs in a minimally intrusive manner.

In our prior work [325], we developed a taxonomy for passive sensing modalities, separating them into continuous and interaction sensing modalities, as discussed below. Throughout many chapters of this thesis, we use data from both these modalities for training machine learning models. Hence, we use the term “multimodal mobile sensing” to refer to both continuous and interaction sensing modalities.

Continuous Sensing Modalities. These modalities involve the unintrusive collection of user data without requiring explicit interactions or system events. These modalities utilize sensors embedded

in smartphones to provide continuous data streams, with sampling rates and logging frequencies determined based on research needs and performance considerations. The *accelerometer* and *gyroscope* data are widely used modalities, and they capture the movement patterns of individuals by leveraging the common practice of keeping smartphones in pockets, bags, or hands. Studies have employed three main approaches: utilizing raw accelerometer data for statistical analysis [55, 454, 30], employing external APIs like the Google Activity Recognition API [183] for activity inference [545, 150], and developing custom activity recognition algorithms [417, 544]. The *proximity* sensor detects the presence of objects near the phone's screen and has been used to determine phone placement or presence in pockets [30]. *Location* sensing, perhaps the most prevalent modality, has been used to derive various behavioral patterns such as travel distances, time spent at specific locations, and points of interest [77, 30, 418]. *Ambient light* sensors have been employed in studies to measure lighting conditions, potentially influencing mental well-being [305, 30, 545, 288]. The *audio* sensor captures soundscapes and conversations, with applications ranging from emotion inference to sleep pattern detection [300, 418, 468, 544]. *WiFi* and *Bluetooth* data are predominantly used for understanding context, such as determining physical proximity and co-location [454, 305, 30, 545].

Interaction Sensing Modalities. Interaction sensing modalities do not rely on physical sensors, yet capture valuable data about users' interactions with smartphones, hence capturing users' phone usage behavior in different situations [325]. The hypothesis is that phone usage behavior is closely tied to other everyday life activities, especially in young adults. Hence, the expectation is that phone usage behavior would capture cues regarding more complex daily life behaviors [359, 4]. These modalities are software-based and exploit events triggered within the smartphone system. The informativeness of these methods depends on users' phone usage patterns. Examples include phone calls, messages, app usage, browsing history, calendar activities, typing events, touch events, and lock/unlock events. These data have often been harnessed as direct proxies for either behavioral or socio-psychological traits in phone users. *Phone call* and *message* events have been extensively used in studies to capture virtual sensing modalities related to the socio-psychological traits of young adults. Research has explored their connections to mental well-being [285], and even inferred behaviors like drinking episodes [454]. Various features, such as the number of calls, messages, call duration, and message length, have been derived from these modalities. These features have been linked to stress, emotions, and drinking behaviors, showcasing their potential as proxies for socio-psychological traits [468, 453, 60, 61]. *App usage* and *browsing history* offer insights into users' behavioral patterns. Categorizing smartphone apps by type, and considering app launch events and usage duration, provides a fine-grained understanding of users' daily routines and circadian rhythms [454, 136, 135, 359]. Additionally, browsing history can be leveraged to generate meaningful features [454, 285]. Phone usage events, including *screen on/off times*, *typing dynamics*, *touch events*, and *lock/unlock actions*, have shown strong correlations with behavioral and socio-psychological traits [5, 453, 30, 454, 545]. For example, the screen on/off times have been linked to stress levels and phone usage behavior [453], with features such as screen unlocks per minute differing between drinking and non-drinking episodes. In summary, interaction sensing modalities have frequently been employed to generate features as proxies for psychological and behavioral traits.

Self-Reports

The use of self-reports plays a pivotal role in in the wild mobile sensing studies to capture data regarding various aspects of everyday life [325, 393]. Often, in many studies, self-reported aspects have been considered as ground truth when training machine learning models. These data can be categorized into various *types* and *trigger contexts*, shedding light on the diverse approaches researchers take to

capture data.

Types of Self-Reports. Structured questionnaires and Ecological Momentary Assessments (EMA) are the most prevalent methods employed to collect explicit self-report data [468, 454, 55, 357, 475, 544]. These techniques involve posing questions to users, either about their current state or past experiences, which range from mood and emotions to eating and sleeping behaviors. Researchers often design questionnaires based on established models and psychological principles to ensure the robustness of the data collected [325]. In addition, photos and videos have also been used in the past as means of self-reporting [454, 469]

Trigger Contexts of Self-Reports. Self-reporting can be initiated in various contexts. The in-situ approach prompts users to report on their current situation or feelings, offering accurate real-time insights [285, 5, 544, 55]. This method has been adopted to capture mood, alertness, eating habits, and more. In contrast, retrospective reporting asks users to recollect past experiences, such as their mood earlier in the day or what they ate for lunch [30, 357, 453, 67]. This method trades real-time accuracy for convenience and allows for a more comprehensive view of behaviors. Some studies also encourage users to self-initiate reporting, either in an in-situ or retrospective manner, with or without reminders. Reminders, often in the form of push notifications [454, 468], are used to nudge users to provide self-reports, enhancing compliance and participation. However, the frequency and timing of reminders must be carefully managed to strike a balance between engagement and user comfort.

In conclusion, self-reporting plays a pivotal role in mobile sensing studies. While the ultimate goal is to gather as few (or none) self-reports from users as possible, achieving this requires the study and development of various algorithms that rely on self-reports. Researchers employ a range of techniques, encompassing structured questionnaires and multimedia integration, to comprehensively capture behaviors, emotions, and contexts. The choice of trigger context, whether in situ, retrospective, or self-initiated, influences the granularity and accuracy of the collected data. By harnessing these techniques, researchers can gain invaluable insights into the intricate interplay among behaviors, contexts, and well-being. In this thesis, we explore various types of self-reports, and discuss their effect on model performance.

1.2 Motivation

By combining multimodal mobile sensing data and self-reports in the context of in the wild studies, this thesis has two primary motivations. First, to deepen the understanding of how multimodal mobile sensing and machine learning could be used to understand eating and drinking behavior and its interplay with mood, context, and everyday life activities. This is illustrated in Figure 1.1. Second, from a machine learning perspective, to explore the use of generic/one-size-fits models, generalized models, and personalized models for inferences, focusing on generalization and domain adaptation of models. This is summarized in Figure 1.2.

1.2.1 Understanding the Interplay between Eating and Drinking Behavior, Mood, Daily Activities, and Context

Individuals across various age groups exhibit distinct lifestyles, behavioral patterns, cognitive processes, and biological characteristics [433, 102]. Young adults, typically aged between 16 and 35 years, navigate unique life circumstances in comparison to older generations, manifesting in divergent activities,

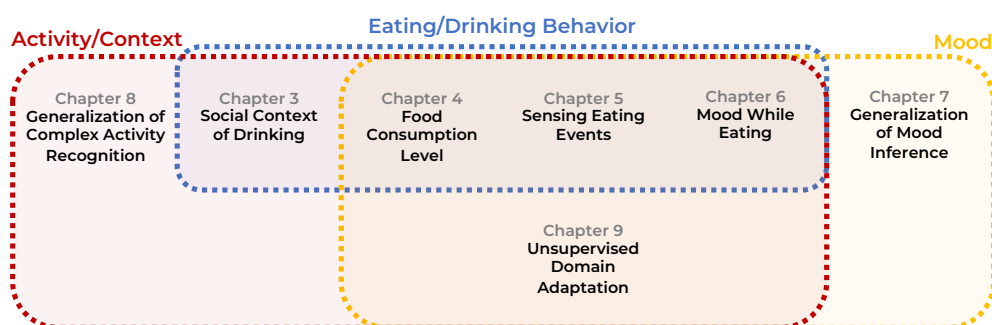


Figure 1.1: High-level overview of application areas covered in the thesis. The three main topics are eating or drinking behavior, activity or content, and mood. Most chapters focus on either two or three of the above application areas.

social interactions, dietary choices, and even their physical and mental health conditions [390, 218, 370, 500, 497, 124]. This demographic group encompasses individuals engaged in pursuits such as education, early career stages, the initial years of marriage, unemployment, or combinations thereof. In light of this life stage, it is recognized that negative mood, anxiety, obesity, alcohol/smoking/drug dependencies, and unhealthy eating habits are prevalent among young adults [203, 204, 395, 497], with the underlying reasons for these issues differing from those affecting individuals in other age groups. For instance, the negative mood among undergraduate students can often be attributed to the demanding academic workload, whereas a person aged 40-50 may experience such conditions due to familial or occupational pressures [220]. Hence, monitoring and inferring such everyday life aspects through the use of multimodal smartphone sensors and machine learning is gaining traction across both clinical and non-clinical research, especially targeting young adults among whom the device usage is high. In recent work, multimodal mobile sensing has been used to model a plethora of everyday life aspects of predominantly young adult populations, focusing on mood [285, 468], stress [453, 74], brain activity [375], alcohol drinking behavior [454, 30], eating behavior [354, 355, 55], and daily activities [494] among many other aspects. However, beyond the current inferences that have been examined, there is a need for deeper understanding regarding eating and drinking behavior and its interplay with mood, social context, and daily activities with the use of mobile sensing, which Chapters 3-6 of this thesis focus on.

1.2.2 Generalization and Personalization of Models

Multimodal mobile sensing for monitoring everyday life behavior offers the potential to significantly impact individuals' well-being and quality of life [325]. However, the deployment of such sensing-based models in the wild comes with challenges that need to be addressed because of the heterogeneity of devices, geographical differences, and individual behaviors. Hence, more often than not, even for simple inferences, generic/one-size-fits-all models do not provide adequate performance, and hence, personalized models are needed [329]. However, model personalization is hindered by the inherent difficulty in obtaining sufficient labeled data from individual users, necessitating the development of techniques that can effectively personalize models within the constraints of limited label availability. Before personalization, which is adapting the model to an individual, an intermediate step could be achieving model generalization to a specific domain or context (e.g., if a model is trained with data collected in the United Kingdom, rather than directly personalizing the model to a user in India,

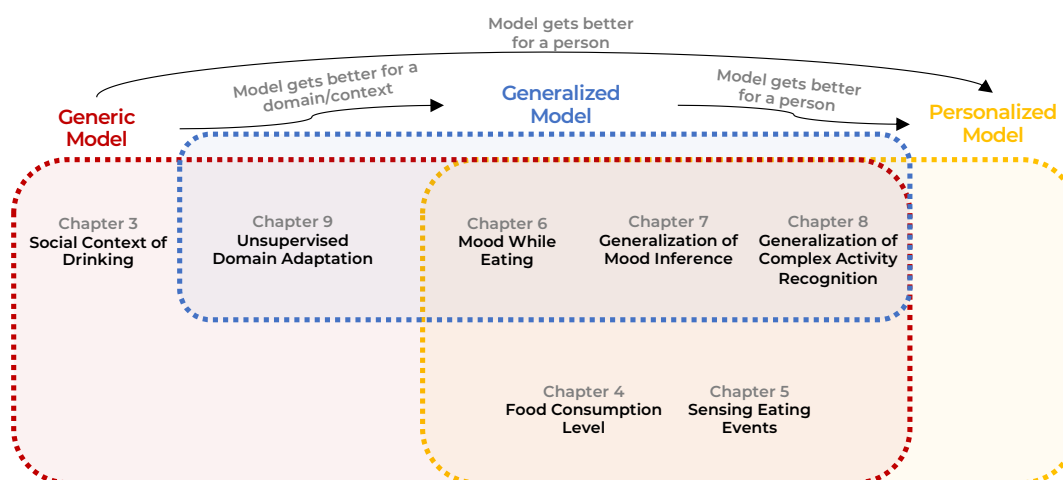


Figure 1.2: High-level overview of machine learning models covered in the thesis. The three main categorizations are chapters examining generic models, generalized models, and personalized models. Most chapters focus on either two or three of the above model types.

the model could be adapted to India for better generalization, which then could be personalized). However, achieving model generalization poses its own set of challenges, given the high heterogeneity observed in user behaviors across various situated contexts, such as different countries and time frames [247, 569]. Previous research in mobile sensing has often overlooked the crucial aspect of model generalization due to limitations related to available datasets and machine learning methodologies [325, 569]. Moreover, conducting longitudinal passive and multimodal sensing studies to gather diverse datasets is a costly endeavor, resulting in previous research relying heavily on surveys and sensor data from specific populations within limited time frames. Often in academic research, these populations consist of individuals associated with the researchers' specific universities, limiting the scope of data collection and model training to a particular geographic and temporal context [325, 569]. To ensure practical, real-world implementation of mobile sensing models, it is imperative to evaluate their performance across diverse datasets, ensuring their generalizability. In case of a lack of generalizability, methods need to be developed to achieve better generalization. Chapters 4-9 of this thesis focus on these aspects from different perspectives.

In summary, this thesis has two primary objectives. First, as application areas, this thesis delves into understanding the eating and drinking behavior of young adults, and its interplay with mood, daily activities, and context by using multimodal mobile sensing data and machine learning (Figure 1.1). Then, from a machine learning standpoint, this thesis delves into using generic one-size-fits-all models, generalized models that cater well to specific populations or contexts, and personalized models that cater well to individuals (Figure 1.2). Hence, the thesis looks into novel inferences regarding everyday life while also examining how deployment challenges such as model generalization and personalization affect those inferences.

1.3 Summary of Contributions and Dissertation Outline

The contributions of this thesis, together with the corresponding chapters, are the following:

- **Datasets (Chapter 2):** We contributed to two mobile sensing data collection efforts that aimed to understand the everyday life behavior of young adults. These data collections were done together with partners from over ten countries in the context of the European Horizon 2020 WeNet project. The first dataset (MEX) was collected in 2019, where over 80 college students in Mexico used a mobile application that collected multimodal sensing data passively, together with self-reports regarding their eating behavior, mood, and context. This study was done for approximately 60 days. The second multi-country dataset (MUL) was collected during the 2020-2021 period, where over 670 college students in eight countries (China, India, Mongolia, Italy, UK, Denmark, Mexico, and Paraguay) used a mobile application that again collected passive sensing data from multiple sensors, while participants provided ground truth regarding mood, context, and everyday life activities (including eating and drinking behavior) with self-reports. The study was done for approximately 30 days. These two datasets provide the basis for most of the analyses done in this thesis. The material of this contribution was originally published in [330] and [324].
- **Social Context of Drinking (Chapter 3):** Prior research links alcohol drinking behavior with social factors like relationship type and group size. Despite its significance, the impact of social context on young adults' drinking behavior in smartphone sensing studies remains underexplored. We investigated weekend nightlife drinking in 241 young adults in Switzerland, over three months, using self-reports and passive smartphone data, in the context of the SNSF Dusk2Dawn project. The used dataset is described in [454]. Analyses revealed accelerometer, location, app usage, bluetooth, and proximity data as indicators of social contexts. We assessed seven social context inference tasks, achieving 75%-86% accuracy in four binary and three ternary classifications. These findings could be used to support alcohol consumption interventions and possibly reduce reliance on self-reports for drink tracking in health studies. The material of this contribution was originally published in [327].
- **Self-Perceived Food Consumption Level (Chapter 4):** While food consumption characterization is well-established in nutrition and psychology, the potential of smartphone sensing in enhancing mobile food diaries remains underutilized. With the MEX dataset, we introduced a novel ubicomp task, inferring self-perceived food consumption (overeating, undereating, usual) with 87.81% accuracy in a 3-class classification using passive smartphone sensing and self-reports. Encouragingly, 83.49% accuracy was achieved for the same task using only smartphone sensing data and meal times. This has the potential to support context-aware and user-friendly mobile food diary apps. We also explored personalization techniques that allow for better performance. The material of this contribution was originally published in [330].
- **Sensing Eating Events (Chapter 5):** While previous work has addressed eating event detection using wearables, utilizing smartphones alone for this purpose remains an open challenge. This chapter introduced a framework for inferring eating versus non-eating events through passive smartphone sensing, evaluated on 58 college students in the MEX dataset. We showed that factors such as time of day, screen on and off events, activity data from accelerometer data, app usage, and location contribute to distinguishing eating events. Using population-level machine learning models, eating events can be inferred with an area under the receiver operating characteristic curve (AUC) of 0.65, which can be further enhanced to 0.81 with user-level and hybrid models, employing personalization techniques. Notably, users exhibit diverse behavioral patterns around eating episodes, necessitating specific feature groups for fully personalized models. The material of this contribution was originally published in [323].
- **Mood while Eating (Chapter 6):** While phone sensor data have been separately used to investigate eating behavior and mood within mobile food diaries and health apps, there is a (1) a lack of studies

on the generalization of mood inference models from passive sensor data to specific contexts like eating; (2) a lack of research examining the intersection of mood and eating using sensor data, and (3) insufficient exploration of model personalization techniques in constrained label settings, common in mood inference. This chapter addresses these gaps by analyzing eating behavior and mood in the MEX and MUL datasets, incorporating passive smartphone sensing and self-report data. Results reveal the decreased performance of generic mood inference models in specific contexts, such as eating, indicating a lack of model generalization. We also introduced a novel community-based personalization approach, achieving mood-while-eating inference accuracy of 80.7%, outperforming traditional methods. The material of this contribution is under review.

- **Generalization of Mood Inference Models (Chapter 7):** Even though mood inference with mobile sensing data has been studied in prior work, limited attention has been given to the cross-country generalization of models. In this chapter, we examined the MUL dataset, collected across eight countries, to explore the impact of geographical diversity on mood inference models. We evaluated approaches tailored to individual countries, continents, cross-country scenarios, and hybrid models in two mood inference tasks. We showed that partially personalized country-specific models performed the best yielding AUC scores of the range 0.78-0.98 for two-class (negative vs. positive valence) and 0.76-0.94 for three-class (negative vs. neutral vs. positive valence) mood inference. Hence, our findings highlight the superior performance of partially personalized, country-specific models and the challenges faced by country-agnostic models, shedding light on model generalization issues across countries. The material of this contribution was originally published in [324].
- **Generalization of Complex Activity Recognition (Chapter 8):** Following a similar analysis strategy to Chapter 7, we analyzed multimodal mobile sensing data and self-reports from 637 college students in five countries (a subset of the MUL dataset) and introduced a challenging 12-class daily activity recognition task. While the generic multi-country approach achieved an AUC of 0.70, country-specific models outperformed it with AUC scores between 0.79 and 0.89. Our results highlighted the importance of diversity-aware methods to advance smartphone sensing research across countries. We also emphasized the lack of model generalization across countries without labeled data from individuals in target domains. The material of this contribution was originally published in [24].
- **Unsupervised Domain Adaptation for Multimodal Sensor Data (Chapter 9):** The challenge of distribution shift hinders real-world deployment of mobile sensing models, leading to lack of generalization as found in Chapters 7 and 8. While a solution for this—domain adaptation has been addressed in computer vision and natural language processing, mobile sensing research has rarely examined unsupervised domain adaptation for multimodal sensor data. To bridge this gap, we introduced an unsupervised domain adaptation method for multimodal mobile sensing using multi-branch adversarial training. Through extensive experiments involving two datasets (MUL and WEEE [166, 18]), three inference tasks, and 14 source-target domain pairs (regression and classification), we demonstrate its effectiveness in unseen domains. Compared to the direct deployment of source-trained models, our approach yields up to 12% AUC improvement in classification tasks and up to 0.13 mean absolute error (MAE) reduction in regression tasks. The material of this contribution is under review.

1.4 Publications

This dissertation is a compilation of six journal publications and one conference publication. Co-primary authorship is marked with * mark.

Journal Papers

- **Lakmal Meegahapola**, Florian Labhart, Thanh-Trung Phan, and Daniel Gatica-Perez. "Examining the Social Context of Alcohol Drinking in Young Adults with Smartphone Sensing". Proceeding of the ACM on Interactive, Mobile, Wearable, and Ubiquitous Technologies (IMWUT/UbiComp). 2021. ACM, New York, USA.
- **Lakmal Meegahapola**, Salvador Ruiz-Correa, Viridiana del Carmen Robledo-Valero, Emilio Ernesto Hernandez-Huerfano, Leonardo Alvarez-Rivera, Ronald Chenu-Abente, and Daniel Gatica-Perez. "One More Bite? Inferring Food Consumption Level of College Students Using Smartphone Sensing and Self-Reports". Proceeding of the ACM on Interactive, Mobile, Wearable, and Ubiquitous Technologies (IMWUT/UbiComp). 2021. ACM, New York, USA.
- **Lakmal Meegahapola**, William Droz, Daniel Gatica-Perez, et al. "Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries". Proceeding of the ACM on Interactive, Mobile, Wearable, and Ubiquitous Technologies (IMWUT/UbiComp). 2023. ACM, New York, USA.
- **Lakmal Meegahapola**, Hamza Hassoune, Daniel Gatica-Perez. "M3BAT: Unsupervised Domain Adaptation for Multimodal Mobile Sensing with Multi-Branch Adversarial Training". Under Review. 2023.
- Wageesha Bangamuarachchi*, Anju Chamantha*, **Lakmal Meegahapola***, Salvador Ruiz-Correa, Indika Perera and Daniel Gatica-Perez. "Sensing Eating Events in Context: A Smartphone-Only Approach". IEEE Access, vol. 10, 2022.
- Wageesha Bangamuarachchi*, Anju Chamantha*, **Lakmal Meegahapola***, Haeun Kim, Salvador Ruiz-Correa, Indika Perera and Daniel Gatica-Perez. "Inferring Mood-While-Eating with Smartphone Sensing and Community-Based Model Personalization". Under Review. 2023.

Conference Papers

- Karim Assi*, **Lakmal Meegahapola***, William Droz, Daniel Gatica-Perez, et al. "Complex Daily Activities, Country-Level Diversity, and Smartphone Sensing: A Study in Denmark, Italy, Mongolia, Paraguay, and UK". Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '23). 2023. ACM, New York, USA.

In addition to the papers mentioned above, I wrote three additional papers as first author during my doctoral studies (not included in the dissertation for space reasons):

- **Lakmal Meegahapola**, Salvador Ruiz-Correa, and Daniel Gatica-Perez. "Alone or With Others? Understanding Eating Episodes of College Students with Mobile Sensing". In Proceedings of the 19th International Conference on Mobile and Ubiquitous Multimedia (MUM '20). 2020. ACM, New York, USA.
- **Lakmal Meegahapola**, Salvador Ruiz-Correa, and Daniel Gatica-Perez. "Protecting Mobile Food Diaries from Getting Too Personal". In Proceedings of the 19th International Conference on Mobile and Ubiquitous Multimedia (MUM '20). 2020. ACM, New York, USA.
- **Lakmal Meegahapola** and Daniel Gatica-Perez, "Smartphone Sensing for the Well-Being of Young Adults: A Review". IEEE Access, vol. 9. 2021.

Finally, I contributed to several additional papers as co-author.

- Laura Schelenz, Ivano Bison, Matteo Busso, Amalia de Götzen, Daniel Gatica-Perez, Fausto Giunchiglia, **Lakmal Meegahapola**, and Salvador Ruiz-Correa. "The Theory, Practice, and Ethical Challenges of Designing a Diversity-Aware Platform for Social Relations". In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '21). 2021. ACM, New York, USA.
- Florian Labhart, Skanda Muralidhar*, Benoit Massé*, **Lakmal Meegahapola***, Emmanuel Kuntsche, Daniel Gatica-Perez. "Ten seconds of my nights: Exploring methods to measure brightness, loudness and attendance and their associations with alcohol use from video clips". PLOS ONE 16(4). 2021.
- Emma Bouton-Bessac, **Lakmal Meegahapola**, Daniel Gatica-Perez. "Your Day in Your Pocket: Complex Activity Recognition from Smartphone Accelerometers". In Proceedings of the 16th EAI Pervasive Computing Technologies for Healthcare (PervasiveHealth). Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 488. Springer, Cham. 2022.
- Alexandre Nanchen, **Lakmal Meegahapola**, William Droz, and Daniel Gatica-Perez. "Keep Sensors in Check: Disentangling Country-Level Generalization Issues in Mobile Sensor-Based Models with Diversity Scores". In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '23). 2023. ACM, New York, USA.
- Nathan Kammoun, **Lakmal Meegahapola**, and Daniel Gatica-Perez. "Understanding the Social Context of Eating with Multimodal Smartphone Sensing: The Role of Country Diversity". In Proceedings of the 25th ACM on International Conference on Multimodal Interaction, (ICMI '23). 2023. ACM, NY, USA.

2 Datasets

The goal of this chapter is to introduce two datasets that were used across many chapters of the thesis. These datasets were collected during my Ph.D. in the context of the European Horizon 2020 project WeNet, with partners in multiple countries. These partners included computer scientists, social scientists, ethicists, design researchers, and software engineers.

2.1 Mexico Dataset (MEX)

The primary objective of this data collection was to investigate links between food consumption, mood, and context using features derived from multimodal smartphone sensing in a cohort of college students in Mexico. The material of this contribution was originally published in [330].

2.1.1 Mobile Application

We used a native Android mobile application called i-Log to collect data from participants [585]. The app was developed at the University of Trento, by using Java, and data were initially stored in a SQLite database on the smartphone. Moreover, the system used Google Firebase as a notification broker to send push notifications. When the phone is connected to a WiFi network and the phone has sufficient battery capacity, anonymized data were uploaded to the Cassandra DB database in secure servers, hence freeing up the internal storage. The app had three main components: (a) a push notification system to prompt users to complete questionnaires, (b) mobile surveys to record self-reports, and (c) a passive smartphone sensing component to log sensor data.

Push Notification System

Given the nature of our study, we decided to use retrospective questionnaires to obtain self-reports from users. A notification strategy is important to get ground truth at the correct time. Hence, with the expectation of users retrospectively reporting their last food intake, we sent notifications to users three times per day, prompting them to fill in the food intake questionnaire. Push notifications were sent during the time slots 10.00 to 11.00 am, 3.00 to 4.00 pm, and 9.00 to 10.00 pm, which was chosen after considering the eating behavior of participants (note that lunch and dinner times in Mexico are later than those in the USA or most of Europe). Further, our expectation was not to collect data regarding all food intakes but to collect as accurate data as possible regarding the three eating periods per day.

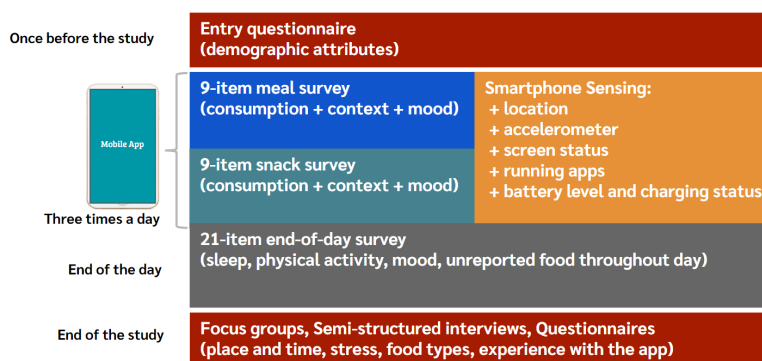


Figure 2.1: Block Diagram of Data Collection

Food Intake Questionnaires

When users clicked a notification, the app opened, and they were asked to indicate whether they wished to report a meal or a snack. The 9-item questionnaire was prepared with user experience in mind such that it can be filled within 1-2 minutes. Initially, we asked people about the number of eating episodes they had within the last four hours. Then we asked users to focus on their last food intake and answer eight more questions that collected data, including how long ago the eating episode had occurred (1-30 minutes ago, 30-60 minutes ago, etc.), food categories (meat, fish, bread, cereals, dairy, etc.), social context of eating (alone, with a date, with a group of friends, etc.), semantic eating location (home, university restaurant, cafe, etc.), and concurrent activities (reading, socializing, watching TV, etc.). Then, two more questions were asked regarding their state of mind, including general mood (valence) at the time of eating (5-point scale from very negative to very positive) and stress level at the time of eating (5-point scale from very stressed to very calm). These questions were adapted from existing literature used to determine mood and stress level [418, 285]. Moreover, similar to Vartanian et al. [527], the app asked users to indicate their self-perceived food consumption level as (1) significantly less food than usual, (2) slightly less food than usual, (3) about the same as usual, (4) slightly more food than usual, and (5) significantly more food than usual.

Passive Smartphone Sensing

The average amount of time users take to provide self-reports is 6-8 minutes per day. To have a fine-grained understanding of behavioral routines connected to eating, the app had passive smartphone sensing capabilities to collect data regarding users throughout the day. (a) *App Usage* - We collected data regarding app usage once every five seconds. Prior research has indicated that app usage behavior could be associated with user context [572, 473] and internal states [597]; (b) *Accelerometer* - Prior research has shown that accelerometer data can be used to derive activity levels of people [514, 277]. Some studies have also associated these activity levels with well-being-related aspects [418, 468]. The app collected data from the accelerometer with a frequency of 20Hz; (c) *Battery Events* - Whether the phone is plugged into a charging port and to what kind of a port it is connected (USB, alternative current, unknown) provide cues regarding the context of users including the availability of charging facilities nearby and also the general phone usage behavior [30]. Moreover, we collected battery level throughout the day at every 1% change in battery level, and each time the charging status changed; (d) *Screen Events* - The app logged details regarding the status of the screen, including when the screen is turned on or off [5, 30]. These details can be used to estimate the amount of time the phone screen is

on around eating events. This is an important attribute when examining eating behavior because prior work has discussed the adverse consequences of excessive phone usage around eating episodes [503, 489, 413]; *(e) Location* - This sensor has been used in many studies to quantify daily movement [581, 37]. The app logged the location of users with both GPS and phone networks based on the availability of sensors once every 10 minutes.

2.1.2 Data Collection and Pre-Processing

Participant Recruitment, Study Approach, and Ethical Considerations

This study considered participants living in San Luis Potosi City in Mexico, which has 1.2 million inhabitants and is home to diverse student populations. We considered college students from two of the main (private and public) universities in the city. People of this age group are tech savvy, and the smartphone coverage among them is high [345]. In addition, young adults are open to changing their behavior, specifically in relation to their eating and physical activity habits [435, 436, 430]. Considering all the above factors, a campaign was launched in June 2019 to announce a mobile sensing data challenge. Two workshops were held in August and October 2019, where the goals of the study were introduced to potential participants, describing how data collection would be carried out, and how data would be used for research. After a basic screening process, interested participants voluntarily filled out a consent form and entered the study, filling out an entry questionnaire (see Figure 2.1). Then, they installed the app on their smartphones. The average age of study participants was 23.4 years (SD: 3.51), the mean BMI was 24.14 (SD: 4.68), and the cohort had 44% males and 56% females. As an appreciation for participating in the study, all of them were rewarded with a gift pack and a t-shirt. Participants had to meet certain criteria: (a) own an Android smartphone and (b) not have eating disorders like bulimia or anorexia.

The initial phase of the study was run from September 3 to October 9, 2019. The second phase ran between November 19 and December 12, 2019. Hence, we collected data for a total of 60 days from 84 university students. We made sure that the data collection was done during a time period with no major examinations or university-level events that would alter the usual behavior of students. The data summary is described in Table 2.1. Phase I involved students from University 1 (Universidad Autónoma de San Luis Potosi), and Phase II was done at University 2 (Universidad Tangamanga). During the data collection campaign, we sent a total of 7898 notifications to users and got 3895 responses (49.3% response rate), and 3278 of these responses corresponded to fully complete reports, including 1911 meal reports and 1367 snack reports. After the mobile data collection phase, we conducted ten semi-structured interviews, one eight-person focus group, and 64 short questionnaires with open-ended questions. The goal was to understand possible links between food consumption and the variables captured in the app, including places, times, stress, and food types. All material was originally in Spanish, then translated to English for reporting purposes and publications. A detailed description of the experimental procedures of the project was reviewed and approved by the local institutional authorities at IPICYT. The institutional review included approval of a Data Protection Impact Assessment document, a Privacy statement and Participant Consent forms, a Data Processing agreement among project partners, and a Declaration of Commitment agreement.

Table 2.1: Summary of Phases Including Workshop Participation and Number of Recruited Volunteers

Phase	# of Days	# of Participants	# of Recruited Volunteers	Population
I	37	32	29 (90.6%)	University 1
II	23	90	55 (61.1%)	University 2
Total	60	122	84 (68.9%)	-

Pre-Processing the Dataset for Analysis

The goal of our analyses (Chapter 4, Chapter 5, and Chapter 6) were to investigate eating episode-level data. Hence, we chose each food intake self-report as a data point in our dataset. To integrate sensor and survey data, we followed an approach suggested in prior mobile sensing literature [468, 55, 454, 30], where for each event of focus, in this case for each eating episode, passive sensing data would be aggregated using a defined *time window*. We selected a time window of one hour, which would mean that for each food intake event, we aggregate passive sensing data half an hour before and after the event starting time. We chose this time window considering prior research regarding characterizing eating events [55] and from a preliminary analysis regarding food consumption level (explained in Section 4.6). We started the procedure by finding the adjusted eating time because self-reports were done retrospectively. As mentioned in the previous section, we asked users "how long before they had the last meal". Using the answer to this question, we adjusted the timestamp of each food intake report to estimate the actual time of the eating episode. As an example, if the time of the self-report is 2pm, the answer to the question is 30-60 minutes ago (on average $30+60/2 = 45$ minutes ago), and the adjusted time of eating is estimated as 1.15pm (2pm - 45 minutes). Hence, using the one-hour time window, each eating event would be considered a one-hour eating episode. If the adjusted eating time is denoted by T , the time window would be one hour from $T - 30$ minutes to $T + 30$ minutes. Next, we describe how each data modality was processed to associate it with eating episodes. All the derived features are summarized in Table 2.2. An extended description of passive sensing features is provided in Table A.3

Accelerometer: Following an approach similar to [55, 402, 404], for each 10-minute slot of the day, we generated features (aggregated sum of all values and sum of absolute values) using accelerometer value for axes x , y , and z . Then, depending on the adjusted time of an eating event (T), we considered three 10-minute bins before that eating episode ($T-30$ to $T-20$, $T-20$ to $T-10$, and $T-10$ to T), and three 10-minute bins after the start of the eating episode (T to $T+10$, $T+10$ to $T+20$, and $T+20$ to $T+30$). This way of pre-processing led to the creation of 18 features using accelerometer values. We use abbreviations to name the features generated using this methodology: (a) *abs* - calculated using absolute values of the accelerometer data; (b) *bef* - feature is calculated considering data before T , from $T-30$ to T ; and (c) *aft* - feature is calculated considering data after T , from T to $T+30$.

Apps: App usage has been commonly used to understand human behavior in prior work [454, 285, 404]. We selected the ten most frequently used apps in the dataset. Then, during the hour associated with the eating episode, we determined whether each of those apps was used or not, hence resulting in binary values for features in feature group *App*.

Location: Using location traces, we calculated the radius of gyration (a commonly used metric in mobile sensing [581, 37, 77]) within the hour of consideration associated with the eating episode. Moreover, for each user, we generated stay regions throughout the whole day. Hence, using self-report labels (home, university, etc.), we generated labels for passively sensed stay regions of users, and we call that feature as *location* in our analysis. Moreover, for the location feature, we only used location degraded in precision for location privacy reasons (keeping only four decimal points).

Table 2.2: Summary of the 56 features used in the analysis. Feature Group describes the type of features, and 'Examples' are some feature names. The number in parenthesis next to categorical indicates the number of categories in categorical features.

Feature Group - # of Features	Description	Type	Example Features
Activity (ACC) - 18	Features derived using accelerometer	numerical	acc_x , acc_xabs , acc_z bef
App Usage (APP) - 10	Features derived using app usage	categorical(2)	facebook, instagram, etc.
Food Category (FOOD) - 15	Features derived using category of food	categorical(2)	fats & oils, meat, eggs, fish
Battery Events (BAT) - 2	Charging status Battery level of the phone	categorical(2) numerical	charging status battery level
Location (LOC) - 2	Radius of gyration calculated using location Location of the person	numerical(1) categorical(12)	radius of gyration location
Temporal (TIME) - 2	Hour of the day and minute of the day	numerical	hours, minutes
Screen (SCR) - 1	Number of screen on/off events	numerical	screen events
Psychological (PSY) - 2	Derived using mood and stress self-reports	categorical(5)	mood, stress
Context (CON) - 3	Self-reported contextual details about eating	categorical	activity, social, location
Overeating (OVER) - 1	Self-reported food consumption level	categorical(5)	overeating

Screen: Using screen-on/off events in the dataset, we calculated the number of times the screen was turned on during the time slot, similar to prior literature [5, 30].

Battery: Similar to [30], we calculated the average battery level and also whether any charging events were detected during the time of the eating episode. Battery and Screen events are used as proxies for smartphone usage behavior [30, 5].

2.2 Multi-Country Dataset (MUL)

This was an extension of the MEX dataset collection and was done in a larger set of institutes. The mobile app and experimental protocol were improved compared to the previous data collection based on the lessons learned. Hence, we collected passive smartphone sensing and self-report data from participants about their everyday life behavior and well-being. The ultimate goal of this pilot was to study their behavior, including aspects such as activity, social context, eating behavior, and mood from a mobile sensing standpoint and also to consider various diverse aspects that could potentially affect sensing-based inferences (ranging from geographical region and gender to personality and values). The study is summarized in Figure 2.2. The study design consisted of two main components: (a) LimeSurvey component to collect survey responses during pre and mid-study phases and (b) Mobile sensing app to collect sensor data and self-reports. A technical report regarding the study procedure and future plans for dataset access will be made available in [175]. The material of this contribution was originally published in [324].

2.2.1 LimeSurvey Questionnaires

Survey responses were captured from participants with three questionnaires sent to them before and during the pilot at three different times. This was done to ensure that the burden on participants was reasonable. These questionnaires were administered through the LimeSurvey platform [145].

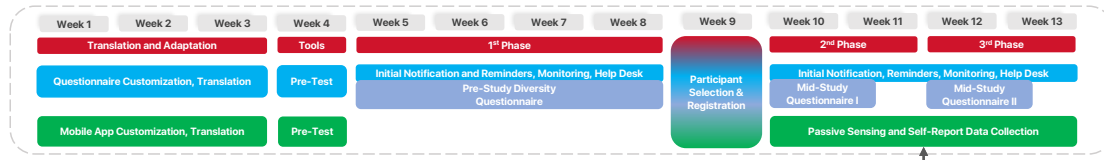


Figure 2.2: High-level overview of the study.

Pre-Study Diversity Questionnaire

The primary objective of this questionnaire was to capture the diversity attributes of participants from different perspectives. As the first step, basic demographic information was captured, including gender, age, sex, degree program, and socioeconomic status. Then, in an attempt to capture the psychosocial profile, the 20-item Big Five Inventory (BFI) [138] and Basic Value Survey (BVS) [187] were administered. Finally, there were several questions regarding social relationships (virtual and real) and cultural consumption that they were interested in.

Mid-Study Questionnaire I and II

The objective of the first questionnaire was to gather more detailed information about personality using the Jungian Scale on Personality Types [233] and Human Values Survey [461]. In addition, questions regarding physical activity and sports, cooking and shopping habits, transport methodologies, and cultural activities were captured. The second questionnaire consisted of the Multiple Intelligences Profiling Questionnaire [512].

2.2.2 Mobile Application

An Android mobile application was used to capture the everyday behavior of participants using short in-situ self-report questions. The app was developed such that data would be stored in an SQLite database on the phone, and later, when the phone is connected to a Wifi network, data would be uploaded to the main server and free up the local phone storage. In addition, the app could send push notifications by using Google Firebase as a notification broker. Hence, the three main components of the application are: (1) a push notification system that would send periodical reminders to participants to fill in self-reports; (2) mobile time diaries to capture self-reports; and (3) a smartphone sensing component to collect passive sensing data from multiple modalities.

Push Notifications

Given the nature of the study and the requirement to capture behavioral and situational data in a particular moment, the app sent reminders for participants to fill in in-situ self-reports regarding their everyday life behavior around 20 times throughout the day. In addition, start-of-the-day and end-of-the-day questionnaires were administered at the beginning and end of the day. When a notification was not clicked, and a participant did not complete the self-report within two hours, the notification expired, and a new notification would be sent later. This allowed us to keep track of participant compliance (e.g., how many self-reports were answered from the total number of notifications sent).

Time Diaries and Start/End-of-the-Day Questionnaires

The start-of-the-day questionnaire was sent to participants at 8 am each day. It only had two questions with five-point Likert scales (very good to very bad): (i) sleep quality and (ii) expectations about the day. The end-of-the-day questionnaire was sent to participants at 10 pm and asked them (a) to rate their day (five-point Likert scale; very good to very bad), (b) if they had any problems during the day (open response), and (c) how did they solve them (open response). The time diary was sent to users once every 30-60 minutes. While this allowed capturing longitudinal behavior granularly, it also introduced user burden. Therefore, the time diary was designed to minimize user burden and reduce completion time. Hence, after several iterations of discussions, only four questions were included in this component: (i) current activity: 34 activities including eating, working, attending a lecture, etc.; (ii) semantic location: 26 categories including home, workplace, university, restaurant, etc.; (iii) social context: 8 categories including alone, with the partner, family member/s, friends, etc.; and (iv) current mood: five-point Likert scale to capture the valence of the circumplex mood model [445] similar to LiKamwa et al. [285], with an emoji-scale.

2.2.3 Data Collection and Pre-Processing

The app collected sensor data from a range of sensors passively. Hence, sensor data included continuous sensing modalities such as accelerometer, gyroscope, ambient light, location, magnetic field, pressure, activity labels generated by the Google activity recognition API, step count, proximity, and available Wi-Fi and bluetooth devices. Interaction sensing modalities included application usage, typing and touch events, on/off screen events, user presence, and battery charging events.

Participant Recruitment, Study Approach, and Ethical Considerations

The primary objective of this study was to capture data from diverse student populations. While many facets of diversity could be captured by experimenting within the same country, it is difficult to study geographical diversity in such a way. Hence, we conducted mobile sensing experiments in eight countries representing Europe, Asia, and Latin America. Details regarding the data collection are mentioned in Table 2.3. According to prior work in mobile sensing, many studies have focused on Europe and North America, but not much research has been conducted in other world regions [325, 402]. Hence, conducting the same study with the same protocol in multiple countries allows us to study different inferences and geographical diversity in a novel sense. The study was conducted in the following phases.

Translation and Adaptation. In this phase, each site received the English version of the questionnaires and the app, including time diaries and the list of sensors to be collected. These tools were evaluated and adapted, in coordination with all the partners, to the specific context (e.g., invitation letters, type and amount of incentives for the participants of the mobile app, privacy and ethics documentation, etc.). Some countries made minimal changes to better adapt the questionnaire to the local situation or academic organization. Concerning the standard scales mentioned above, the translations were completed by a forward translator from the original English version and then validated via panel and back-translation processes by independent translators. In addition, whenever a validated questionnaire translation was available, we used it (e.g., the Big Five traits questionnaire is readily available in several languages). After translation and adaptation, the tools were tested locally. A first test was conducted to check and validate the translations and evaluate the tools' usability. A second test was conducted

Chapter 2. Datasets

Table 2.3: Participants of the mobile sensing data collection (countries named in alphabetical order).

Country	University	Participants	μ Age (σ)	% Women	# Self-Reports
China	Jilin University	41	26.2 (4.2)	51	22,289
Denmark	Aalborg University	24	30.2 (6.3)	58	10,010
India	Amrita Vishwa Vidyapeetham	39	23.7 (3.2)	53	4,233
Italy	University of Trento	240	24.1 (3.3)	58	151,342
Mexico	Instituto Potosino de Investigación Científica y Tecnológica	20	24.1 (5.3)	55	11,662
Mongolia	National University of Mongolia	214	22.0 (3.1)	65	94,006
Paraguay	Universidad Católica "Nuestra Señora de la Asunción"	28	25.3 (5.1)	60	9,744
UK	London School of Economics & Political Science	72	26.6 (5.0)	66	26,688
Total/Mean		678	24.2 (4.2)	58	329,974

by sending the questionnaires to a small sample of participants, both project partners and students from various universities. As far as questionnaires were concerned, approximately 30 participants were involved. This test was also used to ascertain the completion times. Concerning the mobile app, a two-week validation test was carried out.

Invitations, Pre-Study Diversity Questionnaire, and Participants. This was the first of the three phases of the data collection. This phase started by sending an email containing the survey description, the invitation to the first questionnaire, and information on the second part of the data collection (sensing component) via university mailing lists. This invitation was then reiterated through four weekly reminders to all students who still needed to complete the survey. Over 20,000 college students were contacted with mailing lists in the initial recruitment phase. Out of the set of people who were contacted, 13398 participants filled in the pre-study diversity questionnaire. Then, a subset of the eligible participants was selected to participate in the second part of the study, which was done with the mobile app. The requirements for the selection were two-fold: (i) having consented to the processing of personal data – this required participants agreeing to release mobile data collected during the study after anonymization, and (ii) owning an Android smartphone compatible with the app.

Mid-Study Questionnaire I, II and Mobile Sensing app. Of all the participants who completed the pre-study diversity questionnaire, 678 participants were chosen for the next phase with the mobile sensing app. This deployment was done between September and November 2020. The average age of study participants was 24.2 years (SD: 4.2), and the cohort had 58% females. They were sent emails with a specification manual to download and install the mobile sensing app. In addition, the participants completed the mid-study questionnaire I. Reminders were sent after one week for participants who still needed to complete the questionnaire. Then, participants completed time diaries, and sensing data were passively collected in the mobile app. After two weeks of mobile sensing app usage, the mid-study questionnaire II was sent to participants via email. After sending out this questionnaire, two more weeks of mobile sensing data collection were conducted. Daily reports were produced to facilitate monitoring the time diary survey and identify possible problems, including (1) the number of notifications each participant responded to and (2) the amount of data collected by the individual sensors. Using this information, local field supervisors could contact the inactive participants every three days and support them as needed. A further element of contact was the daily sending of the results of a daily prize, which was an additional incentive for participants.

Incentive Design. An incentive scheme was designed to motivate participants to complete time diaries and provide sensing data. Incentives included monetary prizes for participants who completed at

Table 2.4: Summary of 105 features extracted from sensing data, aggregated around activity self-reports using a time window. A detailed description of sensing modalities is provided in Appendix A.

Modality	Frequency	Features and Description
Location	1 sample per minute	radius of gyration, distance traveled, mean altitude
Bluetooth [low energy, normal]	1 sample per minute	number of devices (the total number of unique devices found), mean/std/min/max rssi (Received Signal Strength Indication – measures how close/distant other devices are)
WiFi	1 sample per minute	connected to a network indicator, number of devices (the total number of unique devices found), mean/std/min/max rssi
Cellular [GSM, WCDMA, LTE]	1 sample per minute	number of devices (the total number of unique devices found), mean/std/min/max phone signal strength
Notifications	on change	notifications posted (the number of notifications that came to the phone), notifications removed (the number of notifications that were removed by the user) – these features were calculated with and without duplicates.
Proximity	10 samples per second	mean/std/min/max of proximity values
Activity	2 samples per minute	time spent doing activities: still, in_vehicle, on_bicycle, on_foot, running, tilting, walking, other (derived using the Google activity recognition API [183])
Steps	10 samples per second or on change	steps counter (steps derived using the total steps since the last phone turned on at 10 samples per second), steps detected (steps derived using event triggered for each new step captured on change)
Screen events	on change	number of episodes (episode is from turning the screen of the phone on until the screen is turned off), mean/min/max/std episode time (a time window could have multiple episodes), total time (total screen on time within the time window)
User presence	on change	time the user is present using the phone (derived using android API that indicate whether a person is using the phone or not)
Touch events	on change	touch events (number of phone touch events)
App events	10 samples per minute	time spent on apps of each category derived from Google Play Store [285, 454]: action, adventure, arcade, art & design, auto & vehicles, beauty, board, books & reference, business, card, casino, casual, comics, communication, dating, education, entertainment, finance, food & drink, health & fitness, house, lifestyle, maps & navigation, medical, music, news & magazine, parenting, personalization, photography, productivity, puzzle, racing, role playing, shopping, simulation, social, sports, strategy, tools, travel, trivia, video players & editors, weather, word, not_found

least 85% of time diaries (e.g., 20 Euro in Italy, 150 Kr in Denmark, etc.), cash prizes for multiple daily winners randomly chosen from each pilot (e.g., five winners were given a prize of 5 Euro in Italy, 5 MNT in Mongolia, etc.). In the end, three winners from each country were randomly chosen for a larger prize (e.g., 150 Euros per person in Italy, 150 Sterling Pounds in the UK, etc.). Incentives in all countries were designed by considering each country's socioeconomic status and expecting all participants to be compensated and motivated equally.

Ethical Procedures. All the survey activities and results at each site complied with the national ethical privacy-protecting laws and guidelines, hence getting approvals from respective ethical review boards. In addition, all the experiments, including non-European pilots, were compliant with the General Data Protection Regulation (GDPR) [534]. Additionally, for non-European experiments, the activities and results have been developed to comply with those of a European country for compliance purposes. More specifically, Italian legislation was selected as the reference.

Pre-Processing the Dataset for Analysis

In feature engineering, interpretability was a key factor, as all the features were defined in a meaningful manner. Similar to prior work in ubicomp, we used a time window-based approach for matching sensor data to self-reports [468, 285, 330]. While different time windows can be chosen based on the application scenario, this thesis presents results with a dataset created using a time window of 10 minutes. Hence, if the self-report occurred at time T , sensor data would be considered from $T - 5$ minutes to $T + 5$ minutes. However, we also considered other time windows, such as 2, 4, 15, and 20 minutes. However, results showed that the 10-minute time window performed better for the task in Chapter 7. This could be because shorter time windows do not capture enough behaviors and contexts around self-reports to make a meaningful prediction regarding target attributes. Prior work has also shown that larger time windows can capture a high amount of information about user behaviors [30, 55]. However, we can not use very large windows above 20 minutes because it would lead to a situation where sensor data segments for self-reports might overlap, leading to data overlap between samples. Therefore, throughout this thesis, we present results with a ten-minute time window. Why each sensing modality was chosen has been discussed extensively in many prior studies on mobile sensing for behavior modeling and well-being [468, 247, 454, 285, 330, 30, 55, 544, 325]. The modalities and features crafted from each modality are summarized in Table 2.4.

3 Examining the Social Context of Alcohol Drinking of Young Adults with Smartphone Sensing

According to prior research, the type of relationship between an individual consuming alcohol and their social environment (friends, family, spouse, etc.), as well as the group size (alone, with one person, with a group), is linked to various aspects of alcohol consumption, including quantity, location, motives, and mood. Despite the acknowledged significance of social context in shaping young adults' drinking behavior, smartphone sensing research in this domain has been relatively limited. This chapter examines the weekend nightlife drinking habits of 241 young adults in a European country, utilizing a dataset that comprises self-reports and passive smartphone sensing data spanning three months. Employing multiple statistical analyses, we demonstrate the informativeness of features derived from various sensor modalities, such as accelerometer, location, app usage, Bluetooth, and proximity, in discerning different social drinking contexts. We introduce and assess seven social context inference tasks based on smartphone sensing data, achieving accuracy rates ranging from 75% to 86% in both binary and ternary classifications. Additionally, we explore the feasibility of identifying the gender composition of a friend group using smartphone sensor data with accuracies exceeding 70%. These findings provide promising support for future alcohol consumption interventions that incorporate users' social context more effectively and reduce reliance on self-reports when creating drink logs for self-tracking tools and public health studies. It is also worth noting that this research was done in the context of the SNSF Dusk2Dawn project, and the data collection effort is mentioned in [454]. The material of this chapter was originally published in [327].

3.1 Introduction

In western countries, alcohol consumption is a leading risk factor for mortality and morbidity [256]. The consumption of several drinks in a row, commonly referred to as binge drinking or heavy drinking, can lead to many short-term adverse consequences not only for the person drinking (e.g., accidents, unprotected sex, or injury [261]) but also at the family and community levels (e.g., violence, drunk driving [272, 259]). On a larger time frame, heavy alcohol consumption can also lead to long-term consequences, such as poor academic achievement, diminished work capacity, alcohol dependence, and premature death [313]. Adolescence and early adulthood appear as a particularly critical period of life for the development of risky alcohol-related behaviors since heavy alcohol consumption in late adolescence appears to persist into adulthood [557]. In order to limit excessive drinking among adolescents and young adults, it is essential to understand the etiology and antecedents of drinking occasions [257]. Prior work in social and epidemiological research on alcohol has emphasized the

importance of the social context in shaping people's alcohol use and motives [320, 156, 257] in the sense that the consumption of alcohol or not, and the amounts consumed, vary depending on the presence or absence of family members [407, 518, 519, 400], of friends or colleagues [347, 380, 407, 132], and of the spouse or partner [279, 275, 291]. Additionally, a recent literature review showed that although the type of company is generally not a significant direct predictor of alcohol-related harm, young adults tend to experience more harm, independent of increased consumption, when they drink in larger groups [488].

Recent developments in ambulatory assessment methods (i.e., the collection of data in almost real-time, for example, every hour, and in the participant's natural environment [475, 481]) using smartphones made it possible to assess the type and the number of people present over the course of real-life drinking occasions [260, 262]. Compared to cross-sectional retrospective surveys traditionally used in alcohol epidemiological and psychological research, this type of approach allows the capture of the interplay between drinking behaviors and contextual characteristics at the drinking event level in more detail [260]. For instance, evidence shows that larger numbers of drinking companions are associated with increased drinking amounts over the course of an evening or night [508, 480], and that this relationship is mediated by the companions' gender [509]. By repetitively collecting information from the same individuals over multiple occasions, ambulatory assessment methods are able to capture a large diversity of social contexts of real-life drinking occasions (e.g., romantic date with a partner, large party with many friends, family dinner) with the advantages of being free of recall bias and of participants serving as their own controls.

In addition to the possibility of capturing in-situ self-reports, smartphone-based apps have the potential to provide just-in-time adaptive interventions (JITAI) and feedback [362, 325]. Feedback systems primarily rely on identifying users' internal state or the context that they are in to offer interventions or support (feedback) [255, 486]. Leveraging these ideas, recent studies in alcohol research have used mobile apps to provide interventions to reduce alcohol consumption using questionnaires, self-monitoring, and location-based interventions [194, 25, 579]. Furthermore, mobile sensing research has used passive sensing data from wearables and smartphones to infer aspects that could be useful in feedback systems, such as inferring drinking nights [454], inferring non-drinking, drinking, and heavy drinking episodes [30], identifying walking under alcohol influence [241] and detecting drunk driving [119]. Hence, given that the characteristics of the social context have been identified as central elements of any drinking event, it appears as a central target for inferring drinking occasions. However, to the best of our knowledge, mobile sensing has not been widely used to automatically infer the social context of alcohol-drinking events. Consider the following example to further understand the importance of identifying social context using mobile sensing. If an app could infer a heavy-drinking episode (as shown by [30]), it could provide an intervention. However, there is a significant difference between drinking heavily alone or with a group of friends [478, 177]. Drinking several drinks in a row alone might indicate that the person is in emotional pain or stressed (also known as "coping" drinking motive) [257, 478]. However, drinking several drinks is common when young adults go for a night out with friends [177]. In a realistic setting, for a mobile health app to provide useful interventions or feedback, the knowledge of the social context and knowing that the user is in a heavy-drinking episode could be vital. Hence, understanding fine-grained contextual aspects related to alcohol consumption using passive sensing is important and could also open new doors in mobile interventions and feedback systems for alcohol research.

Further, there is a plethora of alcohol tracking, food tracking, and self-tracking applications in app stores that primarily rely on user self-reports [335, 328, 454]. Even though gaining a holistic understanding of eating or drinking behavior is impossible without capturing contextual aspects regarding such

behaviors, prior work has shown that people tend to reduce the usage of apps that require a large number of self-reports and tend to use health and well-being applications that function passively [325]. Mobile sensing offers the opportunity to infer attributes that otherwise require user self-reports, hence reducing user burden [325, 330, 328]. In addition, mobile sensing could infer attributes to facilitate search acceleration in food/drink logging apps [234]. The social context of drinking alcohol is a variable that could benefit from smartphone sensing in an alcohol-tracking application. As a whole, the idea of using smartphone sensing, in addition to capturing self-reports, is to gain a holistic understanding of the user context passively, which could otherwise take a long time span if collected using self-reports. Considering all these aspects, We address the following research questions:

RQ1: What social contexts around drinking events can be observed by analyzing self-reports and smartphone sensing data corresponding to weekend drinking episodes of a group of young adults?

RQ2: Can young adults' social context of drinking be inferred using sensing data? What are the features that are useful in making such inferences? Are social context inference models robust to different group sizes?

By addressing the above research questions, our work makes the following contributions:

Contribution 1: Using a fine-grained mobile sensing dataset that captures drinking event-level data from 241 young adults in a European country, we first show that there are differences in self-reporting behavior among men and women, regarding drinking events done with family members and with groups of friend/colleagues. Next, using various statistical techniques, we show that features coming from modalities such as accelerometer, location, bluetooth, proximity, and application usage are informative regarding different social contexts around which alcohol is consumed.

Contribution 2: We first define seven social context types, based on the number of people in groups (e.g., alone, with another person, with one or more people, with two or more people) and the relationship between the participant and others in the group (e.g., family or relatives, friends or colleagues, spouse or partner). Then, based on the above context types, we evaluate four two-class and three three-class inference tasks regarding the social context of drinking, using different machine learning models, obtaining accuracies between 75% and 86%, with all passive smartphone sensing data. In addition, we show that models that only take inputs from single sensor modalities, such as accelerometer and application usage, could still perform reasonably well across all seven social context inferences, providing accuracies over 70%.

The chapter is organized as follows. In Section 3.2, we describe the background and related work. In Section 3.3, we describe the study design, data collection procedure, and feature extraction techniques. In Section 3.4 and Section 3.5, we present a descriptive analysis and a statistical analysis of dataset features. We define and evaluate inference tasks in Section 3.6. Finally, we discuss the main findings in Section 3.6, and conclude the chapter in Section 3.7.

3.2 Background and Related Work

3.2.1 The Social Context of Drinking Alcohol

While there are numerous definitions for the term social context in different disciplines, in this chapter, we borrow the concept commonly used in alcohol research [264, 111, 257, 320], which refers to either

one or both of the following aspects: (1) *type of relationship*: the relationship between an individual and the people in the individual's environment with whom she or he is engaging, and (2) *number of people*: the number of people belonging to each type of relationship, with whom the individual is engaging. By combining the two aspects, a holistic understanding of the social context of drinking of an individual can be attained.

The consumption of alcohol is associated with different contextual characteristics. These characteristics include the type of setting (e.g., drinking location), its physical attributes (e.g., light, temperature, furniture), its social attributes (e.g., type, size, and sex-composition of the drinking group, ongoing activities), and the user's attitudes and cognition [320]. Applied to real-life situations, this conception underlines the changing nature of the drinking context, in the sense that the variety of situations during which alcohol might be consumed is rather large. For instance, across three consecutive days, the same person might drink in a restaurant during a date with a romantic partner, join a large party at a nightclub with many attendees, and finally, join a quiet family dinner at home.

Among all contextual characteristics, the composition of the social context is a central element of any drinking occasion, since the consumption of alcohol is predominantly a social activity for non-problematic drinkers [478]. Among adolescents and young adults, previous literature has shown that amounts of alcohol consumed on any specific drinking occasion vary depending on the type and number of people present [111]. The type of relationship that has received the most attention so far is the presence of friends, in terms of number and of sex composition. Converging evidence shows that the likelihood of drinking [52] and drinking amounts are positively associated with the size of the drinking group [480, 508, 160]. Unfortunately, the group size is generally used as a continuous variable, preventing the identification of a threshold at which the odds of drinking in general or drinking heavily increase. Evidence regarding the sex composition of the group, however, provided mixed results, with some studies indicating that more alcohol is consumed in mixed-sex groups [509, 290] while others suggesting that this might rather be the case in same-sex groups [19]. The influence of the presence of the partner (e.g., boyfriend or girlfriend) within a larger drinking group has not been investigated, but evidence suggests that alcohol is less likely to be consumed and in lower amounts in a couple situation (i.e., the presence of the partner only) [509, 227]. It should be noted that these studies only suggest correlational links between contextual characteristics and drinking behaviors and should not be interpreted as causal relationships.

The presence or absence of members of the family also plays an important role in shaping adolescents' and young adults' drinking behaviors. In particular, the presence of parents and their attitude towards drinking are often described as being either limiting or facilitating factors, but evidence in this respect is inconclusive. For instance, the absence of parental supervision was found to be associated with an increased risk for drinking at outdoor locations and young adults' homes [510], suggesting that their presence might decrease this risk. However, another study shows that parents' knowledge about the happening of a party is negatively associated with the presence of alcohol, but there was no relationship between whether a parent was present at the time of the party and the presence of alcohol [158]. Lastly, parents might also facilitate the use of alcohol by supplying it, especially to underage drinkers [227, 172]. To sum up, evidence on the impact of the presence or absence of parents on young people's drinking appears mixed as this might be related to their attitude towards drinking, with some parents being more tolerant or strict than others [385]. Lastly, it should be noted that the presence of siblings has rarely been investigated, but unless they have a supervision role in the absence of parents, their role within the drinking group might be similar to one of friends.

3.2.2 Alcohol Consumption and Mobile Phones

Mobile Apps for Interventions in Alcohol Research

Many mobile apps in alcohol research focus on providing interventions or feedback to users to reduce alcohol consumption [112, 123, 194]. Crane et al. [112] conducted a randomized controlled trial using the app called "Drink Less", to provide interventions. This app relied on user self-reports, and they concluded that the app helped reduce alcohol consumption. Moreover, Davies et al. [123] conducted a randomized controlled trial with an app called "Drinks Meter", that provided personalized feedback regarding drinking. This app also used self-reports to provide feedback. Similarly, many mobile health applications in alcohol research that provide users with interventions or feedback primarily use self-reports [212, 381, 79]. Regarding sensing, Gustafson et al. [194] deployed an intervention app called ACHES, which provided computer-based cognitive behavioral therapy and additional links to useful websites, and this app provided interventions to users when they entered pre-defined high-risk zones, primarily relying on location sensing capabilities of the smartphone. LBMI-A [141] by Dulin et al. is another study that is similar to ACHES. In summary, alcohol epidemiology research that used mobile apps primarily targeted interventions based on self-reports or simple sensing mechanisms. Even though many studies have identified that self-reports are reasonably accurate for capturing alcohol consumption amounts [289], studies have also stated that heavy-drinking episodes are often under-reported when self-reporting [374]. In addition, unless there is a strong reason for users to self-report, there is always the risk of users losing motivation to use the app over time.

Smartphone Sensing for Health and Well-Being

Smartphones allow sensing health and well-being aspects via continuous and interaction sensing techniques, both of which are generally called passive sensing [325]. This capability has been used in areas such as stress [74, 300], mood [285, 485], depression [545, 77], well-being [288, 269], and eating behavior [55, 328, 330]. If we consider drinking-related research in mobile sensing, Bae et al. [30] conducted an experiment with 30 young adults for 28 days and used smartphone sensor data to infer non-drinking, drinking, and heavy-drinking episodes with an accuracy of 96.6%. They highlighted the possibility of using such inferences to provide timely interventions. Santani et al. [454] deployed a mobile sensing application among 241 young adults for a period of 3 months to collect sensor data around weekend nightlife events. They showed that sensor features could infer drinking and non-drinking nights with an accuracy of 76.6%. Kao et al. [241] proposed a phone-based system to detect feature anomalies of walking under the influence of alcohol. Further, Arnold et al. [22] deployed a mobile application called Alco Gait, to classify the number of drinks consumed by a user into sober (0-2 drinks), tipsy (3-6 drinks) or drunk (greater than six drinks) using gait data, obtaining reasonable accuracies. While most of these studies focused on detecting drinking events/episodes/nights, we focus on inferring the social contexts of drinking events.

Event Detection and Event Characterization in Mobile Sensing

Smartphone sensing deployments can be classified into two based on the study goal [325]: (a) Event Detection (e.g., drinking alcohol, eating food, smoking, etc.) and (b) Event Characterization (characteristics of the context that helps understand the event better – e.g., social context, concurrent activities, ambiance, location, etc.). For domains such as eating behavior, there are studies regarding both event detection (identifying eating events [44, 354], inferring meal or snack episodes [55], inferring food

categories [329]) and event characterization (inferring the social context around eating events [328]). Inferring mood [468, 285] as well as identifying contexts around specific moods [121] has been attempted in ubicomp. However, even though alcohol epidemiology researchers have attempted to characterize alcohol consumption to gain a more fine-grained understanding of drinking, mobile sensing research has not been focused on the social context aspect thus far, even though some studies have looked into event detection [30, 454, 241, 22]. Hence, we aim to address this research gap by focusing on the social context of drinking alcohol using smartphone sensing.

3.3 Data, Features, and Tasks

3.3.1 Mobile Application, Self-Reports, and Passive Sensing

We used a dataset regarding young adults' nightlife drinking behavior from our group's previous work [454]. This dataset contains multimodal passive sensing data from phones and self-reports regarding the drinking behavior of 241 young adults (53% men) in Switzerland during weekend nights, throughout a period of three months, and was collected as a collaboration between alcohol researchers, behavioral scientists, and computer scientists. In this section, we briefly describe the study design, the data collection procedure, and the feature extraction technique. A full description regarding the ethical approval, deployment, and data collection procedure can be found in [580, 455, 454].

Mobile App Deployment

To collect data from study participants, an Android mobile application was deployed, and this app had two main components: (a) *Drink Logger*: used to collect in-situ self-reports during weekend nights (Friday and Saturday nights, from 8.00 pm to 4.00 am next day). The app sent notifications hourly, asking whether users wanted to report a new drink; and (b) *Sensor Logger*: used many passive sensing modalities to collect data, including both continuous (accelerometer, battery, bluetooth, location, wifi) and interaction (applications, screen usage) sensing. The application was deployed from September to December 2014. The study participants were young adults with ages ranging from 16 to 25 years old (mean=19.4 years old, SD=2.5). More details regarding the deployment can be found in [454].

Self-Reports

Whenever they were about to drink an alcoholic or non-alcoholic drink, participants were requested to take a picture of it and to describe its characteristics and the drinking context using a series of self-reported questionnaires [263]. Participants labeled the drink type using a list of 6 alcoholic drinks (e.g., beer, wine, spirits, etc.) and six non-alcoholic drinks (e.g., water, soda, coffee, etc.). Then, in accordance with the definition of social context we adopted in Section 3.2.1, participants reported the type and number of people present for each of the following categories: (a) partner or spouse; (b) family or relatives; (c) male friends or colleagues; (d) female friends or colleagues; and (e) other people (called *type of relationship* in the remainder of the chapter). These five categories were adopted from prior work in alcohol research [264]. Next, for each type of relationship, participants reported the *number of people* using a 12-point scale with 1-point increments from 0 to 10, plus 'more than 10'; with the exception of partner or spouse which could either be absent (coded as 0) or present (1). This scale was designed to measure variations in the social context, following the assumption that the presence of each person counts within small groups but that the additional value of each extra

person is less important within larger groups (e.g., ten or more people). Further, information about participants, including age, sex, occupation, education level, and accommodation, was collected in a baseline questionnaire. Overall, by selecting self-reports of situations when participants reported the consumption of an alcoholic drink, we were left with 1254 self-reports for the analysis.

Passive Smartphone Sensing

To gain a fine-grained understanding of users' drinking behavior, passive sensing data were collected during the same time period when participants self-reported alcohol consumption events. The chosen sensing modalities were Accelerometer (ACC), Applications (APP), Location (LOC), Screen (SCR), Battery (BAT), Bluetooth (BLU), Wifi (WIF), and Proximity (PRO). A dataset summary is given in Table 3.1, and an extensive description is given in [454, 404].

3.3.2 Aggregation and Matching of Self-Reports and Passive Sensing Data

Prior studies that used this dataset primarily considered *user-night* as the point of analysis (e.g., inferring nights of alcohol consumption vs. no alcohol consumption [454], inferring heavy-drinking nights [404]', etc.). However, in this study, we consider drink-level data that is more fine-grained. We prepared the self-report dataset such that each entry corresponds to a drinking event. Then, to combine sensor data and self-reports in a meaningful manner, we used the following two-phase technique, which was adopted from prior ubicomp research [29, 468, 328, 55]:

Phase 1 (Aggregation): We aggregate raw sensor data for every ten-minute window throughout the night. Different techniques were used for the aggregation of sensors. Hence, for a user night, we have 48 ten-minute windows from 8.00 pm to 4.00 am the next day. For each feature derived from each sensor, we have 48 values (6 ten-minute windows per hour X eight hours per night) for a user night. For instance, if there is a feature F_1 derived from sensor S_1 , for each user and for each night, F_1 would have 48 values that represent time windows from 8.00 pm-8.09 pm, 8.10 pm-8.19 pm, 8.20 pm-8.29 pm, until 03.50 am-03.59 am of next day.

Phase 2 (Matching): During this phase, features are matched to alcohol consumption self-reports using a one-hour window (approximately from 30 minutes before the alcohol consumption self-report to 30 minutes after the drinking self-report). For instance, if the drinking was reported at 10.08 pm, we calculate the average ($_avg$), maximum ($_max$), and minimum ($_min$) values for each feature using values corresponding to six ten-minute windows (obtained in Phase 1) from 9.40 pm to 10.39 pm, and match those value to the self-report. The idea behind this aggregation is to capture sensor data around drinking events, expecting that contextual cues could be informative of different social contexts. The summary of passive sensing features is provided in Table 3.1. These features were derived for every ten-minute time slot throughout the night, as discussed in Phase 1 (Aggregation), resulting in a total of 134 features. Then, in accordance with the procedure in Phase 2 (Matching), a one-hour time window around each drinking self-report was considered. By considering the average, minimum, and maximum of the six values corresponding to the one-hour time window, 402 passive sensing features were included for each alcohol-drinking self-report in the final dataset. This two-phase technique is summarized in Figure 3.2. We obtained a dataset after following this technique for all self-reports and sensor features. We removed data points with incomplete sensor data (152 records with unavailable sensor data, mainly location, wifi, or bluetooth data), not enough data for the matching phase (102 records for drinking events that were done between 8.00 pm-8.30 pm and 3.30 am-4.00 am), and

Chapter 3. Examining the Social Context of Alcohol Drinking of Young Adults with Smartphone Sensing

Table 3.1: Summary of features extracted from mobile sensors (134). Sensor data are aggregated for every 10-minute time slot from 8 pm to 4 am. For all the given features, average, minimum, and maximum were calculated during the matching phase, resulting in 402 sensing features for each alcohol consumption event.

Sensor – Feature Type (# of features)	Sensor Description Feature Description
Location – Attributes (10) – Signal (3)	Location data were continuously collected for a time period of 1 minute during each 2-minute time slot. Collected data included data source, longitude, latitude, signal strength, and accuracy. {min., max., med., avg., std.} of avg. of speed and sensor accuracy 3 signal strengths (GPS, network, unknown)
Accelerometer – Raw (15) – Angle (15) – Dynamic (20)	Values from all three axes of the sensor were collected 10 seconds continuously, at a frequency of 50Hz, during every minute. We calculated (a) basic statistics from raw sensor data from the X, Y, and Z-axes [454]; (b) aggregated statistics related to acceleration (m, mNew, dm) and signal magnitude area (mSMA) by combining data from three axes [242, 314, 404]; and (c) angle between acceleration and the gravity vector [404, 454]. {min., max., med., avg., std.} of avg. of xAxis, yAxis, zAxis of accelerometer {min., max., med., avg., std.} of angle of xAxis, yAxis, zAxis with g vector {min., max., med., avg., std.} of mSMA, dm, m, mNew values
Bluetooth – Count (4) – Strength (5)	The list of available devices was captured as Bluetooth logs, once every 5 minutes. Features such as the number of devices around, signal strengths, and empty scan counts were captured. number of bluetooth IDs surrounding devices, records, bluetooth scan count, empty scan count {min., max., med., avg., std.} of bluetooth strength signal of surrounding devices
Wifi – Count (2) – Attributes (10)	The list of available devices was captured as WiFi logs, once every 5 minutes. Features such as the number of hotspots around, signal strengths, and empty scan counts were captured. wifi record , wifi id set {min., max., med., avg., std.} of level, frequency of wifi hotspot
Application – Count (2) – Category (33)	Applications were categorized into 33 groups (e.g., art & design, food & drink, social, games, etc.) based on the categorization provided in Google Play Store [184]. Using the categories and running apps, statistics such as the total number of running apps and the number of running apps based on categories were calculated [285, 454, 404]. app count, app record normalized 33-bin histogram of 33 application categories
Proximity – Count (1) – Distance (5)	Several statistical featured were derived using raw values of the proximity sensor. proximity records {min., max., med., avg., std.} of distance from phone to objects
Battery – Status (5) – Level (5) – Count (2)	Battery levels were captured once every five minutes, and status changes were captured whenever a change occurred. Hence, several features, including battery full, discharging, charging, battery level, and whether the phone is plugged-in or not, were derived. 5 battery statuses {min., max., med., avg., std.} of battery level count of battery records and plugged times
Screen – Usage (1)	Screen data were recorded whenever the screen status changed. Using the captured data, we derived the percentage of screen-on time. percentage of screen on time

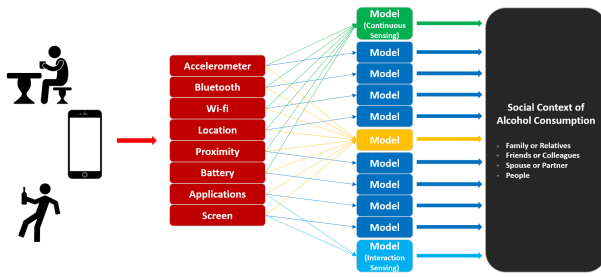


Figure 3.1: A schematic diagram representing the summary of the study

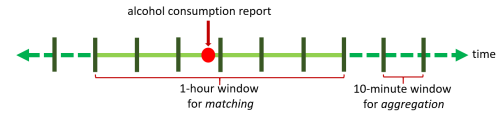


Figure 3.2: Diagram summarizing the two-phase technique for combining self-reports and sensing data

self-reports that were produced while visiting other countries (59 records, when the participant traveled while being in the study). The final dataset contained 941 complete drinking reports with sensor features.

3.3.3 Deriving Two-Class and Three-Class Social Context Features

In Section 3.2, we described how social contexts such as being with/without family members, friends/colleagues, and spouse/partner could be associated with drinking behavior. In addition, under Section 3.3.1, we described the type of social contexts reported by participants. Among them, features such as *with male friends/colleagues*, *with female friends/colleagues*, and *with family members* had twelve-point scales, and *with partner/spouse* had a two-point scale. However, for the purpose of this analysis, we reduced the twelve-point scale to low-dimensional scales (two-point and three-point), with the objective of capturing social context group dynamics that are meaningful in terms of drinking events such as: being alone, with another person, or in a group of two or more. We followed the following steps.

First, except for the feature *with partner or spouse*, which is already two-class, we minimized the scale of other features to two-classes and three-classes. For two-class features, the values could be either zero or one, whereas – *zero*: the participant is not with anyone belonging to the specific social context; and *one*: the participant is with one or more others belonging to the specific social context (hence, in a group). For three-class features, the values could be either zero, one, or two as follows – *zero*: the participant is not with anyone belonging to the specific social context; *one*: the participant is with one other person belonging to the specific social context (hence, in a group of two people); *two*: the participant is with two or more people belonging to the specific social context (hence, in a larger group).

Then, we derived several new features using the existing features:

- *without friends/colleagues vs. with friends/colleagues* (two-class): this aggregated features about men and women friends/colleagues into a single two-class variable by discarding the sex demographic attribute of friends/colleagues.
- *without friends/colleagues vs. with another friend/colleague vs. with two/more friends/colleagues* (three-class): this aggregated features about the men and women friends/colleagues into a single three-class feature.
- *without people vs. with people* (two-class): this feature combines all the two-class social contexts

Table 3.2: Summary of social contexts in the final dataset.

Social Context	Classes	Interpretation
family _{two}	2	without vs. with one/more family members/relatives
partner _{two}	2	without vs. with the partner/spouse
friends _{two}	2	without vs. with one/more friends/colleagues
people _{two}	2	without vs. with one/more people
family _{three}	3	without vs. with one vs. with two/more family members/relatives
friends _{three}	3	without vs. with one vs. with two/more friends/colleagues
people _{three}	3	without vs. with one vs. with two/more people

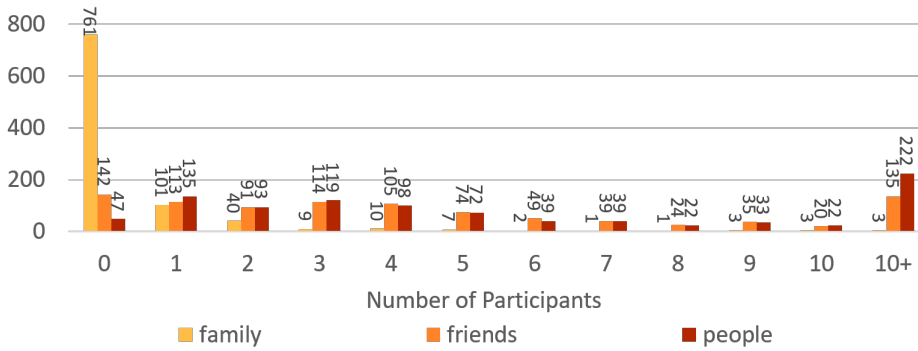


Figure 3.3: Distribution of original self-report features (family, friends) and a derived feature (people)

to estimate the overall two-class social context of the user.

- *without people vs. with another person vs. with two/more people* (three-class): this feature combines all the other three-class social contexts and the two-class feature *with partner/spouse*, to estimate the overall three-class social context of the user.

The final set of social context features used for this study are summarized in Table 3.2. In accordance with the definition of social context proposed in Section 3.3.3, these features capture two aspects. First, they capture the relationships between the study participant and people engaging with the participant during alcohol consumption. Second, they capture group dynamics for each relationship (e.g., alone, with another person – small group of two people, with two/more people – comparatively large group, etc.). According to prior work in alcohol research, both perspectives are important to obtain a fine-grained understanding about drinking behavior [320, 43]. The summary of our analytical setting is presented in Figure 3.1.

3.4 Descriptive Analysis (RQ1)

In this section, we provide a descriptive analysis regarding self-reports using demographic information, to understand the nature of the aggregate drinking behavior of participants.

Self-Report Distribution Breakdown Based on Social Contexts: Figure 3.3 provides a distribution of self-reports. Self-reports for partner is not shown here because it is only a binary response, and is included in Figure 3.5. Figure 3.5 and Figure 3.6 provide a distribution of self-reports for four two-class

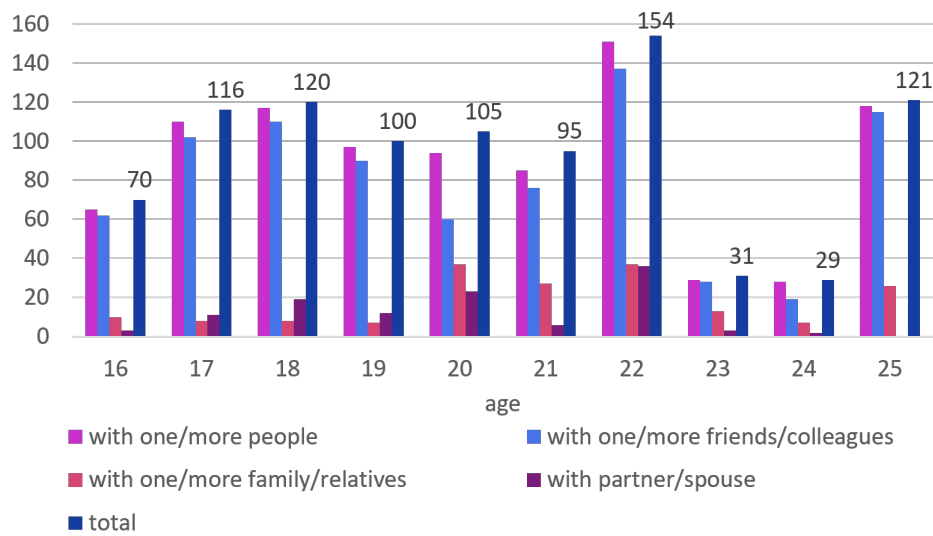


Figure 3.4: Self-Reports in terms of Age

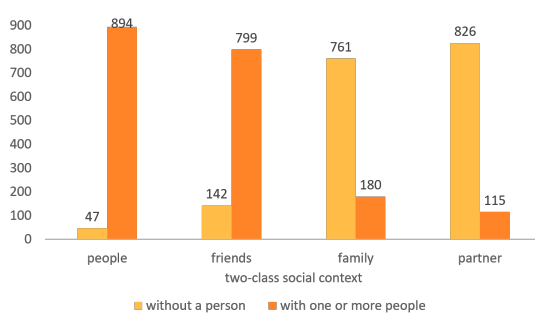


Figure 3.5: Distribution of two-class social contexts

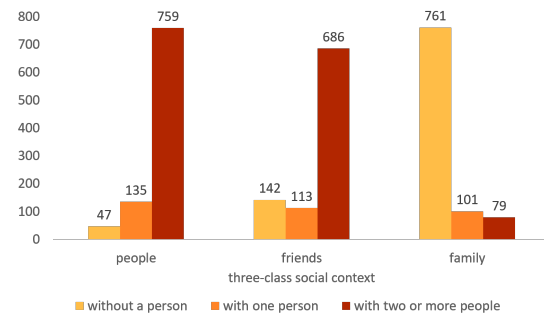


Figure 3.6: Distribution of three-class social contexts

social contexts and three three-class social contexts, respectively. Results in Figure 3.5 show that only 47 (5.0%) drinking occasions were done alone as compared to 894 (95.0%) occasions that were done with one/more people. Out of these 894 reports, 799 (89.4%) were reported to have happened with two/more friends/colleagues. According to Figure 3.6, these 799 reports consist of 113 (14.1%) reports that were done with one friend/colleague and 686 (85.9%) reports that were done with two/more friends/colleagues, hence in a larger group. In summary, participants consumed alcohol while being alone only on a small portion of occasions. This result is comparable to prior alcohol research, which shows that solitary drinking episodes are less frequent as compared to other social contexts [43]. Moreover, the presence of two or more friends/colleagues was reported well over more than half of all drinking occasions ($686/941 = 72.9\%$). This result too is in line with prior work that states that young adults tend to drink alcohol for social facilitation and peer acceptance [43].

Self-Report Distribution Breakdown Based on Sex: In Figure 3.7 and Figure 3.8, we present distributions of self-reports, based on sex and social context pairs. Results indicate that social contexts 'people' and 'friends' reported more drinking occasions with one/more people, for both men and

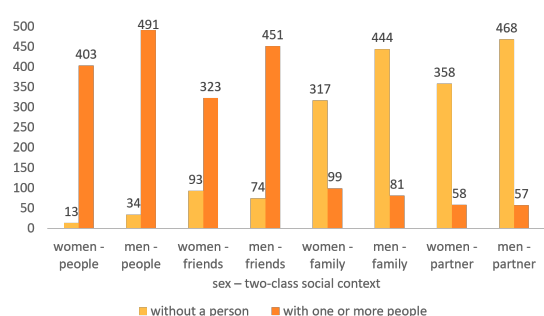


Figure 3.7: Self-Reports in terms of Sex and Two-Class Social Context

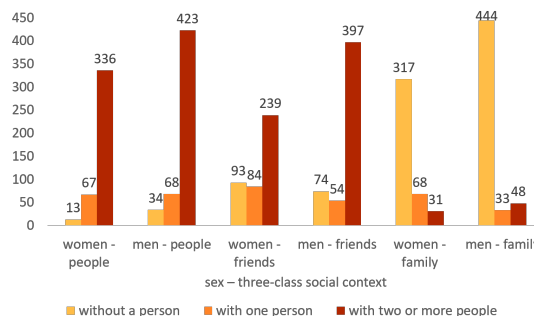


Figure 3.8: Self-Reports in terms of Sex and Three-Class Social Context

women, whereas social contexts 'partner' and 'family' have significantly high numbers of drinking events that were reported to be done alone. In addition, for the social context 'friends', Figure 3.8 shows that the proportion of self-reports in groups of two/more (239) is just over half for women (239/416 = 57.5%), whereas for men, drinking events with two/more friends/colleagues is 75.6% (397/525), which is almost a 20% difference for two sexes. This suggests that men reported a higher proportion of drinking occasions in groups of two/more people. This result is consistent with prior literature that states that men tend to drink in larger social contexts (especially with friends/colleagues) whereas women are less likely to do so [511]. Further, women participants have reported drinking with family members 99 times (99/416 = 23.8%), whereas men only reported to have done so 81 times (81/525 = 15.4%), which is about 9% less than women.

Self-Report Distribution Breakdown Based on Age: As shown in Figure 3.4, participants' age ranged from 16 to 25. Except for ages 23 and 24 (31 and 29 self-reports, respectively), all other ages had over 70 self-reports. Moreover, the highest proportion of situations with one/more friends/colleagues (115/121 = 95.0%) was reported by participants aged 25. The lowest proportion of situations with a partner/spouse (0%) was reported by the same age group.

3.5 Statistical Analysis (RQ1)

3.5.1 Pearson and Point-Biserial Correlation for Social Contexts and Passive Sensing Features

We conducted Pearson (PCC) [501] and Point-biserial (PBCC) [45] correlation analyses to measure the strength and the direction of the relationships between each of the three-class (takes values 0,1, and 2) and two-class (takes values 0 and 1) features and passive sensing features. The results of the top five features with the highest PCC or PBCC with each social context are summarized in Table 3.3. For a majority of social contexts ($family_{two}$, $family_{three}$, $friends_{two}$, $friends_{three}$, $people_{two}$, and $people_{three}$), multiple accelerometer-related features were among the top five features based on the correlation coefficient values. The exception is the social context $partner_{two}$, where application features (e.g. food and drink) were among the top five. However, most of the values suggested, at best, weak positive or negative relationships.

Table 3.3: Pearson Correlation Co-efficient (PCC) and Point-Biserial Correlation Co-efficient (PBCC) for Sensor Features and Social Contexts (two-class and three-class). With the top 5 features for each Social Context are included in the table. p-values are denoted with the following notation: p-value $\leq 10^{-4}$:****; p-value $\leq 10^{-3}$:***

	family		partner		friends		alone	
	Feature (Sensor)	PBCC	Feature (Sensor)	PBCC	Feature (Sensor)	PBCC	Feature (Sensor)	PBCC
two-class	angleXMax_min (ACC)	0.19 (+) ****	yAxisAvgMax_avg (ACC)	0.14 (-) ****	mMean_avg (ACC)	0.22 (+) ****	mMed_max (ACC)	0.19 (+) ****
	angleXMed_min (ACC)	0.19 (+) ****	food_and_drink_avg (APP)	0.13 (+) ****	mSMAMean_avg (ACC)	0.22 (+) ****	mSMAMed_max (ACC)	0.19 (+) ****
	angleYAvg_min (ACC)	0.18 (+) ****	food_and_drink_min (APP)	0.12 (+) ****	mMed_max (ACC)	0.22 (+) ****	mMean_max (ACC)	0.19 (+) ****
	angleXAvg_min (ACC)	0.18 (+) ****	food_and_drink_min (APP)	0.12 (+) ****	mSMAMed_max (ACC)	0.22 (+) ****	mSMAMean_max (ACC)	0.19 (+) ****
	angleYMax_min (ACC)	0.18 (+) ****	zAxisAvgMin_avg (ACC)	0.12 (+) ****	mMax_avg (ACC)	0.21 (+) ****	dmListStd_max (ACC)	0.18 (+) ****
three-class	Feature (Sensor)	PCC	-	-	Feature (Sensor)	PCC	Feature (Sensor)	PCC
	angleXMed_min (ACC)	0.18 (+) ****	-	-	mMean_avg (ACC)	0.24 (+) ****	mMean_avg (ACC)	0.23 (+) ****
	angleXMax_min (ACC)	0.18 (+) ****	-	-	mSMAMean_avg (ACC)	0.24 (+) ****	mSMAMean_avg (ACC)	0.23 (+) ****
	angleYAvg_min (ACC)	0.18 (+) ****	-	-	mMed_avg (ACC)	0.23 (+) ****	yAxisAvgMin_min (ACC)	0.23 (-) ****
	angleYMin_min (ACC)	0.17 (+) ****	-	-	mSMAMed_avg (ACC)	0.23 (+) ****	mMean_max (ACC)	0.23 (+) ****
angleXMean_min (ACC)	0.17 (+) ****	-	-	mSMAMax_avg (ACC)	0.23 (+) ****	mSMAMean_max (ACC)	0.23 (+) ****	

3.5.2 Statistical Analysis of Dataset Features

Table 3.4 shows statistics such as t-statistic [249], p-value [190], and Cohen's-d (effect size) with 95% confidence interval (CI) [267] for the top five features in the dataset for the seven different social contexts. For two-class social contexts, the objective is to identify passive sensing features that help discriminate between: without people (alone) and with one/more people (group). Here, the term group is used because it could either be a small group of two to three people or a large group of more than ten people. Further, for three-class social contexts, the objective is to identify passive sensing features that help discriminate between: (a) without people (alone) vs. with one person (sgroup); (b) with one person (sgroup) vs. with two/more people (lgroup); and (c) without people (alone) vs. with two/more people (lgroup), where sgroup and lgroup stand for small group and large group, respectively. The features are ordered by the descending order of t-statistics and Cohen's-d values. In addition, prior work stated the lack of sufficient informativeness in p-values [577, 276]. For this reason, we calculated the Cohen's-d [429] to measure the statistical significance of features. We adopted the following rule-of-thumb, commonly used to interpret Cohen's-d values: 0.2 = small effect size; 0.5 = medium effect size; and 0.8 = large effect size. According to this notion, the higher the value of Cohen's-d, the higher the possibility of discerning the two groups using the feature. In addition, 95% confidence intervals for Cohen's-d were calculated, and if the interval does not overlap with zero, the difference can be considered as significant [276].

For the social context, family_{three}, features from the bluetooth sensor were among the top five in terms of t-statistic and Cohen's-d for the combination alone vs. sgroup. In addition, all the top five features had Cohen's-d values closer to medium effect size. Further, a total of 122 features had Cohen's-d values above small effect size and confidence intervals, not including zero. For the combinations sgroup vs. lgroup and alone vs. lgroup, the majority of features were from the accelerometer, and two features (video_player and system) were from application usage. In addition, if the hierarchy of the social contexts alone, sgroup, and lgroup are considered, sgroup is in the middle, sandwiched by alone and lgroup, which are further apart; hence, it would be easier to discern between these two groups. This is indicated in the results for the combination alone vs. lgroup, which have higher t-statistics and Cohen's-d values (some around medium effect size) compared to the other two combinations (alone vs. sgroup and sgroup vs. lgroup). Furthermore, for the social contexts friends_{three} and people_{three}, for all three combinations, all features in the top five in terms of both t-statistic and Cohen's-d are from the accelerometer. In addition, for friends_{three}, features in the combination alone vs. lgroup had high t-statistics and Cohen's-d values above medium effect size. In fact, 14 features, all of which are from the accelerometer, had Cohen's-d values above medium effect size. In addition, for people_{three}, 44

Chapter 3. Examining the Social Context of Alcohol Drinking of Young Adults with Smartphone Sensing

Table 3.4: t-statistic (T) (p-value $\leq 10^{-4}$.****; p-value $\leq 10^{-3}$.***; p-value $\leq 10^{-2}$.**), and Cohen's-d (C) with 95% confidence intervals (* if confidence interval include zero). Top five features are shown in decreasing order.

Feature	T	Feature	C	Feature	T	Feature	C	
family_{three}				friends_{three}				
alone vs. sgroup	blueStrengthMax_avg (BLU)	4.41****	blueStrengthMed_avg (BLU)	0.42	yAxisAvgMin_max (ACC)	4.05****	dmListMean_max (ACC)	0.37
	blueStrengthAvg_avg (BLU)	4.21****	zAxisAvgStd_min (ACC)	0.41	xAxisAvgStd_avg (ACC)	3.76***	xAxisAvgStd_avg (ACC)	0.34
	blueStrengthMed_avg (BLU)	4.21****	mMin_min (ACC)	0.41	angleStd_avg (ACC)	3.72***	angleStd_avg (ACC)	0.33
	blueStrengthMin_avg (BLU)	3.87***	dmListStd_min (ACC)	0.40	xAxisAvgStd_max (ACC)	3.65***	dmListMedian_max (ACC)	0.33
	blueStrengthMax_min (BLU)	2.71**	zAxisAvgMean_min (ACC)	0.40	yAxisAvgMax_avg (ACC)	3.55***	angleMin_max (ACC)	0.32
sgroup vs. lgroup	zAxisAvgMean_min (ACC)	3.22**	dmListMean_min (ACC)	0.38	mMedian_avg (ACC)	3.44***	yAxisAvgMin_avg (ACC)	0.38
	zAxisAvgMedian_min (ACC)	3.08**	zAxisAvgStd_min (ACC)	0.36	mMean_avg (ACC)	3.43***	yAxisAvgMean_avg (ACC)	0.35
	anglezMax_min (APP)	2.84**	anglezMax_min (ACC)	0.34	mSMAMedian_avg (ACC)	3.41***	yAxisAvgMedian_avg (ACC)	0.35
	zAxisAvgMax_avg (ACC)	2.71**	system_avg (APP)	0.32	mSMAMean_avg (ACC)	3.41***	mMax_min (ACC)	0.33
	video_players_avg (APP)	2.70**	zAxisAvgMax_avg (ACC)	0.32	mNewStd_avg (ACC)	3.15***	mMedian_avg (ACC)	0.33
alone vs. lgroup	system_min (APP)	5.83****	mMin_min (ACC)	0.49	mMean_avg (ACC)	8.61****	mMean_avg (ACC)	0.56
	anglezMax_min (ACC)	5.64****	zAxisAvgStd_min (ACC)	0.49	mSMAMean_avg (ACC)	8.60****	mSMAMean_avg (ACC)	0.56
	anglezMedian_min (ACC)	5.64****	anglezMedian_min (ACC)	0.48	mMedian_max (ACC)	8.44****	mMedian_max (ACC)	0.55
	angleyMean_min (ACC)	5.55****	anglezMax_min (ACC)	0.47	mSMAMedian_max (ACC)	8.40****	mSMAMedian_avg (ACC)	0.55
	angleyMin_min (ACC)	5.54****	zAxisAvgMean_min (ACC)	0.46	mMax_avg (ACC)	8.40****	mMax_avg (ACC)	0.54
people_{three}				people_{two}				
alone vs. sgroup	mSMAMedian_max (ACC)	3.81***	zAxisAvgMedian_max (ACC)	0.41	yAxisAvgMin_max (ACC)	4.05****	dmListMean_max (ACC)	0.37
	mMedian_max (ACC)	3.81***	xAxisAvgStd_max (ACC)	0.41	xAxisAvgStd_avg (ACC)	3.76***	xAxisAvgStd_avg (ACC)	0.34
	anglezStd_max (ACC)	3.79***	xAxisAvgMedian_min (ACC)	0.40	angleStd_avg (ACC)	3.72***	angleStd_avg (ACC)	0.33
	xAxisAvgStd_max (ACC)	3.63***	xAxisAvgMean_max (ACC)	0.40	xAxisAvgStd_max (ACC)	3.65***	dmListMedian_max (ACC)	0.33
	mMean_max (ACC)	3.60***	anglezMax_max (ACC)	0.38	yAxisAvgMax_avg (ACC)	3.55***	angleMin_max (ACC)	0.32
sgroup vs. lgroup	yAxisAvgMin_min (ACC)	5.30****	mNewStd_avg (ACC)	0.43	anglezMax_min (ACC)	6.82****	mMin_min (ACC)	0.46
	yAxisAvgMin_avg (ACC)	5.11****	mMean_avg (ACC)	0.42	anglezMedian_min (ACC)	6.82****	zAxisAvgStd_min (ACC)	0.46
	yAxisAvgMean_min (ACC)	5.10****	mSMAMean_avg (ACC)	0.42	angleyMean_min (ACC)	6.64****	anglezMedian_min (ACC)	0.44
	yAxisAvgMedian_min (ACC)	5.07****	mSMAMax_avg (ACC)	0.42	anglezMean_min (ACC)	6.49****	dmListStd_min (ACC)	0.43
	yAxisAvgMax_min (ACC)	4.18****	mMax_avg (ACC)	0.41	anglezMax_min (ACC)	6.42****	zAxisAvgMean_min (ACC)	0.43
alone vs. lgroup	yAxisAvgMin_min (ACC)	7.44****	xAxisAvgStd_max (ACC)	0.77	food_and_drink (APP)	4.55****	yAxisAvgMax_avg (ACC)	0.43
	zAxisAvgMin_min (ACC)	6.76****	zAxisAvgMedian_max (ACC)	0.77	food_and_drink (APP)	4.54****	zAxisAvgMin_avg (ACC)	0.33
	xAxisAvgMin_min (ACC)	6.63****	anglezMax_max (ACC)	0.75	food_and_drink (APP)	4.55****	proximityRecord_avg (PRO)	0.29
	yAxisAvgMedian_min (ACC)	6.54****	anglezMin_max (ACC)	0.75	zAxisAvgMin_avg (ACC)	4.17****	yAxisAvgStd_avg (ACC)	0.27
	yAxisAvgMean_min (ACC)	6.31****	mMedian_avg (ACC)	0.73	zAxisAvgMean_avg (ACC)	3.35****	angleStd_avg (ACC)	0.27
partner_{two}				friends_{two}				
-	-	-	-	mMean_avg (ACC)	8.10****	mMean_avg (ACC)	0.52	
-	-	-	-	mSMAMean_avg (ACC)	8.08***	mSMAMean_avg (ACC)	0.52	
-	-	-	-	mMedian_max (ACC)	7.97****	mMedian_avg (ACC)	0.50	
-	-	-	-	mSMAMedian_max (ACC)	7.94****	mSMAMedian_avg (ACC)	0.50	
-	-	-	-	mMax_avg (ACC)	7.89****	yAxisAvgStd_max (ACC)	0.50	

features had Cohen's-d values above medium effect size, and the highest ones were closer to large effect size, meaning that these accelerometer features could discriminate between alone and lgroup social contexts.

For two-class social contexts, $people_{two}$, $family_{two}$, and $friends_{two}$, all features in the top five for both t-statistic and Cohen's-d were from the accelerometer. Further, only $friends_{two}$ had features with Cohen's-d above medium effect size among all four two-class social contexts. However, for $partner_{two}$, several features from application usage (food and drink app usage) were among the top five for t-statistics. In addition, a feature from the proximity sensor had a Cohen's-d of 0.29, which is above a small effect size. In summary, results from the statistical analysis suggest that accelerometer features could be informative of the group dynamic for all the social contexts. In addition, for social contexts related to partner/spouse, app usage behavior and proximity sensors could be informative. Moreover, bluetooth sensors had high statistical significance in discriminating social contexts related to family members.

Table 3.5: Mean (\bar{A}) and Standard Deviation (A_σ) of inference accuracies and the mean area under the curve of the receiver operator characteristic curve (AUC), calculated from 10 iterations, using five different models, for two-class and three-class tasks, with attributes such as family, friends/colleagues, spouse/partner, and alone. Results are presented as: \bar{A} (A_σ), AUC

	Target Variable	Random Forest	XG Boost	Ada Boost	Gradient Boost	Naive Bayes
two-class	baseline	50.0 (0.0), 50.0	50.0 (0.0), 50.0	50.0 (0.0), 50.0	50.0 (0.0), 50.0	50.0 (0.0), 50.0
	family _{two}	86.1 (3.1), 84.8	82.6 (3.4), 73.2	82.5 (4.2), 71.7	83.2 (3.7), 74.2	65.6 (6.9), 67.6
	partner _{two}	87.4 (2.6), 82.6	84.6 (4.6), 74.7	83.5 (3.8), 69.3	84.7 (5.1), 78.8	68.6 (8.2), 64.2
	friends _{two}	80.1 (2.9), 81.3	78.3 (4.5), 75.2	78.0 (4.1), 70.4	78.7 (3.7), 73.7	64.0 (4.2), 61.4
	people _{two}	83.3 (3.2), 79.2	84.1 (3.1), 75.2	82.7 (4.3), 79.5	84.2 (2.8), 76.9	72.3 (4.7), 67.3
three-class	baseline	33.3 (0.0), 50.0	33.3 (0.0), 50.0	33.3 (0.0), 50.0	33.3 (0.0), 50.0	33.3 (0.0), 50.0
	family _{three}	85.9 (2.2), 80.5	81.4 (3.1), 73.2	73.9 (3.7), 63.2	83.0 (2.6), 70.2	63.1 (4.2), 61.8
	friends _{three}	76.7 (2.3), 78.2	77.1 (3.6), 70.1	71.0 (3.1), 68.8	77.6 (2.8), 72.3	62.2 (5.1), 63.5
	people _{three}	78.3 (2.1), 75.3	71.2 (2.6), 69.8	67.1 (3.8), 67.8	73.8 (3.2), 73.1	57.9 (6.7), 67.2

3.6 Social Context Inference

3.6.1 Two-Class and Three-Class Social Context Inference (RQ2)

In this section, we use all the available smartphone sensing features and implement seven social context tasks, using features defined in Section 3.3.3 as target variables. The tasks include four two-class inference tasks and three three-class inference tasks: (1) family_{two}, (2) partner_{two}, (3) friends_{two}, (4) people_{two}, (5) family_{three}, (6) friends_{three}, and (7) people_{three}. In this phase, we used sci-kit learn [391] and keras [91] frameworks together with Python and conducted experiments with several model types: (1) Random Forest Classifier [117], (2) Naive Bayes [432], (3) Gradient Boosting [365], (4) XGBoost [87], and (5) AdaBoost [457]. These models were chosen by considering the tabular nature of the dataset, the interpretability of results, and the small size of the dataset. In addition, we used the leave k-participants out strategy (k = 20) when conducting experiments, where testing and training splits did not have data from the same user, hence avoiding possible biases in experiments. Further, similar to recent ubicomp studies [29, 330, 252], we used the Synthetic Minority Over-sampling Technique (SMOTE) [86] to obtain training sets for each inference task. As recommended by Chawla et al. [86], when and where necessary, we under-sampled the majority class/classes to match over-sampled minority class/classes to create balanced datasets, hence not over-sampling unnecessarily beyond doubling the minority class size. In addition, we also calculated the area under the curve (AUC) (for three-class inferences, one vs. the rest technique, using macro averaging) using the receiver operator characteristics (ROC) curves. All experiments were repeated for ten iterations. We report the mean and standard deviation of accuracies and mean of AUC using results from the ten iterations.

Table 3.5 summarizes the results of the experiments. All the two-class inference tasks achieved accuracies over 80%. Moreover, all the three-class inferences achieved accuracies over 75%. When considering model types, Random Forest classifiers performed the best across five out of the seven inference tasks (family_{two}, partner_{two}, friends_{two}, family_{three}, and people_{three}) and Gradient Boosting had higher accuracies for two inference tasks (people_{two} and friends_{three}). Generally, all models included in the chapter, except for Naive Bayes, performed reasonably well. Further, low standard deviation values suggest that regardless of the samples used for training and testing, the models generalized reasonably well. AUC scores followed a similar trend as the accuracy. These results suggest that passive mobile sensing features could be used to infer both two-class and three-class social contexts related to alcohol consumption, with reasonable performance.

3.6.2 Social Context Inference for Different Sensors (RQ2)

Prior work in mobile sensing has argued for multiple inference models for the same inference task, in the case of sensor failure [576, 330, 454]. For instance, during a weekend night, young adults could be concerned for the battery life of their phone, and could turn-off bluetooth, wifi, and location sensors that drain the battery faster. In such cases, having separate inference models that use different data sources to infer the same target attribute could be beneficial. In addition, prior work has segregated passive sensing modalities into Continuous Sensing (using embedded sensors in the smartphone) and Interaction Sensing (sensing the users' phone usage and interaction behavior) [330]. Considering these aspects, we conducted experiments for different feature groups based on the sensing modality (accelerometer, applications, battery, bluetooth, proximity, location, screen, and wifi) and the following feature group combinations that are meaningful in the context of drinking and young adults [325]:

- **Continuous Sensing (ConSen):** These sensing modalities use embedded sensors to capture context. Examples are accelerometer, battery, bluetooth, proximity, location, and wifi. ConSen contains features from all these sensing modalities, and this feature group combination can measure the capability of the smartphone in inferring the social context of drinking, even if the user does not necessarily use the smartphone, because the considered sensing modalities sample data regardless of the phone usage behavior.
- **Interaction Sensing (IntSen):** These sensing modalities capture phone usage and interaction behavior. Examples include screen events and application usage. In addition, these sensing modalities do not fail often because there is no straightforward way for users to turn off interaction sensing modalities. Furthermore, these sensing modalities consume far less power compared to continuous sensing. In this context, this feature group combination could measure the capability of a smartphone to infer the social context of drinking, based on the way young adults use and interact with the smartphone.

For the two above-mentioned feature groups, we conducted experiments using the same procedure as given in Section 3.6.2. Even though we got results for all models, we only present results for random forest classifiers in Table 3.6 because they output feature importance values which are useful to interpret results in Section 3.6.3, and they provide the results with highest accuracy and AUC values for a majority of inference tasks. Even though the accuracies were well above baselines for both two-class and three-class inference tasks, the lowest accuracies were recorded for SCR. This could either be because of the far too small dimensionality (only three features) or because the features were less informative. For the inference of social context partner_{two}, APP provided the highest accuracy of 82.92% followed by ACC which provided an accuracy of 81.21%. This suggests that the app usage behavior during drinking events is informative of whether participants are with a partner/spouse or not. This could also be related to prior work regarding *partner phubbing* [434, 92] that could lead to relationship dissatisfaction and disappointment. People might try to avoid phubbing (hence use the phone less/differently than normal) when they are with their partner/spouse. Furthermore, except for this inference, for all other social context inferences, the highest accuracies were obtained using ACC (in the range of 71.52% to 83.33%). This suggests that physical activity levels and movement dynamics around drinking events could be used to infer social contexts such as family (family_{two} and family_{three}), friends (family_{two} and family_{three}), and people (people_{two} and people_{three}). In addition, results from the AUC followed a similar trend to accuracies. For two-class inferences, except for SCR, all other modalities reported AUC scores above 70%. However, for three-class inferences, except for ACC, all other modalities reported AUC scores below 70%. Further, except for SCR, standard deviation scores were reasonably low for

Table 3.6: Social Context Inference accuracy breakdown for sensor type based feature groups and feature group combinations using Random Forest classifiers. Both the mean (\bar{A}) and standard deviation (A_σ) of accuracies from cross-validation are reported in addition to the mean area under the curve (AUC) from receiver operating characteristics graph (ROC)

Feature Group (# of features)	two-class				three-class		
	family _{two} $\bar{A}(A_\sigma)$, AUC	partner _{two} $\bar{A}(A_\sigma)$, AUC	friends _{two} $\bar{A}(A_\sigma)$, AUC	people _{two} $\bar{A}(A_\sigma)$, AUC	family _{three} $\bar{A}(A_\sigma)$, AUC	friends _{three} $\bar{A}(A_\sigma)$, AUC	people _{three} $\bar{A}(A_\sigma)$, AUC
Baseline	50.0 (0.0), 50.0	50.0 (0.0), 50.0	50.0 (0.0), 50.0	50.0 (0.0), 50.0	33.3 (0.0), 50.0	33.3 (0.0), 50.0	33.3 (0.0), 50.0
ACC (150)	83.3 (2.4), 80.2	81.2 (3.1), 79.6	74.9 (3.0), 72.5	81.1 (3.1), 81.4	82.6 (1.9), 78.5	71.5 (2.5), 70.1	72.4 (2.7), 72.0
APP (105)	82.9 (3.5), 80.7	82.9 (3.0), 79.2	74.3 (2.4), 76.7	80.5 (2.5), 81.1	81.4 (2.4), 81.5	69.5 (3.1), 69.1	71.9 (2.3), 70.2
BAT (36)	78.7 (2.8), 76.7	77.5 (3.6), 77.3	71.5 (3.0), 73.6	78.1 (3.3), 72.1	77.5 (2.7), 75.6	66.1 (3.1), 67.8	68.1 (2.6), 68.4
BLU (27)	74.6 (2.8), 72.1	75.5 (3.6), 73.8	69.0 (2.8), 70.3	74.0 (3.1), 73.1	69.3 (2.9), 68.8	59.4 (2.7), 61.4	60.5 (2.8), 66.4
PRO (18)	74.1 (3.1), 71.9	75.6 (2.3), 74.5	69.9 (2.8), 68.2	75.8 (2.7), 76.8	70.4 (2.2), 73.2	59.1 (3.8), 61.9	60.7 (2.4), 62.9
LOC (39)	79.2 (3.0), 77.1	78.9 (2.7), 78.2	74.1 (3.2), 76.1	77.6 (2.6), 76.9	77.2 (2.6), 76.4	67.0 (3.3), 69.1	69.5 (2.9), 68.7
SCR (3)	68.3 (4.5), 61.1	69.7 (4.7), 60.3	64.5 (4.6), 62.8	71.9 (3.1), 67.2	62.2 (5.6), 60.7	54.1 (4.3), 55.2	54.8 (5.5), 56.1
WIF (36)	77.5 (2.9), 78.1	77.1 (2.9), 76.9	68.8 (3.7), 70.2	75.3 (3.9), 76.1	73.1 (1.9), 73.0	61.6 (2.9), 63.6	64.5 (2.8), 67.1
ConSen (306)	85.7 (2.6), 82.1	86.8 (2.9), 80.8	79.5 (3.2), 80.2	82.9 (2.1), 81.4	85.3 (1.9), 76.8	76.7 (2.7), 76.9	77.9 (3.0), 76.1
IntSen (96)	83.3 (2.0), 80.1	83.1 (2.5), 81.7	76.5 (2.9), 76.3	81.6 (2.5), 81.5	82.3 (2.7), 78.2	71.4 (2.7), 71.2	73.2 (2.7), 72.7
ALL (402)	86.1 (3.1), 84.8	87.4 (2.6), 82.6	80.1 (2.9), 81.3	83.3 (3.2), 79.2	85.9 (2.2), 80.5	76.7 (2.3), 78.2	78.3 (2.1), 75.3

all the other inferences, suggesting that inference results hold regardless of the training and testing splits. High standard deviations for SCR could be because of the low number of features, which was also reflected in low accuracies and AUC scores.

Feature group combinations ConSen and IntSen provided similar accuracies for all inference tasks even though ConSen achieved slightly better than IntSen for each inference. While ConSen outperformed ACC and APP for all the inferences, IntSen had slightly lower accuracies for social contexts family_{three} (82.36%) and friend_{three} (71.44%) as compared to ACC, which had accuracies 82.60% and 71.52% for the respective inferences. Standard deviation scores for both ConSen and IntSen were low. In addition, AUC scores too were above 70% for all cases, which is a reasonable result. Finally, the results suggest that IntSen could provide reasonably high accuracies as compared to ConSen in case of sensor failure, and in the worst-case scenario, ACC provides fair accuracies for all the inference tasks, which is satisfactory given that it is just one sensing modality.

3.6.3 Feature Importance for Social Context Inferences (RQ2)

In Figure 3.9, we show the top twenty feature importance values for each inference presented in Section 6.2. These values were captured from the output of trained random forest models when using all features. We obtained feature importance values for all features in each iteration, and report the mean value for each feature. The sensor modality that was present throughout all seven inferences was the accelerometer (ACC). This is congruent with the results presented in the statistical analysis (Section 3.5). This suggests that physical activity levels and phone motion dynamics could help infer different types of social contexts of drinking occasions. This makes sense for certain situations because it is highly unlikely that a person would drink and dance alone on a weekend night, while this might happen when people are in larger groups (with both family and friends).

The second most common modality across all inferences is location (LOC). Features that capture the speed of the phone (speedMedian_avg, speedMax_avg, etc.), accuracy of the signal (accuracyMin_max, accuracyMean_avg, etc.), and signal type and strength (signalGps_max, signalNetwork_avg, etc.) are present across all inferences. Especially, for both two-class and three-class inferences regarding family, location features regarding GPS signal strength and speed filled up a majority of the top five features. This suggests that location-related features have captured certain differences with regard to group dynamics in the social context of family. Even though interaction sensing modalities (APP, SCR) were

Chapter 3. Examining the Social Context of Alcohol Drinking of Young Adults with Smartphone Sensing

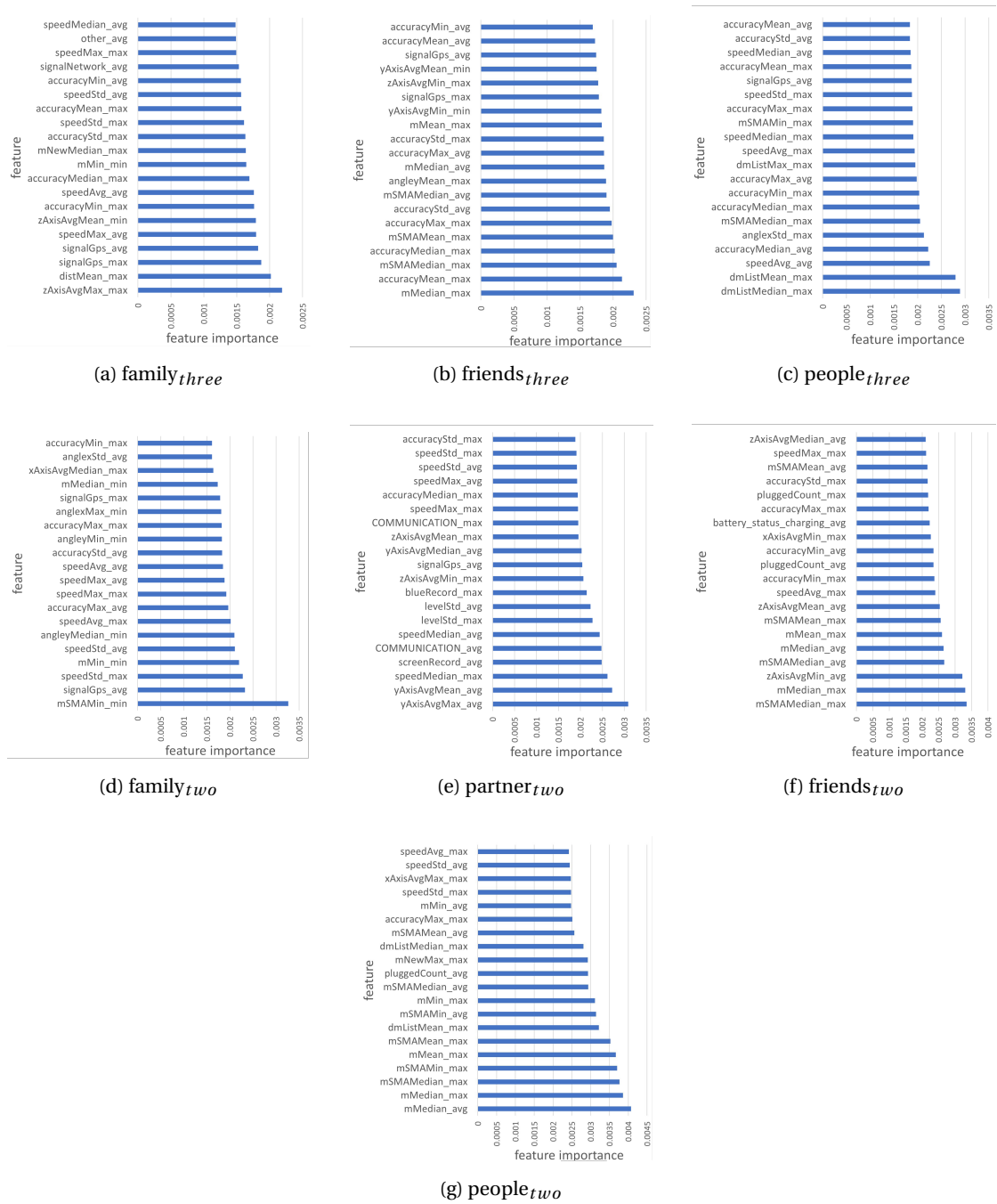


Figure 3.9: Feature importance values from random forest classifiers with all features, for different social contexts.

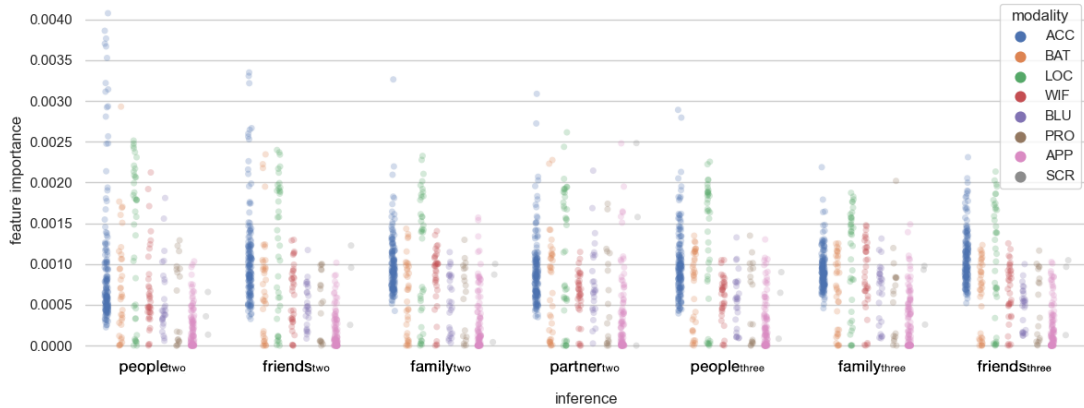


Figure 3.10: Feature importance value distributions for different sensing modalities in different inferences when using all features

not present among all the social contexts, partner_{two} had several features (COMMUNICATION_avg, COMMUNICATION_max, etc.) regarding communication app usage (e.g. viber, whatsapp, messenger, etc.) and also screen usage (screenRecord_avg). Given interaction sensing modalities capture the phone usage behavior, this suggests that people use their phone differently when they are drinking alcohol with their partner as opposed to not being with him or her.

In Figure 3.10, we plot a distribution of feature importance values for all social context inferences, grouped by different sensing modalities. This provides an overview of the informativeness of sensing modalities in making inferences. The most sparse distribution across all inferences came from the ACC, for the social context people_{two} . Overall, the accelerometer produced the most informative feature, for all seven social contexts. Location features had comparatively high values for all seven social contexts. Even though location features were not among the highest for any inference, mean feature importance for location modalities was even higher than for accelerometer features (because the location feature distribution is negatively skewed). In addition, except for WIF, all other modalities had comparatively sparse and wider distributions for the context partner_{two} . To sum up, the takeaways from this analysis are: accelerometer features (ACC) were informative for all inferences, location features (LOC) were generally informative across all inferences too, application usage (APP) and screen usage (SCR) features (interaction sensing) were informative for partner_{two} while not being comparatively informative for other inferences, and except for Wifi features (WIF), all other features had wider distributions for partner_{two} .

3.6.4 Effect of Varying Group Sizes (RQ2)

In the previous analyses, we considered group dynamics as follows: (a) two-class social contexts - *without vs. with one/more people* and (b) three-class social contexts - *without vs. with one person vs. with two/more people*. Hence, while the two-class inference mostly relates to the absence of a particular type of people, the three classes effectively tried to infer the presence of groups of varying sizes (with one person, with two/more people). If we consider the three-class inferences, *with one person* means that it is a group of two people including the participant, and *with two/more people* means that the group has a minimum size of three people, including the participant; hence both classes

t

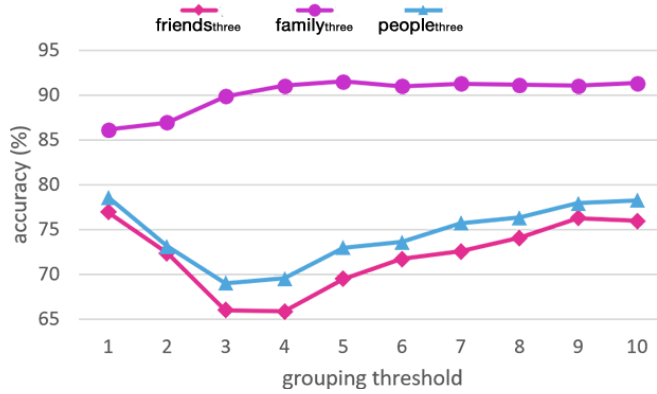


Figure 3.11: Three-Class Social Context Inference Accuracies for Different Grouping Thresholds

capture different group sizes, with the former being a small group, and the latter being a larger group in comparison. Given that there is no gold standard regarding the definition of the size of the drinking group (as highlighted in Section 3.2), in this section, we aim to change the size of these two groups by changing the threshold called *grouping threshold*, which was always equal to *one* in previous sections (e.g., without vs. with *one/less* people vs. with *two/more* people), for three-class inferences. To this aim, we increase the value of the grouping threshold from one to ten, to investigate how it affects the inference accuracy. One and ten were chosen as the highest and lowest thresholds because those were the highest and lowest values available in self-reports to define three classes.

We conducted the evaluation with the three three-class inferences using the same approach mentioned in Section 3.6.1, and the results are summarized in Figure 3.11. For $friend_{three}$ and $people_{three}$, inference accuracies decreased when increasing the grouping threshold, meaning that the model was not good at discerning the three classes when the threshold was around three (alone vs. with three/less people vs. with four/more people) and four (alone vs. with four/less people vs. with five/more people). However, when increasing the threshold further, the accuracies increased back to the same level as when the threshold was equal to one. What this means is that the random forest classifier is not performing well when the small and large groups are defined by thresholds in the range of three to five. This result is not surprising because any nightlife-related activities available for a small group of people (they might find a table to fit altogether in a pub or a restaurant, they might easily travel with a cab, they might all gather in a living room) would result in a large heterogeneity of sensor data as compared to a larger group (e.g., ten or more people). This is because of the differences in behavior when people are in large groups, as opposed to small groups. Consequently, this would result in a lower inference accuracy when social contexts with three, four, or five people are in both the small group and the large group classes of the three-class inference. Consider an example where the grouping threshold is three, where samples with a group of three people would fall into the small group, and samples with a group of four or more people would be included in the large group of the three-class inference.

According to the distribution given in Figure 3.3, for the variable $friends_{three}$, when 114 samples (group of 3) and 105 samples (group of 4) fall into small and large groups in the inference, both classes have homogeneous sensor data, hence making it difficult for the model to discriminate between the classes. Conversely, the range of activities is smaller for larger groups due to its size, resulting in a lower heterogeneity of sensor data within groups, and, consequently, higher inference accuracy for higher grouping thresholds. For example, consider the grouping threshold of ten, where the small group

would have ten or fewer people, and the large group would have eleven or more people. According to Figure 3.3, for the variable $people_{three}$, there are 221 samples of eleven or more (clearly a large group), and over 300 samples of groups with three to ten people, with a majority of data coming from small group sizes such as three (135 samples), four (93 samples), and five (119 samples). This leads to heterogeneous data between small and large groups because the small group consists of data predominantly from groups of three, four, or five, and the large group is predominantly containing groups of 10+ people. This makes it easier for the model to discriminate between the three classes, hence leading to higher accuracies. On the other hand, for $family_{three}$, increasing the threshold had the opposite effect, and increased the performance of the models. Again, this might be explained by the lower diversity of choices of activities and contexts to be sensed in family contexts, which tend to be highly routinized. Finally, results suggest that models performed reasonably well for all $family_{three}$ inferences regardless of the grouping threshold. In addition, except for grouping thresholds from three to five, for all other thresholds, $friends_{three}$ and $people_{three}$ showed reasonable performance with accuracies over 70%. Hence, according to this analysis, having different grouping thresholds seems a valid design choice depending on the application and the use case.

3.7 Discussion

Features. It is worth noting that for modalities such as ACC and LOC, we generated simple statistical features that do not need extensive processing of the dataset. If we consider the ACC, while features proved to be informative in inferring different social contexts, the only set of features we used are statistical features from the three axes, angles between the gravity vector and axes, and aggregate features that combine the values of three-axes (Section 3.3.1). It is also worth noting that these features are less interpretable in the context of alcohol consumption. For example, the feature $mMedian_max$ had the highest feature importance value for $friends_{three}$, as shown in Figure 3.9a. While this feature represents the overall acceleration of the phone at a time period closer to the drinking event, it is difficult to interpret it compared to more interpretable features such as step count or activity type. If such features were derived using the accelerometer data, the interpretation could have been much simpler. However, we were not able to derive them due to limitations in the original dataset (sampling frequency, lack of gyroscope data, etc.). Future work could consider using low-power consuming libraries such as Google Activity Recognition API to obtain activity types and native step counters available in modern smartphones to obtain step counts, hence obtaining more interpretable features. In addition, researchers could also look into using other sensing modalities such as ambient light sensors, typing and touch events, and notification clicking behaviors.

Ethical Considerations. The goal of this chapter is to support public health research. Hence, it is essential to be aware of ethical implications. For public health, the inferences done in this work are anonymous in the sense that no identities of individuals are inferred when inferring social contexts. However, certain social contexts such as 'being with a partner' could be more sensitive, because identifying the presence of such people could potentially reveal sensitive information about them, even though they might not have agreed to have their location indirectly reported. Given that social context is relational, it is critical that during data collection, social companions (friends, family, etc.) agree that their presence is reported (even as an aggregate). Future studies should consider these aspects. Furthermore, for future interactive health systems that would be used by individuals and their health providers, it is fundamental to have clarity on who could access inferred data regarding social contexts, given their sensitive nature. Further, running social context inferences on devices, rather than on servers, would help preserve the privacy of users and others interacting with them. More generally,

participants' respect for privacy and well-being should be the guiding lights of any future design of mobile health systems regarding alcohol consumption.

Limitations and Future Work. We prepared the drinking event level dataset (in Section 3.3.2) without assuming any relationship between two drinking events that occur consecutively, hence, we considered alcohol drinking events to be independent of each other. However, in reality, there could be a relationship between the drinking events of the same person during the same night. Understanding such relationships is a complex problem, and it needs further examination. Another limitation of our work is that it does not capture complex relationships among family members. For example, young adults might prefer drinking with their brother or same-age members of the family, whereas they might not feel comfortable drinking with their parents. In addition, the perception of parents and other family members could differ significantly, and it could affect the drinking behavior in the vicinity of family members. Furthermore, the partner's/spouse's perception of alcohol consumption is another variable that was not captured during this chapter. These aspects need further investigation. In addition, it is worthwhile to note that, inferring the social context of drinking does not directly help overcome health problems. This is not the intention of this work, as it would oversimplify the problem. However, inferring the social context of drinking would assist or complement other inferences such as drinking occasions, drinking nights, drink vs. drunk in ubicomp and alcohol research [454, 30, 29, 405, 404, 194, 141]. In this respect, the inference of social context might help to provide meaningful and context-aware interventions that might decrease the amounts consumed, and, as a consequence, less adverse alcohol-related consequences. The design of such interventions is beyond the scope of the chapter.

Another important aspect is the choice of time windows for aggregation and matching phases. Even though we presented results for the dataset obtained with a ten-minute time window for aggregation and a one-hour time window for matching, we conducted evaluations with different time windows. We obtained the best results using these time windows, and hence, considering space limitations, we only presented results for these windows. It is worth noting that the time window would affect the number of self-reports included in the study. For instance, if the matching time window is two hours, we need to discard all self-reports from 8.00 pm to 9.00 pm because we would not have enough sensor data for reports done between those time windows. The same applies to drinking events done between 3.00 am and 4.00 am. Further, it is worth noting that, regardless of the time window and the resulting dataset size, we obtained inference results comparable to the ones we presented in Table 3.5, with differences of the range 0.4% (best-case scenario) to 12% (worst-case scenario).

An important topic for future work is the drinking motives of young adults. As we discussed in Section 3.2, drinking motives could be the primary driving factor why young adults choose specific social contexts to drink [258, 257]. Hence, examining the associations between such drinking motives and smartphone sensing data could further advocate the idea of building holistic mobile health systems that consider not only alcohol consumption but also other factors associated with the event. Furthermore, even though there were multiple comparisons in the statistical analysis, we did not use Bonferroni correction for p-values [531]. Hence, the results with p-values should be interpreted with caution.

Importance of Diversity-Awareness. The drinking behavior of people differ significantly depending on age, sex, drinking culture, beverage preferences, as well as how people perceive drinking alcohol [192, 30, 454]. For example, in some Asian countries, drinking alcohol might not be socially accepted while it is a societal norm in Europe and North America [483, 34]. Hence, it is worth pointing out that this study regarding the drinking behavior of young adults in Switzerland is exploratory, and the results cannot be directly assumed as being representative of the drinking behavior in other countries. Recent work has

highlighted the importance of considering diversity awareness when building social platforms using machine learning models and mobile sensing data [247, 458].

3.8 Conclusion

In this chapter, we examined the weekend drinking behavior of 241 young adults in Switzerland using self-reports and passive smartphone sensing data. Our work emphasized the importance of understanding the social context of drinking, to obtain a holistic view regarding alcohol consumption behavior. With multiple statistical analyses, we show that features from modalities such as accelerometer, location, bluetooth, and application usage could be informative about social contexts of drinking. In addition, we define and evaluate seven inference tasks obtaining accuracies of the range 75%-86% in two-class and three-class tasks, showing the feasibility of using smartphone sensing to detect social contexts of drinking occasions. We believe these findings could be useful for ubicomp and alcohol epidemiology researchers in implementing future mobile health systems with interventions and feedback mechanisms.

4 Inferring the Food Consumption Level of College Students with Smartphone Sensing

While the characterization of food consumption level has been extensively studied in nutrition and psychology research, advancements in passive smartphone sensing have not been fully utilized to complement mobile food diaries in characterizing food consumption levels. In this chapter, we used the MEX dataset described in Chapter 2 regarding the holistic food consumption behavior of 84 college students in Mexico. The dataset was collected using a mobile application combining passive smartphone sensing and self-reports. We showed that factors such as sociability and activity types and levels have an association with food consumption levels. Finally, we defined and assessed a novel ubicomp task by using machine learning techniques to infer self-perceived food consumption level (overeating, undereating, eating as usual) with an accuracy of 87.81% in a 3-class classification task by using passive smartphone sensing and self-report data. Furthermore, we show that an accuracy of 83.49% can be achieved for the same classification task by using only smartphone sensing data and time of eating, which is an encouraging step towards building context-aware mobile food diaries and making food diary-based apps less tedious for users. The material of this chapter was originally published in [330].

4.1 Introduction

Many young adults show a tendency to adopt unhealthy eating practices during college years when they undergo significant lifestyle changes such as leaving home, meeting new friends, starting a career, and developing relationships [386, 416]. Even though young adults are relatively healthy compared to other older populations, unhealthy eating habits at this age could lead to adverse health outcomes such as cardiovascular diseases, overweight conditions, and obesity in the long term [182, 386, 15]. Due to these reasons, researchers in nutrition, behavioral science, and psychology are extensively studying causes and contexts of food consumption, especially among college students [221, 416, 582, 558]. Moreover, prior research in these domains has linked factors such as social context [213], eating location [164], availability and types of food [504], and psychological aspects [211] to food consumption behavior. With increasing smartphone coverage among young adults and the availability of a plethora of mobile health (mHealth) applications [345], smartphones have become a ubiquitous tool that can help young adults adhere to healthier food consumption practices [325].

The most common use case in mHealth apps is fitness tracking, where such apps keep track of daily activity levels *passively* by providing insights to users in terms of step count, activity levels, and activity types [185]. "Food and Nutrition" is another major category of mHealth applications [345], and many

widely used commercial apps such as MyFitnessPal [360], Lose It! [297], Apple Health [1], and Samsung Health [451] allow users to keep their food intake as a mobile food diary, providing basic statistical insights so people can adhere to healthier eating patterns. However, in their current state, these applications do not yet make full use of smartphone sensing capabilities to provide additional insights (with the notable exception of the camera, which is used to photograph food), even though food consumption level is directly related to many aspects that can be sensed passively such as stress [300], mood [285, 418], activities [525, 146], and sociability [357, 61]. Even though attention has been given to visually identifying food types via mobile apps [448], characterizing meal and snack eating behavior using smartphone sensing and self-reports [55], and identifying eating and overeating events using wearable sensing [588], relatively less research has been conducted to leverage smartphone sensing capabilities to obtain data from people to analyze food consumption levels [337].

To our knowledge, while food intake recognition has been studied in ubicomp research [55], the specific overeating phenomenon has not been studied using passive smartphone sensing and self-report datasets. Using such a rich combination of data sources allows us to analyze the eating behavior of college students using knowledge from nutrition and mobile sensing research by associating food consumption levels to aspects such as mobile app usage, location, activity levels, sociability, and food types. This approach allows for comparisons with findings about self-perceived food consumption levels in prior nutrition and behavioral science research (which validates some of the observed trends), and also provides novel insights regarding techniques to build mobile food journaling systems that leverage passive sensing to identify behaviors of college students associated to overeating, and to provide them with valuable insights and interventions regarding their food consumption. In this chapter, we address the following research questions.

RQ1: What behaviors and contextual patterns around food consumption levels can be observed by analyzing everyday eating episodes of a group of college students obtained via passive smartphone sensing and self-reports?

RQ2: Can self-perceived food consumption level be inferred using contextual data obtained through a mobile application?

By addressing the above research questions, this chapter makes the following contributions.

Contribution 01: As described in Chapter 2, we conducted a new mobile data collection campaign in Mexico with 84 college students to capture their eating behavior, including food consumption levels (eating as usual, overeating, undereating). This is important given studies of populations in Latin America are not common in ubicomp research. During our study, we collected 3278 self-reports, including 1911 meal reports and 1367 snack reports for two participant cohorts spanning 37 and 23 days, respectively. We conducted a descriptive data analysis of the collected dataset to show insights that confirm previous findings in nutrition and psychology research regarding food consumption levels. This conformity allows us to demonstrate the viability of using smartphone sensing to better characterize and understand details regarding eating episodes throughout longer time spans. For example, conforming to prior research, we show that factors such as sociability, stress, mood, and food category could affect the overeating behavior of people, and prior mobile sensing research has directly linked these aspects to passive sensing features.

Contribution 02: We defined and evaluated a novel ubicomp task by inferring eating more than usual ("overeating"), less than usual ("undereating"), and "as usual" episodes using passive smartphone

sensing and self-report data, with an accuracy of 87.81% in a 3-class classification task. We also show that an inference accuracy of 83.61% can be achieved for the same 3-class classification task even if food category-related features are not used in model training. Moreover, we show that the same inference task can be done with an accuracy of 83.49% by only using passive smartphone sensing features and time of eating. These results show that in the set of participants, food consumption level was not only driven by food type or category as assumed in traditional food journals and calorie-based calculations but also by other contextual factors associated with eating episodes. Moreover, results from this chapter illustrate the potential of using passive smartphone sensing alone or together with mobile food journals towards building context-aware food-related mobile systems.

4.2 Defining Food Consumption Level

In prior research, food consumption level has had both objective interpretations (nutrition science-based) [490, 561, 64] and subjective ones (nutrition and psychology-based) [442, 527, 153, 524, 561, 539], and there is no unique way to define it [210, 295]. The *objective* food consumption level attempts to capture the exact calorie consumption during eating episodes from a purely nutritional standpoint. In lab studies, calorie intake is pre-calculated before offering food to participants [64]. Under this objective interpretation, a person should eat only as much as is necessary to offset her/his caloric demands, and overeating occurs if the food intake exceeds this amount [210]. Many currently available mobile food diaries such as MyFitnessPal [360], Samsung Health [451], and other research studies [129] attempt to capture this attribute using self-reports by requesting the users to enter each food type and the amount they eat. Even though the target here is to capture the objective calorie intake, there is, by design, a subjective element because users self-report it, and it is known that people often fail to report the volume/weight of a dish accurately [561]. However, even if caloric intake is correctly reported and calculated, defining food consumption level as overeating and undereating according to this approach is complicated according to Herman et al. [210, 527], because it depends on a plethora of factors (a) individual factors such as metabolic rates, activity levels, age, gender, height, weight; and (b) measurement factors, i.e., the unit of calculation for overeating is usually caloric deficit per day (nutritionists often do it at the day, week, or meal/snack episode level). This implies that for the same person, eating the exact same amount of food on a more active day could be overeating on a slow day. Hence, the process gets more complex as factors add on, and it could get particularly difficult and inaccurate if overeating and undereating episodes are determined based on self-reports that reflect food types and volumes. In addition, a recent study by Jung et al. [234] emphasized how currently available mobile food logging systems can be troublesome to users because of the tedious manual data entry process, hence leading to low adoption rates.

Contrary to the objective view of food consumption level, nutrition researchers have also widely used *subjective* measures to capture food consumption levels of people by considering the psychology of food consumption [442, 295, 527, 524, 523, 504, 244, 528, 439], also known as self-perceived food consumption level. This view is primarily based on the idea that, when you ask people whether they overate or not, more often than not, the answer would be based on an eating episode level and the self-perceived amount of food they have eaten [210]. This measure has often been used as a proxy for the actual amount of food people have eaten. Field et al. [153] showed that self-perceived food consumption levels can be similar to real food consumption levels, and these self-reports are valid to determine bulimic episodes in adolescents. Moreover, Williamson et al. [561] examined the relation between self-reported caloric intake (similar to self-reports regarding caloric intake in MyFitnessPal and Samsung Health) and self-perceived overeating, concluding that there is a positive relationship

between the two variables for all four groups of people they considered: (1) suffering from bulimia nervosa, (2) compulsive binge eaters, (3) obese, and (4) not having any of the three previous conditions. Due to the above-mentioned factors, many prior studies have used self-perceived food consumption level as a proxy to the objective food consumption level, although we are not aware of any study that establishes detailed guidelines of when self-reported subjective overeating and objective overeating coincide or not. Furthermore, prior work in nutrition research suggests that adverse behavioral and emotional effects of overeating arise not only after people eat an objectively large amount of food but even when people think that they have overeaten (self-perceived overeating) compared to their prior beliefs or current social context [295, 408, 409]. This is why many studies regarding psychology and eating behavior consider self-perceived food consumption level to be an important attribute, especially when considering eating as a holistic process to understand eating behavior [57, 515, 229].

Williamson et al. [561] captured overeating episodes by asking participants to report their perception of whether they overate (a binary choice). In a study by Ruddock and Hardman [442], self-perceived food consumption level was examined using a three-level coding system (eating more-than/less-than/as usual). Moreover, Vartanian et al. [527] used a five-point Likert scale (1-5) in their study regarding food consumption levels where 1, 3, and 5 corresponded to "ate much less than I normally eat", "ate similar to the amount I normally eat", and "ate much more than I normally eat" respectively. By *normal* or *as usual*, what these studies meant is in comparison to their past behavior, and how they perceive societal norms regarding normal food intake. Following this literature, in this chapter, we define self-perceived food consumption level as "*eating more than (overeating condition), less than (undereating condition), or roughly the same (as a usual condition) amount of food during an eating episode, in relation to the person's own estimated consumption, beliefs, and norms*". Hence, from here onwards in this chapter, we use the terms "food consumption level", "overeating", "eating as usual" or "undereating" to denote the self-perceived and self-reported attributes.

4.3 Related Work

4.3.1 Internal and Contextual Factors Affecting Overeating

If we consider food consumption level, prior research has shown several factors that affect it such as psychological aspects, sociability, activity levels and types, and food types [477, 205, 244, 389, 210, 556, 209]. A recent review discussed the complex relationship between psychology and food intake, describing that stress and mood affect food consumption behavior, and it can lead to both overeating or undereating depending on external or psychological stressors [477]. A recent article explained that stress can lead to a lack of appetite in the short term (due to the high levels of hormone epinephrine secreted by adrenaline glands) that would lead to undereating [559]. On the flip side, long-term exposure to stress can cause people to overeat with the initial expectation of overcoming stress and a bad mood that is driven by hormonal activity (due to another hormone released from the adrenaline gland called Cortisol, which increases appetite). Many other studies confirm these findings; some show that positive and calm affect, influenced by contextual factors (social gatherings, celebrations, partying), can lead to overeating in the short term [64, 72, 389], while other studies show that longer-term exposure to stress can lead to eating disorders [474, 97]. Further explaining the relationship between human psychology and overeating, Vartanian et al. [528] described that many people perceive the eating amount to be highly driven by internal factors such as hunger, feeling, and emotions rather than external factors such as social context. In another study, Bongers et al. [63] showed that overeating is not only driven by mood and negative affect but also external contextual factors. Studies also suggest

that people who overeat do not self-perceive it at the moment and usually feel it sometimes after consuming the food [210]. This could be due to the rapid nature of food consumption and the time the body takes to generate the sensation of being full [28, 449, 210]. Considering this fact, we designed our mobile surveys retrospectively rather than in-situ [463] as described in Section 2.1.2. In this chapter, we focus on using a mobile food diary to identify short-term food consumption levels with eating episode-level data.

4.3.2 Mobile Health Apps to Analyze Eating Behavior

Nutrition researchers have carried out mobile food diary-based studies to understand compliance among young adults and children [8, 450], claiming that there is potential for young people to independently record their food intake using photos. Most of the mobile food diary-based studies rely on users taking and labeling pictures and manually entering calorie levels [337]. This is a tedious process that can disengage people due to boredom [234]. Hence, studies say that the focus of modern food diaries has been to reduce the effort by users to enter details, such that the benefits of having such diaries outweigh the effort [20]. Following this, a direction pursued by mobile sensing and computer vision researchers is to identify food categories and estimated calorie intake with photos [230, 338, 65, 284]. However, most of these applications still lack the accuracy to be applicable in any unrestricted setting. As Min et al. studied in a survey [337], these studies have focused on food recognition and use heterogeneous data sources such as social media and recipe datasets. In contrast, inferring food consumption levels using contextual passive sensing data and self-reports is not a widely attempted task in food computing research due to its challenging nature, starting from obtaining relevant data with accurate ground truth.

Even though mobile food diaries have been studied extensively as stand-alone applications, food intake is not necessarily a stand-alone activity and is associated with factors such as psychology, activity levels, social context, and the general behavior of people as described in Section 4.3.1. Moreover, prior research highlights the importance of building diaries that are not only aware of the food intake but also the contextual, psychological, and social modeling of the eating episode [354, 114, 80]. Hence, our focus is to find associations between perceived food consumption levels and features from passive sensing and self-reports, using the intuition that representing an eating episode by merely the food intake and type is not enough, but all the contextual and individual cues captured via passive sensing features should be considered.

4.3.3 Mobile Sensing to Analyze Eating Behavior

The use of mobile sensing in food-related studies can be categorized into two main types: (1) detecting and identifying eating events and time of eating using contextual sensing and (2) after eating events are detected, identifying food types and characterizing eating events by identifying associations with user context, food consumption levels, and other attributes.

Detecting and identifying eating events has been primarily approached using wearable sensing and smartphone sensing. Rahman et al. [420] used wearable sensing to predict about-to-eat moments with a recall of 77%. Bedri et al. [44] studied a wearable system called EarBit to detect chewing moments, with 90.1% and 93% accuracies in lab settings and outside-lab settings, respectively. Thomaz et al. [506] used a wrist wearable to detect eating episodes, reporting F1-scores of the range 70%-76% in two lab-based experiments. In summary, these previous works attempt to identify the time of eating.

When the time of eating is detected, the next step is to characterize eating events [339]. Madan et al. [307] used mobile phone sensing to understand the food consumption of US university students for a period of nine months and concluded that healthy food consumption patterns are related to the health behavior of other people that they associate with (sociability). Seto et al. [469] used smartphone sensing to identify eating behavior using self-reports and passive sensing and concluded that the eating environment affects eating patterns. Biel et al. [55] deployed a mobile sensing application to track the eating behavior of 122 Swiss university students. They gathered over 4440 eating events and performed an eating occasion inference (meal vs. snack) task with an accuracy of 84% using data such as location, time of the day, and time since the last food intake. Vu et al. [535] laid out a vision for wearable food intake monitoring systems and emphasized the importance of identifying overeating moments. Zhang et al. [588] used a wrist wearable to detect overeating episodes based on the number of feeding gestures. They worked with the assumption that the higher the number of feeding gestures, the higher the calorie intake and objective food consumption level, and they showed correlations between objective calorie consumption and the number of feeding gestures. In this chapter, we consider eating events as holistic [57, 515, 55], and attempt to understand self-perceived food consumption levels using passive smartphone sensing and self-reports. We show that food consumption level is related not only to the type of food but also to sociability, mood, stress, and other behaviors, some of which can be captured or estimated using passive sensing features. In summary, this chapter differs from previous work in five aspects: (1) we use smartphone sensing combined with mobile food diaries; (2) we consider eating as a holistic event instead of just considering the type and amount of food; (3) we show that the self-perceived food consumption level can be inferred using smartphone sensing and self-reports; and (4) we provide preliminary evidence of using just *passive mobile sensing* features and time of eating, as informative features for the inference of food consumption level of college students.

4.3.4 Study Objective, Hypothesis, and Dataset

The primary objectives of this chapter were to investigate links between food consumption level and features derived using passive sensing and to leverage such links to automatically infer food consumption level, as summarized in Figure 4.1. Prior literature has shown that passive sensing features can be used to infer psychological and contextual aspects such as stress [468, 300], mood [285], activity types [55, 185, 451], sociability [51, 26, 197, 196], and food types [339, 469, 55]. In addition, a plethora of prior nutrition and behavioral science studies have linked the above aspects to food consumption levels [556, 210, 209, 528] as discussed in Section 4.3. Knowing that smartphone sensing features have shown correlations to certain attributes, which have also been connected to food consumption levels as shown in Figure 4.1, our objective is to leverage these relationships studied in prior literature to use passive sensing for inference of food consumption levels. Further, as mentioned in Chapter 2 under MEX dataset, similar to Vartanian et al. [527], the app asked users to indicate their food consumption level as (1) significantly less food than usual, (2) slightly less food than usual, (3) about the same as usual, (4) slightly more food than usual, and (5) significantly more food than usual. For our analysis, we define options 1 and 2 as "undereating", option three as "as usual", and options 4 and 5 as "overeating" in accordance with the definition we laid out in Section 4.2. With the MEX dataset, we used mobile sensing features described in Table A.3.

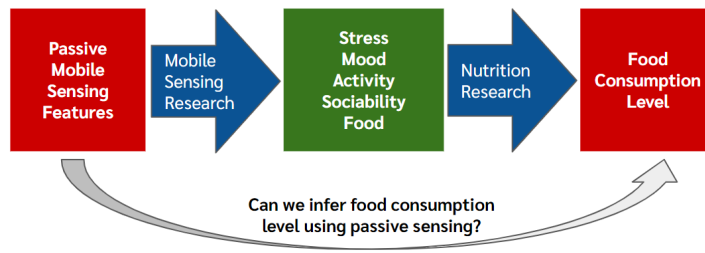


Figure 4.1: Objective of the Study

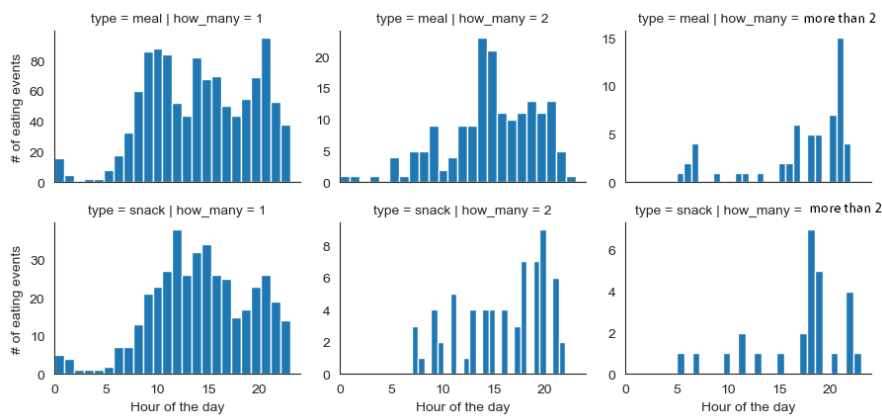


Figure 4.2: Bar charts for food intake reports with hour of the day in the x-axis and number of eating events in the y-axis. "how_many" is the self-reported value from the food intake questionnaire regarding how many eating episodes have occurred in the past four hours.

4.4 Descriptive Data Analysis (RQ1)

4.4.1 Food Types and Food Consumption Level

Figure 4.2 shows how different types of foods (meals in the top row, and snacks in the bottom row) were reported. When the number of meals is 1 (top-left), the figure shows three clear peaks at around 10am, 2pm, and 8pm that correspond to breakfast, lunch, and dinner times, respectively. The peak for lunch is visible when the number of meal events is 2, the dinner time peak is visible in the top-right figure. Moreover, when considering plots for snack intakes (bottom row), the figures for "how_many" values of 2 and more than 2 are sparse, while in the plot for "how_many" having 1, there are two peaks, including one around noon, which is also a time period between breakfast meal at 10am and lunch meal peak at 2pm.

In the dataset, the total number of meal episodes are distributed as 746 (55%) as usual, 319 (24%) undereating, and 285 (21%) overeating. Hence, the number of undereating and overeating episodes are comparable. For snacks, the numbers are distributed as 247 (49%) as usual, 190 (38%) undereating, and 63 (13%) overeating. Hence, the percentage of overeating episodes during snacks is relatively low compared to undereating, which is reasonable given that usual snacks are smaller in size. In addition, the percentage of undereating snack episodes is significantly higher (1.5 times the percentage) than for undereating meal episodes. Moreover, Figure 4.5 shows that a higher number of overeating events are

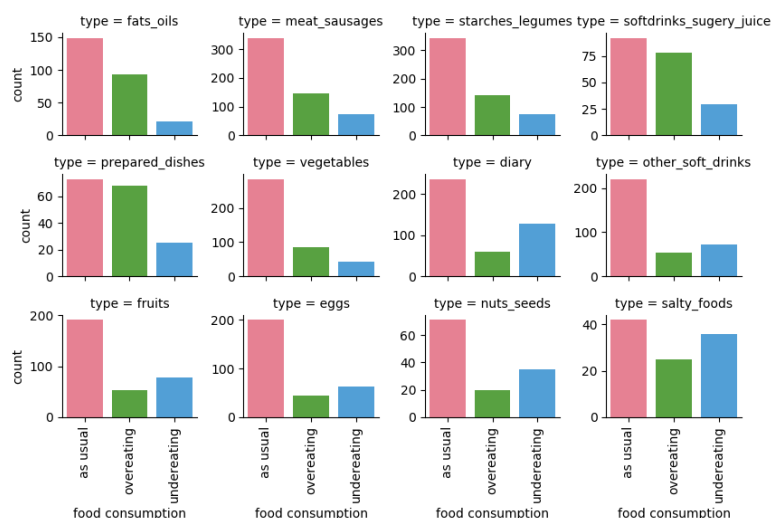


Figure 4.3: Bar charts for ten different food categories. In each subplot, the x-axis indicates the three classes regarding food consumption level, and the y-axis indicates the total number of reported cases.

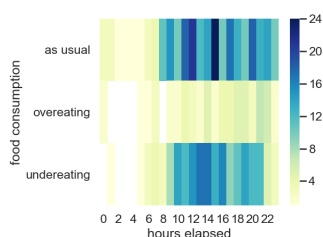


Figure 4.4: Heatmap for hourly food consumption episodes considering **snack** intake. "hours elapsed" denotes the # of hours from the start of the day.

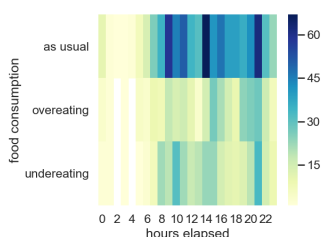


Figure 4.5: Heatmap for hourly food consumption episodes considering **meal** intake. "hours elapsed" denotes the # of hours from the start of the day.

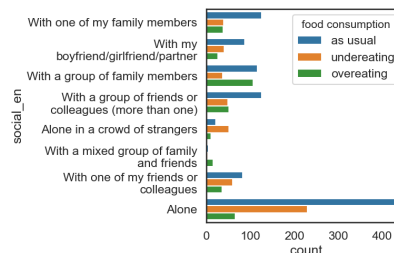


Figure 4.6: Bar chart representing the number of eating events related to different social contexts and associated food consumption levels.

found closer to lunch and dinner meal peaks.

As shown in Figure 4.3, when considering food categories and consumption levels, overeating was associated 93 times (35.4%) when "fats and oils" were present in the food as compared to undereating, which was reported only 21 times (7%). Food type "meat and sausages" was also associated to overeating almost twice the number of times as compared to undereating. For the food type "soft drinks or sugary juices", overeating episode reports (46%) were almost equal to the number of "as usual" episodes (39%). Moreover, for prepared dishes, the relation among "as usual", "overeating", and "undereating" was 3:3:1, which suggests that participants were highly likely to have reported "overeating" or "as usual" when eating prepared dishes, that are typically found in restaurants, bars, or cafes. These results match prior literature that links overeating to eating out in restaurants where food is often more palatable [322]. Moreover, our results are consistent with prior research that also suggests people reported overeating in the presence of fats and oils, sugary drinks, and meat [226, 538].

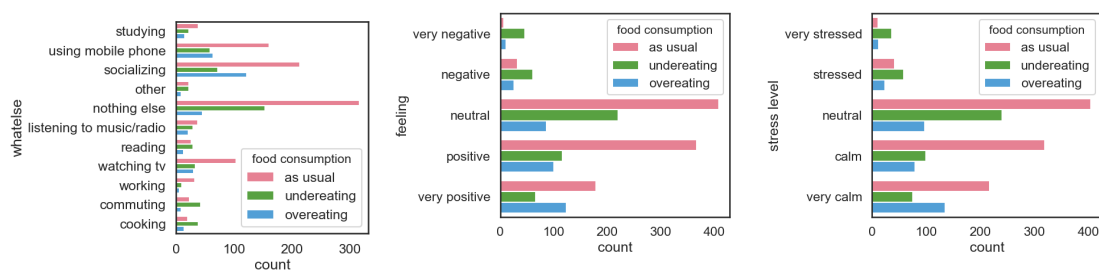


Figure 4.7: Bar chart represent- Figure 4.8: Mood while eating and Figure 4.9: Stress while eating and ing reported concurrent activities count of reported instances. done while eating. count of reported instances.

4.4.2 Sociability and Food Consumption Level

We considered 8 contexts of sociability in our analysis. 3 social contexts correspond to groups (with family and friends, family members, friends or colleagues), 3 contexts correspond to eating with another person (with one family member, boyfriend/girlfriend/partner, one friend/colleague), and the remaining 2 contexts (alone, alone in a crowd) correspond to eating alone. Figure 4.6 shows the distribution of food consumption episodes across different social contexts, and it indicates that all group eating episodes had a higher number of overeating cases (substantially for family; slightly for the other two) compared to undereating episodes. On the other hand, when eating alone, 230 reports indicated undereating, and in the social context "alone in a crowd of strangers", the percentage of undereating episodes (61.4%) is higher than the "as usual" percentage (25.3%). In summary, these results suggest that when participants were in groups, they had more overeating episodes as compared to undereating episodes. In eating behavior research, this phenomenon has been called *social facilitation* [209, 210]. These results match prior research that discusses eating as a holistic social event and how people tend to eat more food than they usually eat when they are in more social contexts [556, 64, 209, 210]. On the other hand, when alone, the number of undereating episodes rose significantly. Further, when participants were with another person (partner, family member, friend), more undereating episodes were reported compared to overeating in all cases. This could be because of the effect called *impression management*, which deals with lowering food intake to please one's companions [209, 210]. Moreover, apart from face-to-face sociability, widespread mobile adoption has made people sociable in the online sphere as well. Some studies have examined the differences and similarities of online and real-world sociability of people in different countries [423]. In our dataset, apps such as facebook, whatsapp, and instagram showed different usage levels around the three food consumption levels (e.g., usage of social apps such as facebook and instagram have shown a higher frequency closer to overeating reports as opposed to undereating reports). These observations conform with prior literature that has indicated associations between social media use and food consumption [516, 537, 403].

4.4.3 Mood, Stress, and Food Consumption Level

According to Figure 4.8, when the mood was negative or very negative, people tended to report undereating. This result is puzzling, although some literature in nutrition and psychology has suggested that loss of appetite has been indeed reported when people go through short-term episodes of negative moods [559, 466]. The number of reported overeating episodes increases when going from a neutral to a positive mood. Our analysis shows that out of all the overeating reports with a very positive emotion (124 reports), the majority (47.58%) reported home as the eating location. Overall, the number of

reports with negative or slightly negative moods (183 reports) is low compared to positive and very positive moods (952 reports). As shown in Figure 4.9, a similar overall trend can also be seen for the reported stress level. This similarity is reflected in the high and positive Pearson correlation coefficient (0.77, $p < 10^{-5}$) between mood and stress [48].

4.4.4 Concurrent Activities and Food Consumption Level

Figure 4.7 shows the types of concurrent activities done while eating as reported by the participants. The "Nothing Else" category was reported 514 times, and it had the lowest percentage of overeating reports (8.2%) out of all the different concurrent activities. As the literature suggests, this might be because when participants are fully focused on the eating event, they are mostly aware of their food consumption level [435]. Moreover, results suggest that participants reported more overeating episodes as compared to undereating when they socialized while eating, which is consistent with prior research that suggested social eating as an instance when people overeat [389, 64]. Furthermore, using mobile phones while eating also corresponds to more overeating episodes than undereating. Watching TV had about the same numbers for overeating and undereating. This result is consistent with a recent study that explored the relationship between screen time (mobile or TV) and unhealthy diets [489]. Moreover, another recent study suggested that distracted eating can lead to overeating [435], which indeed happens when eating while using the phone or watching TV. This is also why studying screen time and app usage are important when analyzing overeating episodes with a holistic approach. Moreover, activities such as commuting, cooking, and reading had higher percentages of undereating as compared to both overeating and as usual. Recent reports regarding commuting and overeating suggest that long commutes could significantly influence overeating behavior [104, 225].

4.5 Statistical Analysis (RQ1)

4.5.1 Pearson and Point-Biserial Correlation for Self-Report Features

We conducted Pearson [501] (for mood, stress) and Point-biserial [45] correlation analysis (for social context, food types and food categories because they are binary values) to understand feature relationships between self-report features and food consumption level, and some results are summarized in Table 4.1. Mood, Stress, and Sociability showed correlation values above +0.29, indicating close to moderate relationships. This is once again conforming to prior nutrition studies such as [556, 210, 209, 528], that linked psychological aspects and sociability to food consumption level. On the other hand, we observed a correlation value of +0.22, which is a weak positive linear relationship between food type and food consumption level. Food categories such as meat and sausages, food starches and legumes, fats and oils have values above +0.3, indicating weak positive linear relationships with food consumption levels. Similar findings were previously reported in some nutrition studies [226, 538] where they reported overeating in the presence of categories of foods such as fats and oils and meat and sausages.

4.5.2 Statistical Analysis of Dataset Features

Table 4.2 (left) shows statistics such as t-statistic [249], p-value [191], and Cohen's-d (effect size) with 95% confidence interval (CI) [266] for all the features in the dataset for two groups: **Overeating** and **Undereating (OverUnder)**. Hence, the objective is to identify features that would allow to discriminate

Table 4.1: Pearson and Point-Biserial correlation analysis for some self-report features and food consumption level.

Features	Value	Features	Value
food consumption level	1 (+)	food meat sausages	.31827 (+)
mood	.35927 (+)	food fats oils	.32637 (+)
stress	.29719 (+)	food starches legumes	.30412 (+)
social context	.29165 (+)	food prepared dishes	.24144 (+)
		food softdrinks sugery juice	.25710 (+)

Table 4.2: Comparative statistics of 20 features across classes "overeating" and "undereating" (OverUnder) and "overeating" and "as usual" (OverUsual): t-statistic, p-value (* if $p > 0.01$, ** if $p > 0.1$, *** if $p > 0.5$), and Cohen's-d with 95% confidence intervals. Features are sorted based on the decreasing order of t-statistics.

OverUnder				OverUsual			
Feature	Group	t-statistic	Cohen's-d [95% CI]	Feature	Group	t-statistic	Cohen's-d [95% CI]
social context	CON	12.35	0.85, [0.71, 1.00]	social context	CON	5.77	0.53, [0.35, 0.72]
fats or oils	FOOD	10.03	0.66, [0.52, 0.80]	microsoft launcher	APP	5.01	0.40, [0.21, 0.58]
meat or sausages	FOOD	9.46	0.64, [0.50, 0.78]	mood	PSY	4.64	0.41, [0.23, 0.60]
mood	PSY	8.76	0.61, [0.47, 0.75]	prepared dishes	FOOD	4.37	0.37, [0.19, 0.55]
starches or legumes	FOOD	8.72	0.59, [0.45, 0.73]	fats or oils	FOOD	4.35	0.38, [0.20, 0.57]
stress	PSY	8.31	0.58, [0.43, 0.72]	softdrinks or sugery juices	FOOD	3.97	0.34, [0.16, 0.52]
softdrinks or sugery juices	FOOD	7.42	0.49, [0.35, 0.63]	stress	PSY	3.94	0.35, [0.17, 0.54]
prepared dishes	FOOD	6.87	0.46, [0.32, 0.59]	minutes elapsed	TIME	3.57	0.34, [0.16, 0.52]
vegetables	FOOD	6.63	0.44, [0.31, 0.58]	hours elapsed	TIME	3.52	0.33, [0.15, 0.52]
microsoft launcher	APP	5.58	0.34, [0.20, 0.48]	screen on count	SCR	2.52*	0.22, [0.04, 0.40]
acc z abs bef	ACC	4.41	0.31, [0.17, 0.45]	meat or sausages	FOOD	2.09*	0.19, [0.01, 0.37]
acc z abs	ACC	4.38	0.30, [0.17, 0.44]	facebook	APP	2.00*	0.18, [-0.00, 0.36]
facebook	APP	3.21	0.22, [0.08, 0.35]	acc z bef	ACC	1.85*	0.17, [-0.01, 0.36]
acc z abs aft	ACC	3.14	0.22, [0.08, 0.36]	salty foods	FOOD	1.68*	0.15, [-0.04, 0.33]
minutes elapsed	TIME	1.95*	0.13, [-0.00, 0.27]	starches or legumes	FOOD	1.64**	0.15, [-0.03, 0.33]
whatsapp	APP	1.88*	0.13, [-0.01, 0.27]	acc z	ACC	1.61**	0.15, [-0.03, 0.33]
hours elapsed	TIME	1.82*	0.26, [0.12, 0.39]	acc z abs bef	ACC	1.52**	0.15, [-0.03, 0.33]
instagram	APP	1.32**	0.09, [-0.05, 0.23]	acc y	ACC	1.41**	0.13, [-0.05, 0.31]

between "overeating" (n = 348) and "undereating" (n = 509) episodes. Table 4.2 (right) shows the same set of statistics considering the two groups **Overeating** and **As Usual (OverUsual)**. The objective here is to identify discriminating features for two groups "overeating" (n = 348) and "as usual" (n = 993) episodes, which are closer in range to each other. In both tables, the features are ordered by the descending order of t-statistics. Moreover, since p-values are not sufficiently informative [577, 276], we additionally calculated Cohen's-d [429] to help understand the statistical significance of the features. To interpret Cohen's-d, we used a commonly used rule of thumb: small effect size = 0.2, medium effect size = 0.5, and large effect size = 0.8. Moreover, we calculated 95% confidence interval for Cohen's-d. The higher the Cohen's-d value, the stronger the possibility of discriminating the two groups using the considered feature. If the confidence interval does not include zero, the confidence interval of the effect size depicts the reliability of the effect size. If the confidence interval is on the positive side, the feature could be promising for discrimination among the groups. Moreover, the narrower the confidence interval, the higher the reliability of the calculated effect size.

In Table 4.2 (left), social context is the feature with the highest t-statistic, and it has a large effect size too. Hence, it shows that social context during an eating episode is the most discriminating feature with regard to OverUnder. In Table 4.2 (right), social is also the feature with the highest t-statistic, which suggests that it is the most discriminating feature for OverUsual. Moreover, this feature's t-statistic and effect size has dropped from 12.35 and 0.85 to 5.77 and 0.53 from OverUnder to OverUsual. This suggests that the social feature has a higher discriminating capability for OverUnder as opposed to OverUsual.

When considering food types, "fats or oils" was the feature from the FOOD feature group with the highest t-statistic and effect size for OverUnder. However, a place down in OverUsual, whereas prepared dishes had a t-statistic of 4.36 and less than medium affect size. Even features such as "meat or sausages", "softdrinks or sugary juices", and "starches and legumes" switched positions in the two tables showing that features that affect the discrimination of the three food consumption level classes can differ in relevance. However, "starches or legumes" and "salty foods" contain zero in the corresponding confidence interval for OverUsual, suggesting that they might not be reliable features to distinguish between these two classes. Another feature, "vegetables" appeared in OverUnder with medium effect size, but disappeared from the Table for OverUsual. In both tables, mood, and stress showed differences across groups. The effect sizes were 0.61 (resp. 0.57) in OverUnder, and 0.41 (resp. 0.35) in OverUsual. This result again tallies with previous findings in nutrition and behavioral psychology research as discussed in Section 4.3.

If we consider app usage, "facebook" and "microsoft launcher" are discriminant of the two classes in OverUnder with high t-statistics and small, medium effect sizes, respectively. Even though "whatsapp" and "instagram" appear in OverUnder, negative values in the confidence interval suggest that those are not reliable features for discrimination. When considering activity level derived from accelerometer, OverUnder contains three important features from z-axis, that have slightly high t-statistics and small-to-medium effect sizes. However, in OverUsual, these features have gotten smaller effect sizes with negative values in confidence intervals. Interestingly, minutes elapsed and hours elapsed features from the group TIME have moved from low t-statistics and unreliably small effect sizes in OverUnder to higher t-statistics and at least small effect sizes in OverUsual. This suggests that the time of the day at which food intake has been reported would be important in deciding between "overeating" and "as usual" episodes, but not between "overeating" and "undereating" episodes. Moreover, screen events feature showed medium effect size and high t-statistics in OverUsual. It also supports findings from prior literature that screen time is an important variable in determining "overeating" and unhealthy dieting habits [489]. Further, features related to battery events (BAT) and radius of gyration (derived using location data) did not seem to be significant enough features to discriminate food consumption levels of consideration in either table.

4.6 Food Consumption Level Inference (RQ2)

4.6.1 Three-class Food Consumption Level Inference

The three-class inference task uses different subsets of features in the training set, and calculates classification accuracy, precision, and recall. The target classes were *overeating*, *undereating*, and *as usual*. We used python with scikitlearn and keras in this phase, and we conducted experiments using several model types (in the decreasing order of accuracies for inference task G5): random forest, naive bayes, gradient boosting, neural networks, XGboost, AdaBoost, and support vector classifiers. However, considering space limitations and aspects such as interpretability and model personalization, we present inference results for two models as follows:

(a) Random forest classifier (RF) with ntree values between 50 - 500: we got the highest accuracy values for inference tasks using RFs. More importantly, RFs models output the feature importance values used in inference, hence enabling us to understand and interpret the results.

(b) Multi-layer Perceptron Neural network (NN) with 3-4 layers with relu activations, dropout for regularization, and binary cross entropy loss function (number of layers, number of nodes in intermediate

layer/s changed depending on the feature group and the input dimensions; hyper-parameter tuning was done for each feature group with the goal of obtaining the best model for the task): NNs provided with reasonable accuracies, and allow transfer learning without much sophistication [384]. Hence, we present results for NNs here because we use them for transfer learning to personalize models in Section 4.6.2.

We followed the leave k -participants out strategy ($k = 15$) for all the experiments when preparing the dataset, where training and testing splits did not have data from the same user, hence avoiding this possible source of bias in the evaluation procedure. Moreover, when preparing the dataset, we made sure that the classes are balanced by up-sampling the minority classes and down-sampling the majority class to get a balanced dataset of 2400 records. The baseline for experiments is 33.3% since the classes were balanced in all inference tasks. We conducted experiments for individual feature groups and meaningful feature group combinations:

Self Reports without FOOD (G1): This corresponds to self-reports that would not be available in a traditional mobile food diary, such as reports regarding eating context (sociability, concurrent activities), psychological state (mood, stress) together with the time of eating. The objective is to show that even without capturing the types and amounts of food, it is still possible to infer food consumption levels. An envisaged application scenario of this inference is where the mobile health app simply captures these few self-reports instead of all the food consumption details, hence making the user experience better in terms of the lower burden of manual data input.

Self Reports (G2): This corresponds to all the self-reported features including food types and categories. In addition to features in G1, this also captures the types of food consumed by participants. This inference would reaffirm the relationship shown in Figure 4.1 with regard to the association between food consumption level and aspects such as mood, stress, sociability, activities, and food.

Passive Smartphone Sensing with TIME (G3): This feature group combination contained a single self-reported variable (time of eating), and a set of passively sensed features without any user input, such as accelerometer, app usage, location, screen usage, and battery events. Importantly, this group reflects an envisaged mobile health application usage scenario where participants only report that they ate, hence capturing the time of eating, without typing all the details about the food types and amounts, sociability, and concurrent activities, hence making it less tedious. In addition, prior work has examined the use of passive mobile sensing features to determine the time of eating [420, 44, 506], which is a separate open research question. Hence, this feature group combination denotes an envisaged use case that depends on *near-passive* sensing.

All Feature Groups without FOOD (G4): This contained all feature groups except for FOOD. Hence, this set of features would require the same set of user involvement/effort as in G1. As we are following a holistic approach regarding food consumption, the goal here is to evaluate whether only knowing about the contextual factors and sensing data without knowing the food types and amounts could characterize the food consumption levels.

All Features (G5): This used all the available features to demonstrate the potential of a future mobile food diary that is driven by passive smartphone sensing in addition to traditional self-reports. This feature group captures food-related details, contextual and socio-psychological attributes, and passive sensing data.

All Features w/ PCA (G6): Out of the 3 commonly used multi-modal fusion techniques: (a) data, (b) feature, and (c) decision) [471, 122], all inference tasks in this chapter except for G6 used feature-level

Table 4.3: Three-class food consumption inference (overeating, undereating, as usual) accuracy, precision, and recall obtained with a random forest classifier (RF) and a neural network (NN) using different feature group combinations.

Feature Group Name (# of Features)	RF			NN		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Baseline	33.33%	-	-	33.33%	-	-
SCR (1)	40.26%	42.65%	40.97%	31.35%	33.46%	15.91%
APP (10)	47.52%	51.29%	46.74%	45.21%	48.68%	44.05%
PSY (2)	50.82%	52.36%	50.59%	44.88%	51.81%	44.54%
LOC (2)	56.43%	56.31%	56.68%	32.34%	29.63%	33.94%
CON (3)	62.37%	62.27%	62.37%	45.87%	55.08%	45.70%
BAT (2)	63.03%	61.58%	62.69%	50.03%	40.36%	51.64%
FOOD (15)	65.34%	66.11%	65.38%	60.72%	60.83%	60.78%
TIME (2)	67.65%	67.14%	67.01%	56.30%	56.04%	56.14%
ACC (18)	76.89%	83.89%	69.76%	47.52%	47.01%	57.89%
G1: CON + PSY + TIME (7)	81.19%	81.45%	80.91%	62.19%	62.39%	62.11%
G2: CON + PSY + TIME + FOOD (22)	82.50%	82.61%	83.56%	66.68%	67.29%	67.25%
G3: ACC + APP + LOC + SCR + BAT + TIME (35)	83.49%	83.19%	82.84%	73.26%	73.33%	72.20%
G4: ACC + APP + LOC + SCR + BAT + CON + PSY + TIME (40)	83.61%	83.99%	83.57%	79.20%	79.26%	79.17%
G5: All Features (55)	87.81%	87.97%	88.37%	82.17%	82.19%	82.95%
G6: All Features w/ PCA (principle components=4) (4)	83.53%	83.46%	83.54%	76.67%	76.94%	76.53%

fusion, that feeds a processed feature map into a classifier. In G6, we examined feature extraction and dimensionality reduction with principal component analysis (PCA), which fuses the features before feeding them into the classifier.

Results of the experiments (Table 4.3) show that RFs perform better than NNs across all feature groups and evaluation measures. Hence, in this section, only the results from RFs are discussed. Table 4.3 shows that individual feature groups such as ACC (accelerometer data), TIME (time of the day), FOOD (types of food consumed – similar to a traditional food diary), BAT (battery events), and CON (context when consuming food) have accuracies above 60%, while the highest accuracy of 76.89% corresponded to ACC feature group. This suggests that activity levels derived from the smartphone can be used to distinguish food consumption levels, to some degree. This is justifiable because prior work in nutrition and behavioral sciences has discussed the relation between food consumption levels and activity levels [210, 556]. G2 provides an idea regarding accuracies that can be obtained with currently available mobile food diaries that fully rely on participant self-reports. Accuracy, precision, and recall had values in the range 81%-83%, suggesting that self-reports are performing well.

G3 shows that it has an even higher accuracy when compared to G1 or G2. Given that most prior research in nutrition has relied on self-reports regarding food categories and volumes to analyze food consumption behavior, this result shows that it is worth looking into passive smartphone sensing for cues regarding food consumption levels, given that there are many aspects such as app use, screen time, sociability that relate to the way people consume food in modern societies. Moreover, this result aligns with the hypothesis we presented in Section 4.3.4 and Figure 4.1 with regard to the possibility of inferring food consumption levels primarily using mobile sensing features.

G4, with an accuracy of 83.61%, shows the benefit of considering eating as a holistic event as compared to just food categories and volumes. This accuracy, which is above the accuracy obtained with G2, again shows the importance of the holistic view of eating. Finally, by combining all the feature groups in G5, the model achieved an accuracy of 87.81%, a precision of 87.97%, with a recall of 88.37%, all of which are encouraging. In addition, for G6, we got the best results with 4 principle components, an accuracy of 83.53%, which is higher than G1 and G2, although it is still lower than the corresponding feature level fusion (G5) that used the same set of features. This result suggests that passive smartphone

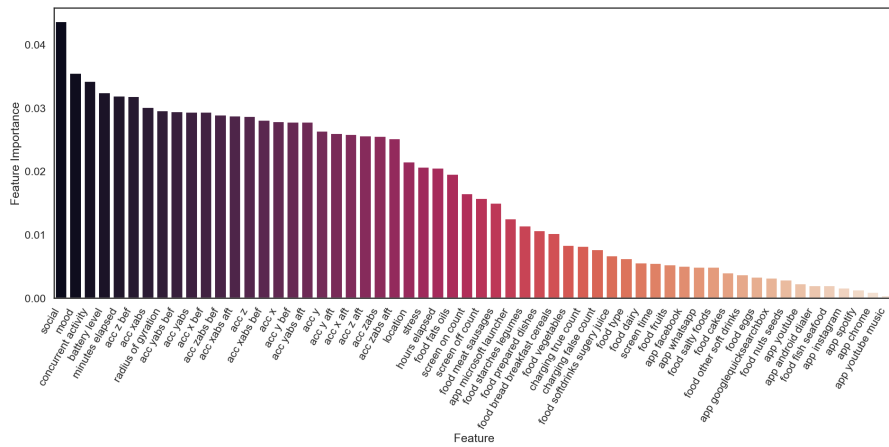


Figure 4.10: Feature Importance generated using the RF for G5 Inference

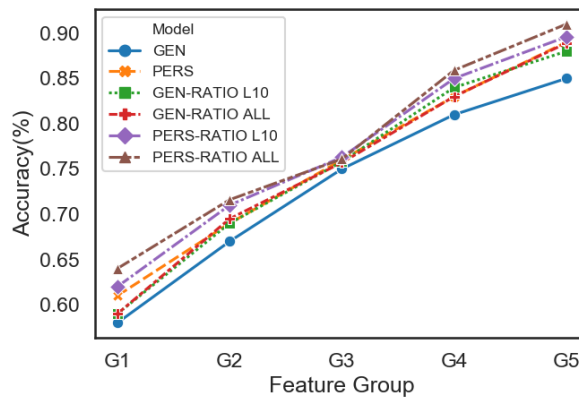


Figure 4.11: Accuracies for different feature group combinations among GEN and PERS.

sensing can be of great value when incorporated into mobile food diaries that are currently based only on self-reports. Further, these results also highlight the potential that passive smartphone sensing has as part of mobile applications for food monitoring with less intrusive usage scenarios.

In Figure 4.10, we plot the importance of features in classifying food consumption levels using RFs for G5 inference, hence containing all the features. Results indicate that social, mood, and concurrent activity are the self-reported features that were among the top 10 features. In addition, passive smartphone sensing features such as battery level, acc z bef, acc x abs, radius of gyration, acc y abs bef, and acc y abs were among the top 10 features in terms of feature importance. This indicates that activity levels and phone battery levels around food intake events are useful as part of a rich model. It is also important to note that even though food category-related features had high t-statistics and effect sizes in Table 4.2, Figure 4.10 shows that those features did not have comparatively high feature importance values in the RF.

4.6.2 Model Personalization

In Table 4.3, we presented inference results without any personalization by only considering eating event-level features, and by not considering any temporal features. In this section, we attempt to personalize inferences by considering aspects regarding inference models (model-wise) and by including additional temporal features to the dataset (data-wise), with the expectation of increasing the accuracy of food consumption level inference. Finally, we show results for NN models because they allow transfer learning in a relatively straightforward manner.

Transfer Learning for Model Personalization

In a prior study, LiKamwa et al. [285] have shown that general models are sufficient when users do not provide enough data for personalized model training. However, when the training dataset size increases, e.g., when collecting data for more days, the personalized models outperformed general models for inferences. In our dataset, we have limited data from individuals, hence restricting the possibility of training separate personal models for all users. Hence, we trained models on general data, and then follow a transfer learning approach to retrain the model with individual data. In our dataset, 11 users reported more than 50 food intake reports. For this task, we used the training procedures and feature group combinations presented in Section 4.6.1. The two model types we used were:

(1) GEN: the model is trained on data from all users. However, as we described in the training procedure in Section 4.6.1, we do not use data from the same individual in training and testing splits. Hence, this model has been trained on a dataset in which the specific user's data is not included. When testing, we used 30% split of user's data.

(2) PERS: We retrained the already trained model using a split (70%) from user's personal data, and test it on the unused 30% split of personal data.

We carried out experiments for the 11 users having more than 50 food reports, and included the average results in Figure 4.11 by repeating experiments for five iterations. Results show that for all feature group combinations, PERS outperformed the GEN with accuracies of the range 0.95% (G3) – 4.08% (G5).

Using Temporal Features for Model Personalization

The inferences done in previous sections did not consider food intake events to be linked to each other, as we considered all the food intake events to be mutually exclusive from even past food intake events of the same person. Adding temporal features could possibly increase accuracy. Hence, we introduce *four* new features: (1) previous food consumption level (*prev_level*) - capturing the previous food consumption level reported by the individual; and (2) three ratios to the dataset that capture the temporal evolution of reported food consumption levels of an individual participant. The ratios are: (2a) overeating ratio (*oer*): the total number of overeating episodes reported in the past, divided by the total number of episodes up to that time; (2b) undereating ratio (*uer*): the total number of undereating episodes reported in the past, divided by the total number of episodes by that time; (2c) as-usual ratio (*aur*): the total number of as-usual episodes reported in the past, divided by the total number of as-usual episodes by that time. Hence, these three ratios capture the participants' reporting history.

Moreover, when calculating the ratios, we examined two temporal windows: (1) Last ten food intakes (L10), which captures the recent food consumption levels; and (2) All previous food intakes (ALL), which captures the overall history of an individual. We assessed the efficacy of these features using the

Table 4.4: Summary of the Qualitative Analysis.

Topic	Insight & Quotes
Stress	Stress can play a key role with regard to eating behaviors. Stress has caused undereating or overeating in participants. “[...] we really eat according to how we feel at that moment.” (Female 30 yo) “[...] when I’m stressed, I tend to skip meals and end up eating later more food than I should, including grease food.” (Female 24 yo)
Food types	Food types and portions eaten by volunteers varied significantly. Variability depends on the place, social context, stress level, and activity load. Food eaten at the school cafeteria and places around campus often have lower quality and variety than food prepared at home. Junk and fatty food are widely available to students, which sometimes fosters unhealthy eating behaviors. “Nutritious things are generally more expensive and take longer to cook. Therefore, we choose to eat fast food.” (Female 22 yo) “The time we dedicate to preparing our food, we want to live at a very fast pace, and we prioritize spending time on social networks. If we spent some of that time preparing our food, we would consume a much smaller amount of fast food.” (Male 24 yo)
Places and Times	Participants mainly eat at home and the school cafeteria. Some prepare their food at home and take it to school, and others go to restaurants or cheap kitchens near school. Eating schedules are mainly determined by study/work load. However, some students only eat when they find the opportunity to do so, or have free time (they have no defined schedules). In general, reported times for breakfast, lunch and dinner are between 6:00 a.m. to 12:00 p.m., 1:00 p.m. to 6:00 p.m. and 9:00 p.m. 00 to 00:00, resp. “I eat when I have free time” (Female, 21 yo) “I mainly eat at the university cafeteria or places around. I sometimes eat at home” (Male 22 yo) “I prepare breakfast and dinner, and my mother cooks lunch for me” (Female 20 yo)

following two tasks:

(1) GEN-RATIO: we re-trained all the models for feature groups G1 to G5 including the four features, for both temporal windows, using the approach presented in Section 4.6.1, and tested using the data from the aforementioned 11 participants.

(2) PERS-RATIO: we re-trained all the personal models for feature groups G1-G5 for 11 participants, for both temporal windows, using the approach presented in Section 4.6.2 by including the four additional features that capture their temporal evolution.

As shown in Figure 4.11, results suggest that using these four features increased the accuracies for all four model-feature combinations (GEN-RATIO L10, GEN-RATIO ALL, PERS-RATIO L10, PERS-RATIO ALL) with accuracies in the range 0.1% to 3.9%. In addition, for the feature `prev_level`, for groups overeating and undereating, we obtained a Cohen’s-d of 0.30, which suggests that this feature is informative of overeating vs. undereating episodes. Moreover, the three ratios had Cohen’s-d values between 0.12 and 0.16, which are less than the small effect size. In summary, the model performance can be increased by using features that consider the longitudinal aspects of data.

4.7 Discussion

4.7.1 Feedback from Participants

We start this section by summarizing some of the feedback obtained from the study participants (semi-structured interviews, focus group, and questionnaire) in Table 4.4.

4.7.2 Passive Smartphone Sensing for Characterizing Food Consumption Levels

Results presented in prior sections confirm that our hypothesis regarding food consumption levels and passive sensing features (Section 4.3.4) is valid. Most smartphones are capable of both continuous sensing (feature groups APP and SCR) and interaction sensing (feature groups such as ACC, BAT, TIME, LOC) for behavioral modeling. Obviously, these sensing modalities do not directly capture the food type or internal aspects that nutrition and behavioral science researchers have linked to food consumption levels in the past. However, in Section 4.6.1, we showed the potential of using passive smartphone sensing to infer food consumption level. What these modalities sense is not the food type or psychological aspects, but the physical activity levels and smartphone usage behavior. Given that physical activity levels have been linked to stress and mood in prior mobile sensing literature [418, 468], we believe these passive sensing modalities contain contextual information that could directly relate to food consumption behavior, and that is the reason why inferring food consumption levels with an accuracy of 83.49% using passive smartphone sensing and time-related features was feasible. This is one of the first studies in this direction, and there are plenty of opportunities to explore *eating as a holistic event* as we did here. Given that computer vision researchers are focused on identifying food intake types and levels using images of the food portion (Section 4.3), we could expect food consumption self-reports to get automated in future mobile food diaries. However, considering that eating is a holistic event driven by many factors, further research to determine food consumption behavior could enable advanced mobile food diaries that do not solely depend on user input to generate recommendations and interventions.

4.7.3 Further Informative Features Regarding Food Consumption Levels.

We acknowledge that the features we generated from passive modalities are simple and easier to interpret when associated with eating episodes. However, there is ample opportunity to build upon these findings, and develop novel features that could discriminate food consumption levels with higher accuracies. For example, in this chapter, all the accelerometer features are single-dimensional and we did not use linear acceleration or 2D resultant acceleration features due to some limitations in the data collection process (not having data from the gyroscope to match accelerometer traces so that gravity biases could be removed [41, 207]). Moreover, when considering app usage behavior, the features we used only determined whether a particular app was used or not during the time window of the eating episode. However, advanced research could be done to determine the usage times of each app during eating episodes, hence obtaining a comprehensive understanding of app usage behavior related to eating. Moreover, using a low-power API such as Google Activity Recognition API to detect activity types could generate new features that might be beneficial in characterizing eating events.

4.7.4 Accounting for Diversity

The eating behavior of people in different countries vary depending on a plethora of factors such the culture, type of food they consume, concurrent activities while eating, and how they perceive events such as eating [529, 563, 254]. Hence, it is important to clarify that the results from the deployment of our application in Mexico are exploratory and not representative of the food consumption behavior of people from other regions. Moreover, if other aspects apart from food are considered, there are already known differences with regard to factors such as sociability [394, 444], activity levels [195, 39], and phone usage [286, 397] in different countries, and these aspects could get reflected in smartphone sensing datasets. For example, a study regarding the sociability of university students in Mexico and

the USA showed that Mexican students perceived themselves to be less sociable compared to how Americans perceived themselves, although in reality, Mexicans were more sociable [423]. Moreover, results show that Americans socialize more in private environments or by interacting through social media. On the other hand, Mexican students preferred to be more social in person with people who are around them [423]. A similar trend was visible in our results as depicted in Section 4.5.2, where social was the feature with the highest t-statistic and effect size among all features used in the analysis. Hence, we could expect differences in passive sensing data obtained from students in these two countries. It is fundamental to consider human diversity in smartphone sensing studies, and we believe more studies should integrate these diversity aspects in the future. Hence, future research could look into deploying mobile food diaries with sensing capabilities in diverse user groups based on ethnicity, culture, and geographic regions. In our opinion, the goal of such studies is to build models for mobile food diaries that generalize well enough to cater to and adapt to diverse user populations. Even though our study is focused on college students of a Latin American country, we believe that this is a first step in this direction.

4.7.5 Limitations and Future Directions

Similar to other prior mobile sensing in-the-wild deployments [55, 454, 468, 30], we used feature-level fusion techniques in the inference task except for G6. However, other decision-level fusion techniques could be explored in future work for similar mobile sensing datasets. Further, this chapter used the MEX dataset from college students in Mexico. Even though the features we used did not contain country-specific details, studies similar to this could be carried out among diverse user groups within Mexico and also in other countries to determine the validity and applicability of our approach to user groups of diverse ages and occupations. Moreover, we prepared the base dataset (in Section 2.1.2) assuming that individual eating episodes are mutually exclusive (similar to prior studies regarding mobile app deployments in-the-wild [418, 468, 454, 55, 285]). By making this assumption, we created a general dataset for the worst-case scenario where no temporal relationship between eating events for a given individual is assumed, hence avoiding any personalization effect due to considering longitudinal features from the same individual. However, in Section 4.6.2, we show how some basic longitudinal features can be used to personalize inferences, hence leading to higher accuracies for food consumption level inferences. We believe it is best to examine these longitudinal aspects in depth in separate studies that explicitly focus on these aspects.

Similar to other studies, in the real-life app deployment, it becomes challenging to verify all the self-reports given by participants [468, 55, 454, 285]. For example, in our study, even if photos of food types were captured, it is difficult to verify other self-reports such as sociability, concurrent activities, mood, and stress. While we used different techniques to monitor the completion of food intake reports in line with prior work [55, 454], we believe that these are challenges that are common, especially during real-life field experiments where subjective reporting is used to collect ground truth labels. In future work, an interesting question relates to possible gender differences in mobile food diaries. Prior literature has suggested gender differences in food choice [552], activity levels [27], app usage [251], and different factors that we discovered in this chapter to be affecting food consumption level. Moreover, the number of participants in our study was 84, which is reasonable given the recent food studies in smartphone sensing by Seto et al. [469] (12) and Biel et al. [55] (122). We collected data from participants for a longer time span compared to Biel et al. (10 days) and Seto et al. (6 days). Further, many other recent smartphone sensing studies with app deployments had participants in the range 50 - 100 [67, 30, 179, 545]. Based on this, we believe that the total number of events we captured

through the study could be enough to answer the research questions we addressed here reasonably. Future work could investigate larger sample sizes, which could allow us to study issues of generality and personalization in more depth.

4.8 Conclusion

In this chapter, we examined the eating behavior of 84 college students using MEX, a dataset consisting of smartphone sensing data and self-reports. We demonstrated that behavioral mobile sensing and self-report features of college students around eating events can be used to infer self-reported food consumption levels. We achieved an accuracy of 87.81% for the 3-class food consumption level inference task, showing the potential of passive sensing together with self-reports in future mobile food diaries. Further, we show that an accuracy of 83.49% can be achieved for the same classification task by only using passive sensing data and time of eating events in an envisaged simplified mobile food diary usage scenario. This performance suggests that smartphone sensing opens the possibility of detecting self-perceived food consumption levels in future mobile health applications, which would need to be validated further with larger and more diverse user populations.

5 Sensing Eating Events in Context: A Smartphone-Only Approach

While the task of automatically detecting eating events has been examined in prior work using various wearable devices, the use of smartphones as standalone devices to infer eating events remains an open issue. This chapter proposes a framework that infers eating vs. non-eating events from passive smartphone sensing and evaluates it on a dataset of 58 college students. First, we show that the time of the day and features from modalities such as screen usage, accelerometer, app usage, and location are indicative of eating and non-eating events. Then, we show that eating events can be inferred with an AUC (area under the receiver operating characteristics curve) of 0.65 using population-level machine learning models, which can be further improved up to 0.81 for user-level and 0.81 for hybrid models using personalization techniques. Moreover, we show that users have different behavioral and contextual routines around eating episodes, requiring specific feature groups to train fully personalized models. These findings are of potential value for future mobile food diary apps that are context-aware by enabling scalable sensing-based eating studies using only smartphones; detecting under-reported eating events, thus increasing data quality in self-report-based studies; providing functionality to track food consumption and generate reminders for on-time collection of food diaries; and supporting mobile interventions towards healthy eating practices. The material of this chapter was originally published in [323].

5.1 Introduction

According to prior work in nutrition science and public health, unhealthy eating practices could lead to severe conditions such as heart disease, diabetes, high blood pressure, and high cholesterol [454, 420, 44]. Hence, understanding the etiology and managing eating behavior is crucial. Fueled by such motivations, researchers have come up with different techniques to detect and monitor food intake, among which keeping food diaries (also known as food journaling and food logging) is one of the most common ones [234]. On an individual level, food diaries help users with self-awareness, self-monitoring, and behavior change, and have also helped people with weight loss goals [578, 234]. At the population level, they help researchers conduct large-scale studies to understand population-level food consumption [234]. Food diaries originated as a pencil-and-paper based technique [13], but in recent times mobile food diaries have become popular, and widely adopted commercial mobile health (mHealth) apps such as Samsung Health [451] and MyFitnessPal [360] allow users to keep food diaries and facilitate mindful eating.

While keeping a food diary has many benefits, it is difficult to sustain the practice of reporting all

food intake over long periods due to a plethora of personal, societal, and technological factors such as forgetting to report food, losing motivation to report, and self and recall biases (e.g, not reporting all eating events intentionally) [330, 234, 578]. Such drawbacks have called for tools and techniques to automatically recognize eating events, as this would allow remind users to report food intake on time. Prior studies in mobile sensing have used wrist wearables [507, 578, 505], jaw-bone wearables [452], earables [53, 44], necklaces [98, 589], and other sensing modalities [420, 333] to detect eating by sensing wrist movements, bites, swallowing, and mastication among many other actions. While most of these techniques have shown promising performance in lab settings, some have also performed reasonably in everyday life conditions. However, these techniques require specific hardware configurations and wearables to be worn, which might be both a hassle for some users and unaffordable for others. Furthermore, wearable-based eating detection systems would have to maintain a connection with smartphones to automatically trigger actions on the phone, which requires bluetooth, wifi, or data connections to be kept turned on. As wearables are known for low battery life, the need to run continuously could drain the battery even faster. In contrast, recognizing eating events directly on the smartphone could address some of these usability and technical issues. Moreover, unlike wearables, smartphone coverage and mHealth app usage are already high in many countries [345]. For example, 96% of young adults aged 18-29 in the United States own a smartphone [343], and Nutrition and Diet apps have become the second most common app category among mHealth app users, just behind Fitness apps [345]. Prior studies in mobile sensing have looked into improving mobile food diaries with context-awareness [330, 55, 328, 329]. However, whether smartphones alone be used to recognize eating events remains an open question. Considering these aspects, sensing eating events on smartphones could provide the following benefits:

Automatic Food Intake Tracking and Reminders: Keeping mobile food diaries manually can be cumbersome as people tend to forget to report [47]. Automatic eating event recognition could, on the one hand, keep track of eating events to provide feedback and remind users to report forgotten eating events. If such inferences were combined with other inferences such as food type [55], social context [328], or food consumption level [330], a holistic food diary could be maintained with minimal user input. This vision has been discussed in prior dietary monitoring [47] and sensing [325] literature.

Self-Report Validation: A challenge in self-report-based questionnaires is under-reporting, i.e., participants failing to report eating events [47]. In the context of mobile food diaries, when an actual eating event is incorrectly considered as a non-eating event, it adds noise to the data, which can be detrimental when training machine learning models. A low-cost smartphone-based system that could estimate if a period contained an actual eating or non-eating moment could filter out highly confident self-reports from noisy reports. This would lead to higher quality labels for public health studies and for training machine learning models in research and commercially available food diaries.

Mobile Interventions: After determining whether someone is eating or not, many subsequent inferences to determine the behavior and context of eating could be made [55, 329, 330, 328]. These inferences could help provide context-aware interventions and feedback to app users [47]. More importantly, sensing eating events with the phone could be done without relying on external wearables, possibly reaching larger and more diverse populations.

Population-Level Studies: Smartphone-only inferences could make it easier for nutrition scientists and dietitians to conduct population-level eating studies among larger populations without the need for additional hardware such as wearable devices. Current population-level eating-related studies are predominantly done using self-reports. Some drawbacks of such studies, related to participant burden and attention limits of self-reports, could be addressed by automatically detecting eating events.

Prior work has discussed the use of wearables for automatic dietary monitoring in population-level studies [47]. Furthermore, a recent study discussed the detection of eating events using wearables to trigger further data capture on mobile food diaries [354]. However, conducting large-scale studies using additional wearable devices is expensive. Hence, the proposed method, based solely on smartphone sensing, could be helpful for nutrition researchers and dietitians to overcome issues in current methods and facilitate the implementation of population-level studies by automatically inferring eating events.

In summary, while the characterization of eating has been attempted in the smartphone sensing domain (social context [328], food consumption level [330], food category and type [55]), the use of smartphones as standalone devices to infer eating events remains as an open issue. In this chapter, we examine whether smartphone sensing features could be used to classify time windows as corresponding to eating vs. non-eating events using user-level, hybrid, and population-level machine learning models [152].

We pose two research questions:

RQ1: What situational contexts and behaviors around eating and non-eating events can be observed by analyzing the everyday eating events of a group of college students obtained via passive smartphone sensing?

RQ2: Can eating and non-eating events be inferred by only using passive smartphone sensing?

By addressing the above research questions, this chapter provides the following contributions:

Contribution 1: We analyzed how passive smartphone sensing features differ for eating and non-eating in everyday life situations. As a case study, a subset of the MEX dataset, which has over 12000 eating and non-eating events provided by 58 college students in Mexico, was used. We showed that features from modalities such as application usage, accelerometer, location, and time of the day are informative of eating and non-eating events.

Contribution 2: We defined and evaluated the task of inferring eating and non-eating events using passive smartphone sensing data, obtaining an AUC of 0.65 with population-level models, which can be increased to 0.81 with user-level models. Moreover, we showed that feature selection plays a key role when training user-level models. Each user might need models that use different features (compared to others) to achieve high performance. This shows the behavioral diversity of people around eating and the need to consider such diversity in building machine learning models that recognize eating events. We also found that hybrid models (partially personalized) perform reasonably well and on par with user-level models. The results illustrate the potential of using passive smartphone sensing for building context-aware and automated mobile food diaries.

This chapter is organized as follows. Section 5.2, describes the background and discusses related work. In Section 5.3, the study design, data collection procedure, and feature extraction techniques are provided. Section 5.4 presents a descriptive analysis and a statistical analysis of the dataset. The inference task is defined and evaluated in Section 5.5. A number of important issues are discussed in Section 5.6. Finally, the chapter is concluded in Section 5.7.

5.2 Background and Related Work

5.2.1 Nutrition Science Perspective

Eating as a Holistic Event. Guided by how and why people eat, Bisogni et al. [57] provided a contextual framework for eating and drinking events, describing them as holistic events with eight interconnected dimensions. The dimensions are food and drink (type, amount, source, how consumed), time (chronological, relative experienced), location (general/specific, food access, weather/temperature), activities (nature, salience, active or sedentary), social setting (people present, social processes), mental processes (goals, emotions), physical condition (nourishment, other status), and recurrence (commonness, frequency, what recurs). The primary idea behind this framework is that situational and behavioral factors guide eating. Further, Jastran et al. [229] showed that eating routines are embedded in the schedules around daily lives related to family, work, and recreation. They also said that repetitive patterns could be found in eating episodes among participants regarding the type of food and the situational context in which food was consumed. Similar ideas were proposed in other studies that also showed that social context, activity levels and types, psychological aspects, location, food availability, and several other situational and behavioral factors could affect eating behavior [477, 210]. Ma et al. [304] found that people in small towns and rural areas tend to travel sizable distances for lunch. In larger towns, people consume lunch at places closer to their workplace (e.g., canteen, close by restaurants, fast-food outlets) because of the limited time available for eating. This shows how eating behavior is related to dimensions such as recurrence, location, and time. Further, some studies examined links between app usage and eating behavior. For example, Turner et al. [516] showed that excessive Instagram usage could be indicative of Orthorexia Nervosa, a condition of having an obsession with maintaining a healthy eating behavior, including a focus on healthy eating, food anxiety, and dietary restrictions.

Even though not conclusive, such studies point towards modern mobile social media playing a role in shaping eating behavior. Hence, all these nutrition and behavioral sciences studies demonstrate how different behaviors and situational factors could affect eating behavior. In addition, they also show the repetitive nature of different dimensions around eating events, which could be helpful in terms of modeling eating behavior using smartphone sensing and machine learning. Moreover, even though there is no concrete definition regarding eating events or episodes [47], going with the terminology regarding holistic eating behavior, throughout this chapter, the term *eating episode* is used as the actual time period of food intake. Moreover, the time period around the food intake also contains behaviors and contexts around the eating episode (e.g., going to the place of eating and coming back, using particular mobile apps before/after/while eating, etc.), that help us to consider eating as a holistic event is termed as *eating event*. Hence, an eating event is the eating episode, and behavior and context before and after the eating episode are captured with a time window. This definition is in line with prior work in mobile sensing that looked into characterizing eating and drinking events [55, 30, 330, 328]. More details regarding how these terms are operationalized can be found in Section 5.3.3.

Situational Context and Behavior as Proxies to Eating Events. Mobile sensing studies collect passive sensing and self-report data that can be broadly categorized into three pillars [325]: person, behavior, and context. What this means is that each sensing modality will be taken as a proxy to a trait that is related to a person (i.e., mood, stress, sociability, age, sex, etc.), behavior (i.e., activities, routines, etc.), or context (i.e., location, social context, environmental context, date and time, etc.). Prior work in smartphone sensing has shown that passive sensing features can be used to infer psychological aspects such as mood and stress (person) [468, 285], activity levels and types (behavior) [417, 451], sociability

and social context (person, context) [51]. In the context of eating behavior, these pillars of data can be mapped to the eight dimensions proposed by Bisogni et al. [57]. Gatica-Perez et al. [167] showed that mobile sensing features and self-reports could be represented using Bisogni's framework to understand eating routines meaningfully. In essence, this means that smartphone sensing features have shown to be promising in inferring attributes that have shown to be part of the eight dimensions related to eating events. Such relationships have been used in prior ubiquitous computing (ubicomp) studies regarding the eating and drinking behavior [30, 167, 55, 262, 454, 327, 330]. Leveraging these relationships, this chapter seeks to examine whether smartphone sensing could be used to directly infer eating events by taking situational context and behavior sensed via smartphones as proxies for eating events. This study objective is summarized in Figure 5.1. Hence, we hypothesize that date and time, application usage, screen usage, activity level, and movements are indicative of eating and non-eating.

5.2.2 Mobile Food Diaries

Food diaries are essential to understanding individual and population-level eating practices [234]. While manual food logging using pencil-and-paper-based techniques could be useful for self-reflection, mobile food diaries allow logging fine-grained details (weight and size of dishes, variety of dishes) about eating episodes more systematically by searching food types, varieties, and sizes in a database and logging them [234]. Some recent mobile food diaries have also looked into easing the process of logging using speech and photos [302]. Going a step further, some studies looked into estimating the calorie intake [377] and also calorie deficit by combining information from food diaries and passively detected physical activities [129]. Some popular mHealth apps that provide food journaling functionalities include Samsung Health, Lose It!, MyFitnessPal, EasyDietDiary, and SparkPeople [234, 330]. However, while food diaries provide many benefits, they also come with a plethora of drawbacks such as tediousness in using and finding the correct food type, difficulty in recording the correct dish size, and losing interest in logging over time, among which one of the most common drawbacks is users forgetting to report eating episodes [234, 55]. Hence, to overcome this barrier, researchers have come up with automated eating detection systems using different mobile sensing modalities [354, 507, 98, 589, 55]. In a nutshell, these systems would detect eating episodes and provide interventions, automatically keep track of events, or remind users to report details about the eating episode on a mobile food diary. The goal of this chapter, too, is similar. However, the main difference is that this study only used smartphones to make the inference, while previous studies used wearables.

5.2.3 Mobile Sensing for Eating Behavior Monitoring

Mobile sensing studies for eating behavior monitoring could be segregated into two main categories [325]: (1) detecting eating events (i.e., time of eating); and (2) characterizing eating events by identifying behavioral and situational context-related routines around eating episodes. Many prior studies using wearable sensing modalities to detect eating episodes fall under the first category. For example, Chun et al. [98] used a necklace-based wearable to detect jawbone and head movements in determining eating episodes. Their technique showed a precision of 95.2% and a recall of 81.9% in controlled studies, a precision of 78.2%, and a recall of 72.5% in a free-living study. Bedri et al. [44] studied an ear-worn wearable system called EarBit to detect chewing moments, with 90.1% and 93% accuracies in lab settings and outside-lab settings, respectively. Further, they showed that they could recognize eating episodes ranging from 2-minute snacks to 30-minute meals. Morshed et al. [354] used a wristwatch-based eating detection system, obtaining an accuracy of 96.5% in detecting eating episodes. They also showed how detecting eating episodes using a wearable could trigger momentary ecological

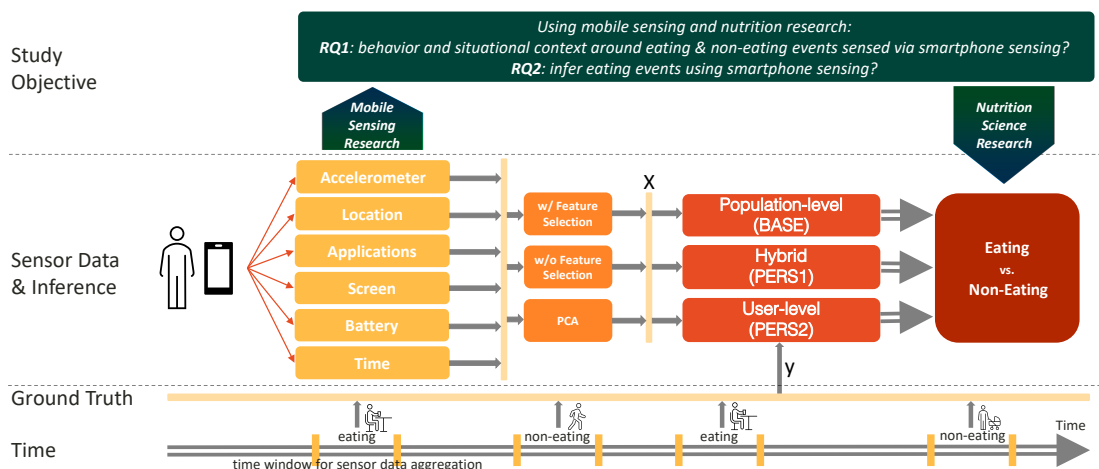


Figure 5.1: Objective of the Study

assessments (EMA) in the smartphone to capture additional contextual and behavioral information about the eating episode. Thomaz et al. [507] also showed that it is possible to detect eating episodes using smartwatch-based inertial sensors with F1 scores of 71.3% and 76.1% in two experiments done in free-living conditions. Moreover, Rahman et al. [420] used a combination of wrist-worn wearables and audio from a mic to predict about-to-eat moments with a recall of 77%. Further, they showed that personalization could increase the recall up to 81%.

Table 5.1 summarizes the differences between other mobile sensing studies and this study. On a fundamental level, while all other studies primarily used wearables, this study uses commodity smartphones. Further, while many prior studies focused on using sensor data streams from one or two wearable sensing modalities, rich and multimodal sensor data streams coming from smartphones are focused on here. In addition, while most studies attempt to detect eating episodes by sensing actions such as hand movement, chewing, bites, mastication, etc. (hence using them as proxies for eating), we seek to understand eating and non-eating events with behavioral and contextual features captured via smartphone sensors and leverage them to detect eating events.

5.2.4 Smartphone Sensing for Eating Behavior Monitoring

Smartphone sensing has not been often used to monitor eating behavior. Even the available studies focused on only characterizing eating events. Madan et al. [307] conducted a study to understand the eating behavior of university students in the United States. They concluded that healthy eating behaviors of individuals are related to the health and well-being of others with whom they associate. In another study, Seto et al. [469] concluded that behavior and context could affect eating patterns. However, these studies did not attempt an inference task using smartphone sensing data. Biel et al. [55] used smartphone-based contextual sensing to characterize eating events by detecting meal vs. snack events using a time window of two hours around eating episodes to aggregate sensor data. They deployed a mobile application among 122 Swiss university students, collecting over 4440 eating episodes. They performed an eating occasion type inference (meal vs. snack) with an accuracy of 84% with features such as time of the day, time since the last food intake, location, and other sensor data. Meegahapola et al. [330] showed that self-reported food consumption levels (eating as usual,

5.3 Data, Smartphone Features, and Definition of Eating/non-eating Episodes

Table 5.1: Summary of Eating Detection Approaches in Mobile Sensing. LB: Lab-Based and IW: In-The-Wild experiment.

Study	N	LB	IW	Modality (Sensor Location)	Sensors	Proxy
Morshed et al. [354]	28	✓	✓	Smartwatch (Wrist)	Accelerometer	Hand Movement
Kadomura et al. [236]	5		✓	Sensing Fork (Fork)	Single-pixel RGB color sensor	Color of the food
					Accelerometer & Gyroscope	Hand Movement
					Three-electrode Conductive Probe	Electrical conductivity of food
Gao et al. [163]	28	✓	✓	Bluetooth headset (Ear)	Microphone	Mastication
						Deglutition
Chun et al. [452]	23	✓	✓	Jawbone-mounted (Jawbone)	Accelerometer & Gyroscope	Mastication
Rahman et al. [420]	8		✓	Microsoft Band (Wrist)	Accelerometer & Gyroscope	Physical Movement
				Affectiva Q sensor (Wrist)	Electrodermal activity sensor	Psychological arousal
				Wearable microphone (Neck)	Microphone	Chewing and swallowing
				Smartphone (Smartphone)	GPS sensor	Location
Thomaz et al. [507]	28	✓	✓	Smartwatch (Wrist)	Accelerometer	Hand Movement
Bi et al. [53]	24	✓	✓	Ear-mounted Sensor (Ear)	Microphone	Chewing Sound
Ye et al. [578]	7		✓	Pebble Smartwatch (Wrist)	Accelerometer	Hand to mouth eating gestures
Mondol et al. [349]	74		✓	Smartwatch (Wrist)	Accelerometer	Hand Movement
					EMG sensor	Motion & EMG activities
Chun et al. [98]	17	✓	✓	Necklace (Nect)	Proximity sensor	Head & Jawbone movements
Thomaz et al. [505]	14	✓		Wrist-wearable (Wrist)	Accelerometer & Gyroscope	Wrist Movements
Zhang, et al. [589]	20		✓	Necklace (Neck)	Accelerometer, Gyroscope, Magnetometer, Proximity, Ambient light	Head Movement and Chewing
Bedri et al. [44]	10	✓	✓	Ear wearable (Ear)	Microphone & Proximity & IMU	Chewing
Merck et al. [333]	6	✓		Smartwatch (Wrist)	Accelerometer, Gyroscope, Magnetometer	Wrist Movements
				Google Glass (Head)	Accelerometer, Gyroscope, Magnetometer	Head Movements
				Ear wearable (Ear)	Microphone	Chewing Sounds
Dong et al. [137]	43	✓		Ear wearable (Ear)	Microphone & Proximity & IMU	Chewing
Our Study	58		✓	Smartphone (Any)	Multimodal	Behavior and Context

overeating, undereating) could be inferred with an accuracy of 83.49% in a three-class inference task using only smartphone sensing features with a one hour time window around eating episodes. In another study, Meegahapola et al. [328] showed that the social context of eating events could be inferred with accuracies above 77% for student populations in two countries. However, all these efforts are towards characterizing eating events and not detecting them.

The uniqueness of the above studies is that similar to Bisogni et al. [57], they consider eating episodes as holistic events that happen amidst different behavioral and situational circumstances, in addition to the main action of eating. Further, they have performed inferences assuming that eating episodes can be detected, including the hour of eating as a feature in inference models. Therefore, the inference can only be made once the eating events are detected. These studies build upon the premise that wearables can detect and smartphones can characterize eating events. In this chapter, we consider eating as a holistic event and attempt an eating event detection task that has not been attempted in prior work. Further, since prior work has shown that smartphones fare reasonably well in characterizing eating events, this work complements them well in showing the potential of using only smartphones for detecting eating events.

Finally, as shown in Figure 5.1, the methodology was rigorously evaluated using different sensor features, feature selection techniques, model types, and personalization techniques.

5.3 Data, Smartphone Features, and Definition of Eating/non-eating Episodes

5.3.1 Dataset

We used the MEX dataset regarding the eating behavior of young adults from our previous work [330], which was described in Chapter 2. A summary of mobile sensing features is also provided in Table A.3.

5.3.2 Ground Truth and Passive Sensing Data

The mobile app was designed to capture retrospective self-reports. It is worth noting that dietary recall techniques are common in eating behavior studies (e.g., 24H dietary recall [292, 81]). Hence, during three timeslots of the day that are a minimum of four hours apart, the mobile application sent a reminder to participants to report food intake (note that a far lesser four-hour window was used in this study to capture eating reports compared to a typical 24H recall). The ground truth responses were:

- (a) Case 1: no food intake within the last four hours.
- (b) Case 2: one food intake within the last four hours.
- (c) Case 3: two or more food intakes within the last 4 hours.

In Case 2 and Case 3, they were asked to report how long ago the last food intake occurred, and the possible answers included 0-30 min, 31-60 min, 60-90 min, 90-120 min, 120-150 min, 150-180 min, 180-210 min, and 210-240 min ago. This report helps to determine an approximate eating time (T_{anc}) as an *anchor* for the last eating episode. For example, if a self-report was done at 8.00 pm, and if the eating episode occurred 31-60 minutes ago, the approximate eating time was about 45 minutes ago (mean of 31 and 60), hence $T_{anc} = 7.15$ pm. Further, in Case 2, except for the time window corresponding to that eating episode, the rest of the times correspond to non-eating. Hence, a maximum of two non-eating events from such self-reports were randomly sampled. Furthermore, in Case 1, the last four hours would correspond to a non-eating period, and a maximum of three-time windows were randomly sampled from the last four hours as T_{anc} for non-eating events. Moreover, it is worth noting that all eating and non-eating events were chosen such that they are *non-overlapping* when sensor data are aggregated with a time window (described in the next sub-section), hence avoiding any biases in the evaluation. While self-reports were captured only three times per day, passive smartphone sensing data were captured throughout the 24 hours. The sensing modalities include the accelerometer (ACC), location (LOC), battery (BAT), screen (SCR), and application usage (APP). A summary of passive smartphone sensing features used in the study is given in Table A.3.

5.3.3 What is an Eating Event?

A $2X$ time window around T_{anc} was used to aggregate sensor data to match a self-report as given in Table A.3. According to Figure 5.2, $2X$ time window means that the inference can be done X minutes after the T_{anc} . Using a larger time window like one hour ($X = 30$ minutes) is common

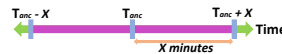


Figure 5.2: Time Window for Sensor Data Aggregation

for ubicomp studies regarding eating and drinking behavior that considers the context and behavior of participants in addition to the actual eating/drinking episode [30, 330, 55, 327]. This does not mean that the time of inference should be 30 minutes after the end of eating. In reality, it could be much less than that because eating episodes can span around 20-30 minutes. Moreover, even though most of the analysis in the next sections focuses on a one-hour time window, the time window can be changed depending on the dataset and application requirements. In Section 5.6, this is further discussed.

With the said time window, for the previous example where $T_{anc} = 7.15$ pm, sensor data would be aggregated from 6.45pm ($T_{anc}-30$) to 7.45pm ($T_{anc}+30$). In case participants reported that they did not have food during the last four hours, and if the self-report time is T_{sr} , T_{anc} was chosen carefully to make sure that a half-an-hour window was present on either side of the T_{anc} , within the non-eating time window ($T_{sr}-30 \geq T_{anc}$ and $T_{anc} \geq T_{sr}-210$). This is to ensure that there are sufficient sensor

5.4 Descriptive and Statistical Analysis of Sensor Data and Eating Events (RQ1)

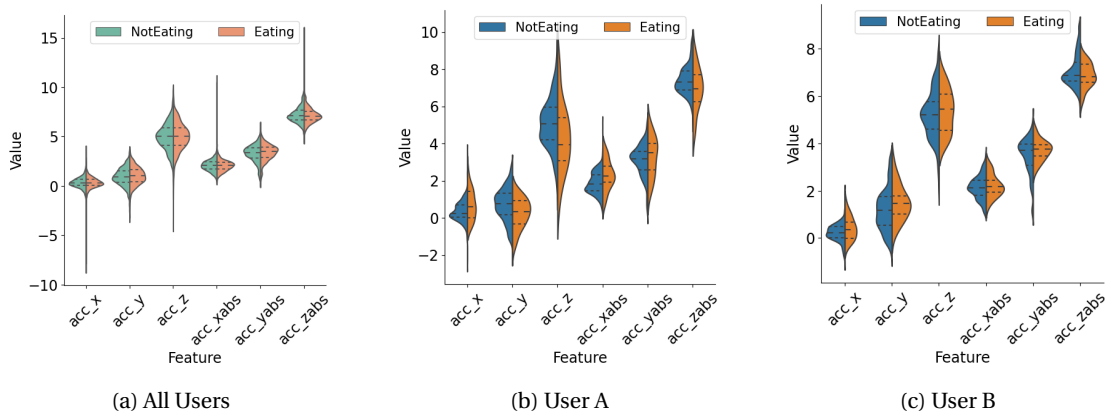


Figure 5.3: Violin plots of six selected accelerometer features for all the users and for two randomly selected users

data (one hour) in a non-eating time window to match the self-report. Again, it is worth noting that all the eating and non-eating events in the dataset contain non-overlapping sensor data. Hence, each event is mutually exclusive, and there is no data leakage. Moreover, there is no clear definition for the terms ‘eating event’ or ‘eating episode’ as these terms have been used interchangeably in different studies [47]. Therefore, for clarity, throughout this study, the one-hour time window reported by the participants as they had food during that time is referred to as an *eating event*. This event consists of the actual *eating episode* (i.e., the time period of actual eating) and of an extended time period that aims to capture the surrounding behavior and context around the eating episode, similar to prior work [30, 55, 330, 328, 327]. Finally, a one-hour time window in which food is not consumed by participants is referred to as a *non-eating event*.

5.4 Descriptive and Statistical Analysis of Sensor Data and Eating Events (RQ1)

Accelerometer Features. Figure 5.3 shows accelerometer data distributions for all users and two randomly selected users. In Figure 5.3a, accelerometer data distributions for eating and non-eating events look similar when all the users are considered in general, with minimal differences between means in all six features. However, for individual users in Figure 5.3b and Figure 5.3c, there are visible differences in accelerometer data distributions. Hence, even though the all-user distribution looks the same for eating and non-eating events, individual-level differences could be leveraged when building inference models by considering within-user differences during eating and non-eating events.

Location Features. Figure 5.4a shows a distribution of eating and non-eating events for different values of radius of gyration for all users. Although, according to the figure, the ratio of eating to non-eating is almost the same for all radius of gyration values other than between 3-4, a relative increase in eating events can be seen. This suggests that a significant amount of movement during an hour corresponds to eating events. This finding is coherent with prior work that said people might travel a sizable distance at noon in search of food, regardless of the geographical location [246, 304]. However, when distributions of two randomly selected users are considered (Figure 5.4b and Figure 5.4c), even though the distribution of *user B* is almost the same as *all-users distribution* in Figure 5.4a, *user A* has a different distribution. That user has eating events only from 0 to 2 radius of gyration values, suggesting

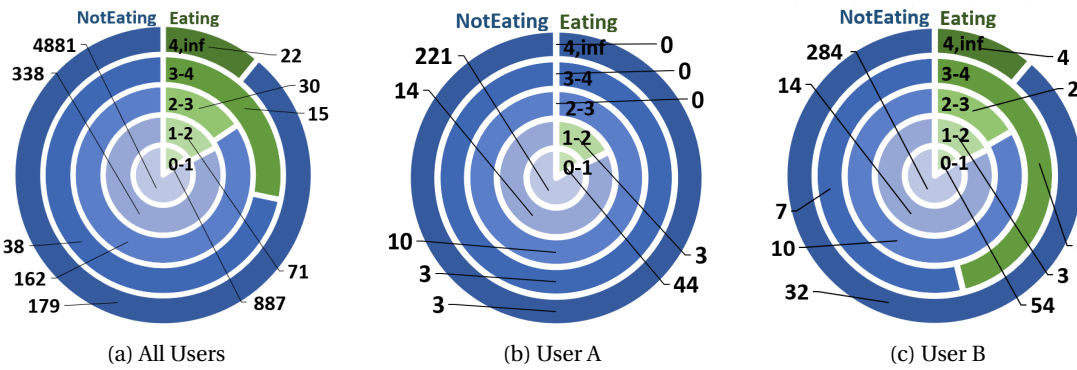


Figure 5.4: Distribution of eating and non-eating events for different radius of gyration values for all the users and for two randomly selected users

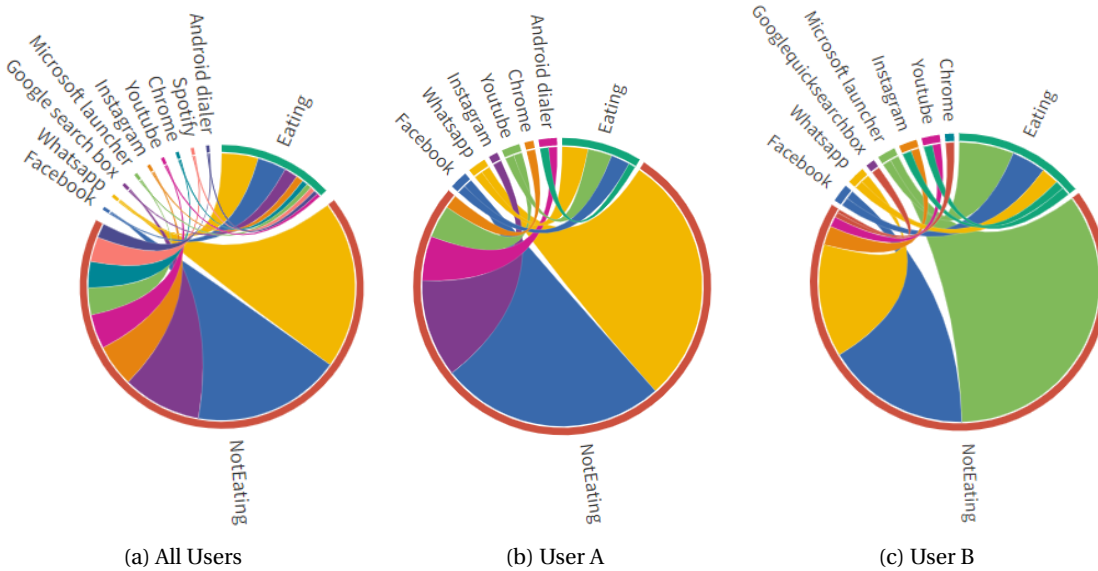


Figure 5.5: Distribution of app usage for eating and non-eating events for all users and for two randomly selected users

that while moving *user A* has not eaten. Therefore, these diagrams capture the behavioral diversity of people in terms of movement during a one-hour window and how such movements correspond to eating and non-eating.

Application Usage Features. Figure 5.5a provides the app usage distribution for eating and non-eating events for all the users. Results suggest that the proportion of app usage for a particular app could differ between eating and non-eating. For example, more people used Spotify than YouTube during eating events, whereas there was more YouTube usage compared to Spotify during non-eating periods. While these population-level statistics look interesting, individual-level app distributions for two random users in Figure 5.5b and Figure 5.5c suggest that individual-level app usage could alter from the population level. For example, *user A* has not used Instagram and Chrome while eating. They have used YouTube more than Facebook while eating whereas used Facebook more than YouTube when not eating. *User B* has also used Chrome only while not eating. These individual app usage differences are in line with prior work in mobile sensing that has shown that app usage behavior could be used to identify users [136]. Furthermore, app usage behavior has shown good performance for other eating and drinking behavior-related tasks [454]. In addition, prior work has proposed that app usage can be

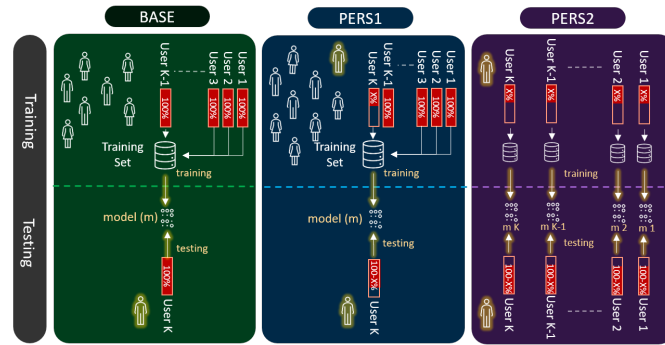


Figure 5.6: Three Phases of the Study

used to gain an understanding of user context [62]. Hence, when combined with findings in prior work, the descriptive results indicate why person-level models could be useful over population-level models to capture these fine-grained behavioral differences among people, especially when considering app usage behavior.

In summary, descriptive analysis of the dataset shows that features from passive sensing modalities could indicate eating and non-eating events, especially when combined with the time of the day. Further, results for some modalities such as APP, LOC, and ACC suggest that only considering data from a single user could capture individual-level behavioral differences with regard to eating and non-eating events, that would otherwise be not noticeable when considering population-level behaviors. In the next section, these feature-level relationships are examined and quantified in more detail using various statistical techniques.

5.5 Detecting Eating Events and Important Features (RQ2)

5.5.1 Two-Class Eating Event Inference

In this section, we used different feature group combinations to infer eating vs. non-eating events using smartphone sensing data. Scikitlearn framework [391] and python is used to conduct experiments in three phases using different model types: (1) Random Forest (RF) [117], (2) Naive Bayes (NB) [432], (3) Gradient Boosting (GB) [365], and (4) AdaBoost (AB) [457]. These models were chosen by considering the tabular nature of the dataset, the interpretability of results (e.g., getting feature importance values), and the small size of the dataset. Further, similar to recent ubicomp work [29], the Synthetic Minority Over-sampling Technique (SMOTE) [86] was used to prepare training sets for each inference task. In addition, the F1-Score and AUC scores, both with macro averaging, are reported. This would give equal emphasis to both classes, hence indicating whether both eating and non-eating classes are classified well. Moreover, the three phases of experiments are described below (named user-level, population-level, and hybrid in line with prior work [152]) and summarized in Figure 5.6.

Population-level (BASE): These models are also called population-level models. The leave-one-out cross-validation strategy that is commonly used in mobile sensing research [285, 330] is used in this phase. The objective is to train a set of users and test a user not seen on the training set. Hence, this is the base accuracy because, from a mobile sensing standpoint, this corresponds to a situation where a new user is starting to use a mHealth app, and the server does not have any data from the user.

Therefore, the machine learning model used on the user's app is a general model trained with data from other users. This is the accuracy that can be expected for a new user without any personalization.

Hybrid (PERS1): This corresponds to a situation where the server has some data from a new user to include in training a partially personalized model. However, the server does not have enough data from that user to train a separate, fully personalized model. From a mobile sensing standpoint, this corresponds to a situation where users have used a mHealth app for some time, hence generating some data for model training. Then, users use the mHealth app that contains the partially personalized model. The generated model is partially personalized because user data has been used in training the model, and the same user's data would be used in testing. When conducting experiments, it was ensured that each training split contained data from other subjects (similar to leave-one-out cross-validation) and 70% data from the target user, and the rest of the data points of that user were used to test the model.

User-level (PERS2): These models are also called user-level models. The server has enough data to train entirely personal models for each user. Hence, testing is done with a model that is trained with the same user's data. This corresponds to a fully personalized scenario from a mobile sensing standpoint [285] where users have used a mHealth app for some time and have produced enough data for the server to generate a fully personalized model. In this case, different users would have different models. Hence, the approach was evaluated by training the model using 70% data from each user, testing with the rest of the data of that user, and finally averaging the results of all the users. These models are also called user-level models.

We conducted all three experiments using several feature group combinations. The first task was to understand whether single-feature modalities could be used for the inference task. This is important because prior work has shown that having multiple models that can make the same inference could be useful for robust mobile sensing systems in situations of sensor failure [454]. For instance, a college student might turn off the location sensor during some hours because that sensor could drain the battery faster. In such situations, having different inference models that use other data sources to make the exact inference is crucial. In addition, the feature group TIME would indicate whether time alone could be a good predictor for eating and non-eating events. Hence, the experiments were conducted for feature groups LOC, SCR, TIME, BAT, APP, and ACC separately.

There are three more feature groups where all features are used. First, the ALL feature group considers all the features available for the inference task. Then, ALL-PCA used principal component analysis [564] to obtain the optimum number of principal components to get the best accuracy using all features. Then, ALL-FS used a sequential forward feature selection algorithm [312] to select the best set of features for a given inference task. For BASE and PERS1, feature group after feature selection is common for all users because there is only one model for all users. However, since there is one model for each user in PERS2, different feature groups are used to train models after feature selection.

Table 5.2 and Table 5.3 summarize the results of experiments. First, in Table 5.2, BASE results are shown for all inference models. As per the results, the best-performing models are different for different feature groups. The highest F1-score of 0.74 and AUC of 0.65 for the BASE came from the RF when using the ALL feature group. These results suggest that BASE scores for the inference tasks are moderate. Then, personalized results (PERS1 and PERS2) are included in Table 5.3. Considering space limitations, this table only contains results from RF because they provided the best performance for PERS1 and PERS2 in most cases. There is a bump in AUC scores for PERS1 compared to BASE in most cases. In addition, for ALL, ALL-PCA and ALL-FS, F1 scores increased by over 6%. ALL-FS with an F1 feature

5.5 Detecting Eating Events and Important Features (RQ2)

Table 5.2: Averaged F1-score (F1) and AUC of 58 users, calculated using four different models, for the BASE of eating event detection task. For ALL-PCA, the number in square brackets (e.g. [c=6], etc.) indicates the number of principal components used for the inference. For ALL-FS, the notation in square brackets (e.g., [F3], [F7], etc.) indicates the name of the feature group. More details about feature groups, including the list of features in each group, are given in Appendix (Table A.1).

Feature Group (# of Features)	RF F1, AUC	NB F1, AUC	GB F1, AUC	AB F1, AUC
Majority Class	0.00, 0.50	0.00, 0.50	0.00, 0.50	0.00, 0.50
LOC (1)	0.72, 0.50	0.72, 0.48	0.72, 0.51	0.72, 0.52
SCR (2)	0.73, 0.52	0.72, 0.54	0.72, 0.53	0.72, 0.54
TIME (3)	0.72, 0.51	0.72, 0.54	0.72, 0.52	0.72, 0.53
BAT (6)	0.71, 0.51	0.72, 0.50	0.72, 0.52	0.72, 0.53
APP (10)	0.72, 0.51	0.70, 0.51	0.72, 0.51	0.72, 0.51
ACC (18)	0.72, 0.55	0.72, 0.55	0.72, 0.54	0.72, 0.52
ALL (40)	0.74, 0.65	0.70, 0.56	0.72, 0.59	0.72, 0.56
ALL-PCA	0.73, 0.53 [c=5]	0.72, 0.53 [c=1]	0.72, 0.52 [c=4]	0.72, 0.52 [c=4]
ALL-FS	0.74, 0.65 [F1]	0.71, 0.51 [F7]	0.72, 0.58 [F3]	0.72, 0.50 [F7]

Table 5.3: Averaged F1-score (F1) and AUC calculated using random forest classifiers, for BASE, PERS1, and PERS2 of the eating event detection task of 58 users. For ALL-FS, the notation in square brackets (i.e. [F1]) indicates the name of the feature group. More details about the feature group, including the list of features in the group are given in the Appendix (Table A.1).

Feature Group (# of Features)	BASE F1, AUC	PERS1 F1, AUC	PERS2 F1, AUC
Majority Class	0.00, 0.50	0.00, 0.50	0.00, 0.50
LOC (1)	0.72, 0.50	0.72, 0.52	0.74, 0.59
SCR (2)	0.73, 0.52	0.72, 0.52	0.73, 0.57
TIME (3)	0.72, 0.51	0.71, 0.52	0.72, 0.52
BAT (6)	0.71, 0.51	0.72, 0.56	0.73, 0.60
APP (10)	0.72, 0.51	0.72, 0.51	0.71, 0.51
ACC (18)	0.72, 0.55	0.75, 0.66	0.79, 0.73
ALL (40)	0.73, 0.65	0.81, 0.81	0.80, 0.73
ALL-PCA	0.73, 0.53 [c=5]	0.79, 0.72 [c=4]	0.81, 0.76
ALL-FS	0.74, 0.65 [F1]	0.81, 0.81 [F1]	0.85, 0.81

set provided the best F1 score of 0.81. Finally, the best F1 score out of all inference tasks (0.85) came from PERS2 with the ALL-FS feature group. Similar patterns could be seen with AUC scores. Since PERS2 averages results from 58 different users, the model for each user had different feature sets after feature selection that provided the best performance. Section 5.5.2 discusses more details about these feature sets. This suggests that while user-level models could help detect eating events with reasonably high performance, depending on the user, it is better to select the best set of features using a feature selection technique. In addition, hybrid models perform fairly well compared to user-level models for most feature groups.

5.5.2 Feature Importance for Eating event Detection (RQ2)

In order to study individual differences further, Figure 5.7 gives BASE feature importance values (Figure 5.7a) and PERS2 feature importance values for three randomly selected users (Figure 5.7b, Figure 5.7c, and Figure 5.7d) based on Gini importance of RF classifiers. These figures illustrate how

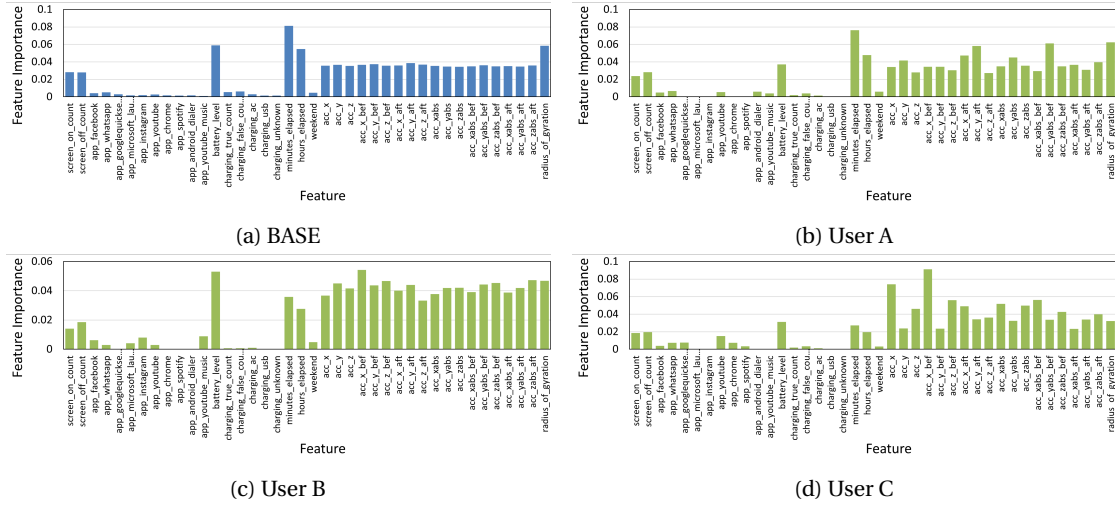


Figure 5.7: Feature importance values from RFs for (a) BASE and (b)–(d) PERS2 models from three randomly selected users.

Table 5.4: Personalized eating event detection accuracy breakdown for random five users in PERS2 with ALL-FS. F1-score (F1) and AUC are shown. The top performing feature group, and modalities included in the feature group are shown with \checkmark . F1-score Bump indicates the F1-score of BASE (Blue), increase in F1-score from BASE to PERS1 (Green), from PERS1 to PERS2 (Orange) for each user.

User	F1, AUC	LOC	ACC	APP	BAT	SCR	TIME	F1-score Bump
1	0.92, 0.82		\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	
2	0.84, 0.59		\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	
3	0.88, 0.91	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	
4	0.81, 0.48			\checkmark		\checkmark		
5	0.98, 0.99		\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	

datasets from different users have varying features that could discern between eating and non-eating. For example, BASE (0.081, 0.055) and *user A* (0.076, 0.048) had high values for the time of the day (minutes_elapsed, hours_elapsed). However, for *user C* (0.027, 0.019), the values for those features are low. Moreover, BASE (0.059) and *user B* (0.053) had high values for battery_level, whereas for *user A* (0.037) and *user C* (0.031) had lower values compared to other features. Another feature that had a clear difference was radius_of gyration. It was high for BASE (0.058), *user A* (0.062), *user B* (0.049), and low for *user C* (0.031). On a feature group level, the ACC feature group had very similar values across all features [0.035,0.040] in BASE, whereas *user A* and *user C* had highly varied feature values, where some ACC feature values are comparatively higher than other features values. These findings point toward why PERS2 models provided significantly higher accuracy than the BASE. This is because PERS2 captures user-specific behaviors regarding eating events.

5.5.3 Effect of Personalization on Individuals

Table 5.4 shows the PERS2 results breakdown for five random users. The maximum F1 score attained by a user here with PERS2 was 0.98. In addition, for five users here and, in fact, for all 58 users for whom PERS2 models were trained, after feature selection, no two users shared the same feature group, which resulted in the highest accuracy. For example, *user 3* used features from all feature groups in the model

to attain an F1 score of 0.88, whereas *user 4* used a selected set of features from only APP and SCR to attain an F1 score of 0.81. In addition, the F1-score bump representation shows that the PERS2 score of different users increased by different amounts compared to BASE and PERS1 results for individual users. For example, for *user 5*, the BASE to PERS2 increase was 24%, whereas for *user 1*, the increase was 2.6%. This shows how the effect of personalization could vary from user to user depending on the chosen features. It is also worth noting that for no user, both the F1-score and AUC decreased when going from BASE to PERS1 to PERS2.

5.6 Discussion and Limitations

We now discuss the implications as well as the limitations of our work.

Time Window for Eating Event Detection. It is worth noting how a chosen time window affects the inference task. Similar to other inferences of human activities in ubicomp research [55, 330, 468, 327, 30], eating event inference is carried out with a time-window based approach, with fixed or variable frequencies, depending on the application. For example, typical activity recognition algorithms that use time windows of 3-10 seconds could run once every second with overlapping sensor data segments or run every thirty seconds to generate sparser inferences. In prior work, Bae et al. [30] used thirty-minute, one-hour, and two-hour time windows for drinking event detection. Meegahapola et al. [330] used a one-hour time window to detect subjective food consumption level inferences. This research is in line with Bisogni's contextual framework [57], which not only considers eating/drinking events but the whole context around them. Similarly, the models discussed in this chapter could also be run with different time intervals, depending on the use case. For example, with a one-hour time window, if the inference task was performed for $T_{anc} = 2.15\text{pm}$ at 2.45pm, and run again for $T_{anc} = 2.17\text{pm}$ at 2.47pm, the granularity would be higher, compared to running the second inference after another hour for $T_{anc} = 3.15\text{pm}$ at 3.35pm. In addition, as shown in the previous section, this inference could be run for shorter time windows, obtaining reasonable inference accuracies. Even though the one-hour time window performed better for this dataset, shorter time windows might perform better for other datasets. Coming back to the previous example, with a shorter twenty-minute time window, for $T_{anc} = 2.15\text{pm}$, the inference could be made at 2.25 pm, hence reducing the time between T_{anc} and inference time. Hence, the time window and inference frequency should be chosen depending on the use case and the available data. Future work could explore how these aspects affect inference performance in more depth. Furthermore, we produced results for different time windows ranging from ten to ninety minutes (not included here due to space limitations). The best average results were obtained for the sixty-minute case, closely followed by fifty minutes and seventy minutes. Hence, only the results for the sixty-minute case were included throughout the chapter. However, as mentioned earlier, different time windows could be used in future work depending on the dataset and features.

Effect of Automatic Inferences on Battery Life. Many prior automatic dietary monitoring inference systems have used wearable devices, which required the device to run inference models continuously at fixed or variable intervals [47]. This could have a significant impact on the battery life of wearables. In addition, wearable devices rarely work alone as standalone devices. Typical wearables connect to smartphones, and smartphone apps are used to provide feedback and interventions to people. In the case of reminding users to fill in self-reports, users typically need to fill in food diaries on the phone as it is challenging to use a food diary on a wearable device. Hence, maintaining this connection with the smartphone could also affect the battery life of both the phone and the wearable. Eating event detection on the phone would reduce most of these issues. In the results section, we showed that

models performed well for ALL, interaction sensing modalities (INTSEN - SCR, APP), and continuous sensing modalities (CONSEN - LOC, BAT, ACC) for both PERS1 and PERS2. While CONSEN would consume far more battery life as it uses accelerometer and location-related features, INTSEN only uses app usage and screen usage, which are just phone usage logs. This makes INTSEN inferences computationally cheaper than those using CONSEN or ALL. However, the set of features available in the dataset for APP and SCR are not rich, and future data collection efforts could look into collecting richer datasets that could further increase performance. Overall, INTSEN could be a low-cost alternative to phone-based eating event tracking, compared to CONSEN and wearable-based tracking. In this sense, INTSEN could be used as a low-power sensing modality to trigger more accurate, high-power CONSEN or ALL-FS models, as discussed in [47].

Interpretation of Eating Event Detection. There are many ways in which an eating event can be interpreted. In this chapter, an hour period that contains an eating episode is considered as an eating event. This was done with the assumption that there is a time period in which participants prepare for eating (going to the place of eating, preparing food, etc.) and move on to other activities after the eating event. Hence, the objective was to capture all such behaviors using sensing modalities. This is in line with the idea of capturing holistic food consumption events similar to prior work in smartphone sensing [55, 57, 30, 327, 330, 327]. However, some previous studies only investigated detecting eating gestures from hands, neck, chews, or mastication, and thus, in many such cases, eating episodes are detected by sensing different phenomena (e.g., hand motion in wrist-based sensing, chewing in earable-based sensing, or jaw bone movement in necklace-based sensing). In addition, some studies have modeled eating event detection as a time-series data analysis problem. In contrast, the approach discussed in this chapter did not consider the time-series nature of the data, and we extracted eating and non-eating events using short-term retrospective self-reports to model inferences using a tabular dataset (Figure 5.1). Due to these differences in sensing and modeling techniques, there is no direct comparison possible with previous work. It is worth noting that there exist subtle differences in how eating events/episodes are defined in different studies, and results should be interpreted with caution by understanding the essence of each individual work.

About to Eat, Eating Now, or Just Ate? Even though previous research has used terms such as eating event detection, eating moment recognition, and eating detection, there is a fundamental difference between *what* is being sensed and *when* it is being sensed. In addition, depending on when sensing occurs, there are differences w.r.t. how such sensing techniques can be used to benefit users. For example, Rahman et al. [420] explicitly mentioned that they are predicting about-to-eat moments to predict eating episodes before they occur. Such inferences are useful for interventions because they can motivate a user not to eat or used to ask users to control their eating amount before an episode occurs. However, this method might not be used for automated food tracking because predicted eating episodes might or might not happen due to interventions. Furthermore, other studies attempt to detect the episode during the actual eating action [507, 44, 98]. In such cases, it is challenging to ask users not to eat because they are already doing it. However, these approaches could be used for certain interventions and automated food tracking. Additionally, assume that there is a need for users to complete mobile food diaries. In this case, it might be less desirable to trigger reminders at the moment itself, as prior work has shown that people do not appreciate it when they are disturbed during eating moments [325, 55]. On the other hand, in our approach, an eating event is detected retrospectively and approximately less than thirty minutes after the end of eating. The proposed technique is less useful for in-the-moment interventions. However, it could be used for interventions regarding future eating events. The proposed technique would also work for automated food tracking, as it would allow reminding users to fill in food diary reports within a short time after eating, hence not disturbing them

during actual eating times. This could also reduce recall bias because it would not be too far from the actual eating episode. Hence, as explained, depending on the time events are sensed, the most appropriate use cases would be different. These aspects should be considered when building future sensing techniques for mobile food diaries.

Other Informative Features. It is important to acknowledge that it was challenging to generate interpretable and meaningful features from the dataset. First, it was clear from the results that ACC features were informative of eating and non-eating events. This means that activity levels are helpful in discerning between the two classes. However, the only available features were statistical features generated from the accelerometer axes, which are hard to interpret. More easily interpretable features such as activity level (i.e., step count) and activity type (i.e., walking, sitting, driving, running, cycling, etc.), which can be captured using activity recognition engines in modern smartphones, were not available in the dataset. Future work should look into capturing such interpretable features, which can be generated with low-power consumption on a smartphone. It was also impossible to determine app usage times for features in the APP group because such features were not available in the dataset. This is another aspect that could be improved when creating mobile sensing applications for future studies. Another challenge in the data filtering and processing phase was missing data, especially from the location sensor. As the location sensor consumes a high amount of power, there is a tendency for participants to turn off this sensor. Finally, researchers could examine other modalities such as touch and typing events, notification clicking behaviors, and continuous sensing modalities such as ambient light sensors, which are typically available in modern smartphones and have shown promise in other smartphone sensing-based behavioral modeling tasks [325].

Importance of Diversity-Awareness. Depending on demographic attributes, lifestyle, and culture, eating behaviors can significantly vary [438, 304]. Early work documented differences between men and women related to eating behavior [438]. Other statistics show country-based differences. For instance, people in European countries like France, Italy, and Spain spend more time a day on average eating and drinking than people in the United States [319]. Our work has studied the eating behavior of a group of college students in Mexico. The results can not be assumed to represent the eating behavior of other age groups from the same country or people from other countries. Prior work has highlighted the importance of considering diversity awareness in social platforms that use mobile sensing and machine learning [247, 458]. Hence, future work needs to be carried out for different age groups and countries.

Additional Assumptions and Implications.

In the first place, our work assumes that human behavior does not change significantly over time with respect to app usage, screen usage, and activities. However, this is not always the case, as the lifestyle and behavior of people could indeed change over a period of time. Future work could look into using time windows (e.g., one month, two months, etc.) when selecting data for training models. Unfortunately, the dataset used in our work did not allow us to capture such behavioral changes because the data collection involved only a few weeks. Examining such temporal behavioral change was not a goal of this work and is out of the scope of the chapter.

In the second place, we assumed that there is no relation between eating and non-eating events of the same person, even within the same day. However, this might not be the case because eating and not eating are temporally linked, e.g., long periods of not eating could increase the possibility of eating. On the flip side, eating a meal right now would increase the possibility of non-eating in the next few hours. With the studied dataset, it was not possible to test such phenomena because only three self-reports

were collected per day. Therefore, all eating and non-eating events might not be present. This topic is open for future investigation. In the third place, another limitation is that eating events could be under-reported for convenience or other reasons (i.e., a person reporting non-eating in a case when an eating event indeed occurred). Even though this might add some noise to the dataset, prior work also suggests that it is difficult to capture and verify all self-reports during real-life experiments [468, 55, 454, 285].

As a fourth issue, general machine learning models with feature level fusion were used in our work, similar to many prior smartphone sensing in-the-wild deployments [55, 454, 468, 30]. Whether other decision-level fusion techniques could be used for this task is open for future investigation. Finally, regarding statistical analyses, even though we used a reasonably larger sample size compared to other ubicomp studies, future work in this domain could examine larger sample sizes to examine the issues of generality and personalization in more depth. Note also that the Bonferroni correction [531] was not used when calculating p-values even though there were multiple comparisons in the statistical analysis. So, the results regarding p-values need to be interpreted with caution.

5.7 Conclusion

In this chapter, we examined the eating behavior of 58 college students in Mexico using self-reports and passive smartphone sensing data. First, it was shown that the time of the day, and features from modalities such as screen usage, accelerometer, app usage, and location are indicative of eating and non-eating events. Then, it was shown that eating and non-eating events can be inferred with an AUC of 0.65 (F1-score of 0.75) using a population-level model, which can be further improved up to an AUC of 0.81 (F1-score of 0.85 for user-level and 0.81 for hybrid models) using personalization techniques. Using feature importance values from classification models and sequential forward feature selection techniques, our work showed that best-performing, user-level models for different users rely on different feature groups. These findings are encouraging for future mobile food diary apps that are context-aware for both user- and population-level use cases.

6 Inferring the Mood-While-Eating with Community Based Personalization

The interplay between mood and eating has been the subject of extensive research within the fields of nutrition and behavioral science, indicating a strong connection between the two. Previous work relied on questionnaires and mobile phone self-reports to investigate the relationship between mood and eating. Additionally, phone sensor data have been used to characterize both eating behavior and mood, independently, in the context of mobile food diaries and mobile health applications. However, limitations within the current body of literature include: *i*) the lack of investigation around the generalization of mood inference models trained with passive sensor data from a range of everyday life situations, to specific contexts such as eating, *ii*) no prior studies that use sensor data to study the intersection of mood and eating, and *iii*) the inadequate examination of model personalization techniques within limited label settings, as we commonly experience in mood inference (i.e., very less negative mood reports compared to positive or neutral reports). In this chapter, we sought to examine everyday eating behavior and mood using two datasets of college students in Mexico ($N_{MEX} = 84, 1843$ mood-while-eating reports) and eight countries ($N_{MUL} = 678, 329K$ mood reports incl. 24K mood-while-eating reports), described in Chapter 2, containing both passive smartphone sensing and self-report data. Our results indicate that generic mood inference models decline in performance in certain contexts, such as when eating. Additionally, we found that population-level (non-personalized) and hybrid (partially personalized) modeling techniques were inadequate for the commonly used three-class mood inference task (positive, neutral, negative). Furthermore, we found that user-level modeling was challenging for the majority of participants due to a lack of sufficient labels and data from the negative class. To address these limitations, we employed a novel community-based approach for personalization by building models with data from a set of similar users to a target user. Our findings demonstrate that mood-while-eating can be inferred with accuracies above 80.7%, surpassing those attained with traditional methods. The material of this chapter is under review.

6.1 Introduction

"Mood" can be defined as "a conscious state of mind or predominant emotion" [554]. Even though it is an internal, subjective state, it often can be inferred from behaviors of individuals [468, 285]. Mood affects many facets of our daily lives. While positive moods increase the likelihood of sociability, creativity, mating, and planning [133], negative moods could alter such behaviors in the short term, and also provide a way to adverse health outcomes in the long term [215, 31]. Another important aspect of our everyday lives that could get altered due to positive or negative mood, is eating (food

choice, overeating and undereating, speed of eating, etc.) [46, 330]. For example, prior work has shown that extreme moods could trigger overeating or undereating depending on contextual circumstances such as location, app usage, food type, etc. [330]. Due to these reasons, understanding the causes and contexts of mood and quantifying mood has been an active research topic in mobile sensing during the last decade [325]. While some studies have used audio sensors in the smartphone [418], others used phone usage patterns [285] and multi-modal sensors [324, 468, 305]. Recent work has also used sensor data from wearables to recognize different moods [280]. These studies show promise for building mobile health systems that could provide feedback and interventions by considering user moods more meaningfully. However, a majority of mood-related studies have focused on everyday life behavior, and there is not much knowledge on the location, social context, and concurrent activities (i.e., studying, working, eating, etc.) done while mood reports were captured using self-reports, as studies that inspect such aspects are scarce [355].

The use of generic one-size-fits-all models to infer everyday human behavior has been a topic of research in the past [325]. However, recent studies in the field of mental well-being and mobile sensing have revealed that such models may not be effective in different contexts, such as varying countries and time periods, due to distributional differences in data, otherwise known as domain shift [324, 569]. In contrast, the development of context-specific models to infer human behavior has also been explored. For instance, Morshed et al. [353] examined the relationship between stress and the workplace utilizing mobile sensor data and self-reported measures. More recent studies advocate the use of passive sensing in the context of psychological well-being during work [369]. Some studies emphasized the need to study mood during the specific context of driving [601, 520]. All these studies, in a way, advocate the idea of studying mood in specific situated contexts. Furthermore, a recent study utilizing phone sensor data gathered from eight countries to evaluate the impact of geographical diversity on mood inference models concluded that country-specific models trained and tested within the respective country perform better than generic multi-country models [324]. Furthermore, previous research has established that longitudinal behavior models lack generalizability across time [569]. Despite the examination of context-agnostic mood inferences in prior studies [285, 468], the effectiveness of such inferences in specific contexts, such as eating occasions, has not been explored before. Thus, further research in this direction is crucial for gaining a deeper understanding of the mood during eating as it would enable the development of more robust passive mood-tracking systems that can generalize well to diverse everyday life occasions, including meal times, which play a significant role in overall health and well-being.

Mental well-being and eating behavior have been the subject of numerous studies, both individually and in conjunction with one another. For example, studies have been conducted to examine the impact of eating behavior on mood [96], obesity and mental health [493], and the practices, psychological states, and routines surrounding eating [551]. The plethora of studies that have investigated the connection between mood and eating suggest that understanding the role of mood in eating, or vice versa, could potentially aid in a deeper understanding of eating behavior and, in turn, lead to more effective practices, interventions, and treatments [170]. Previous research has indicated that negative moods may act as a precipitating antecedent for bulimic eating episodes, while positive moods have been associated with a reduced probability of unhealthy eating episodes [96]. Furthermore, many studies have reported that negative moods may be predictive of disordered eating episodes in college student populations [232, 6, 414, 188]. One study even suggested that mood shifts are the most accurate predictor of disordered eating in college students [189]. Hence, it is worth noting that, in the context of mobile food diaries, having accurate mood predictions around eating times is of great value to track holistic eating behavior, that not only tracks what people are eating—but the overall mental state and

contextual aspects while eating [57, 55, 355, 330, 328].

Now, let's consider smartphone sensing studies that examined eating and mood. Smartphone sensing has demonstrated potential in various mood-related tasks, such as inferring positive and negative moods [468], high and low pleasure and activeness [285], mood instability [356], and displeasure, tiredness, and tension [305]. Similarly, it has also shown promise in eating-related tasks, including inferring eating episodes [467, 507], meal or snack episodes [55], the social context of eating [328], and overeating [330]. However, comparatively less attention has been given to the intersection of mood and eating. Nevertheless, studying this intersection could aid in the development of mobile health applications and mobile food diaries that are context-aware. As an example, the ability to infer an individual's mood during eating events could aid in the provision of targeted feedback or interventions. Previous research has demonstrated that passive sensing-based mobile food diaries can successfully identify meal or snack eating episodes [55]. As we all know, snacking often is unhealthy. However, even if a mobile food diary is able to provide interventions during periods of vulnerability [362] (i.e., a moment of high susceptibility to the negative health outcome – in this case, the person is eating or going to eat a snack), the receptivity (i.e., the ability of the individual to receive and process an intervention) of the individual may vary depending on both internal factors, such as mood and external factors, such as the social context in which the eating event occurs. Therefore, while this is just a single example, incorporating an understanding of an individual's mood during eating events into mobile health applications could aid in the development of more effective and personalized interventions for managing food intake in many different ways.

Finally, in terms of machine learning-based modeling, the traditional approach to mobile sensing studies has been to employ one-size-fits-all models, which have been found to be effective in experimental settings [568, 240, 24, 324]. However, in real-world scenarios, such models often require personalization techniques to account for the heterogeneity of user behaviors [240, 231, 323]. The process of personalization requires a significant amount of data, which may not be available in the initial stages of deployment or even after some time, if users do not provide adequate self-reported data as ground truth [285]. This is referred to as the "cold start problem" or "user-held out" in mobile sensing research [240, 228]. Therefore, developing personalized mood inference models that are robust to user heterogeneity and lack of sufficient data, remains a challenge. In light of these considerations, we pose the following research questions.

RQ1: Do generic mood inference models (trained with all available data regardless of the activity performed by users while reporting mood) work well for specific contexts such as eating (i.e., mood-while-eating)?

RQ2: Can the self-reported mood of college students during eating be inferred using smartphone sensing data with population-level (non personalized), hybrid (partially personalized), and user-level (fully personalized) models? What are the challenges of making such inferences?

RQ3: What measures should be taken to tackle issues such as lack of individual data and cold-start problem in mood-while-eating inference, in building personalized models? Can a community-based model overcome such issues?

By addressing the above research questions, this chapter provides the following novel contributions:

Contribution 1: We conducted an analysis using the MUL dataset collected from 678 participants in multiple countries. This dataset consists of mobile sensing data of participants' everyday moods and

Table 6.1: Terminology and description regarding different model types and Mood-while-Eating. The degree of Personalization increases when going from Population-Level to User-Level.

Terminology	Description
Population-Level	The set of users present in the training set and the testing set are disjoint. This represents the case where the model was trained on a subset of the population, and a new set of users joined the system and started using its model. These models do not use end-user data in building the model and are built with the assumption that a single model will generalize well for a large number of individuals. In practice, this is similar to the leave-one-user-out/leave-k-users-out strategy.
Hybrid	The sets of users in the training and testing splits are not disjoint. Part of the data of some users present in the training set is used in the testing set. This represents the case where the model was trained on the population, and the same people whose data were used in training continue to use the model as a system users. Hence, models are partially personalized. Therefore, models use data from both the individual and others in training. These can also be known as partially personalized models. However, depending on the data coming from others and inter-subject variability, the degree to which personalization works could differ. In practice, this is similar to the leave-one-sample-out/leave-k-samples-out strategy.
Community-Based	Builds on top of hybrid models and uses a part of the target user's data and data from other users who are similar to the target user (a subset of the population used in hybrid models) in training models. These models do not need a lot of data from the target user for personalization because they also rely on users similar to the target user.
User-Level	Also known as fully personalized models that only use target user's data in both training and testing. These models work well for individuals with low inter-subject variability, hence generalizing well for a specific individual. However, these models need a lot of data from each individual for personalization.
Mood-while-Eating	The mood-while-eating corresponds to the instantaneous <i>valence</i> [445] during eating events as reported by study participants on a five-point scale (from very positive to very negative), reduced to a three-point scale corresponding to positive, negative, and neutral classes. Our definition follows prior ubiComp work [285, 468] that used valence for mood inference. The key difference is, we focus on specific eating events and related moods. For more details, see Section 6.2.1.

behavior. First, we trained a generic three-class (positive, neutral, negative) mood inference model. We found that this model does not work similarly across all contexts (i.e., walking, resting, eating, working, etc.) and favors certain contexts ahead of others due to many reasons such as the exigent nature of activities, prevalence, etc. In fact, for the eating context, the mood inference performance was below the overall performance. In addition, when increasing the representation of mood-while-eating reports in the training set for the generic mood inference model, the performance of the model on mood-while-eating reports on the testing set increased. However, overall performance declined, indicating the difficulty for the model to generalize to mood reports captured during different situated contexts. These results point towards the need for context-specific models for mood-while-eating inference.

Contribution 2: We conducted an analysis of the MUL dataset and, in addition, the MEX dataset regarding their everyday eating behavior. Building on prior work in nutrition and behavioral sciences that emphasized the need to study eating behavior and mood in greater depth, we defined and evaluated a novel task: mood-while-eating inference (Table 6.1). This three-class inference task attempts to infer positive, neutral, and negative moods associated with eating events. We show that population-level models (non-personalized) do not generalize well to unseen individuals (accuracy of 36.9% for the Mexico dataset and 43.6% for the multi-country dataset). Then, we show that hybrid models (partially personalized) can perform better than population-level models, with an accuracy bump of 27.4% for the Mexico dataset and 6.7% for the multi-country dataset. We also discuss how training user-level models is difficult in both datasets, with the lack of individual data from a majority of users for the negative class.

Contribution 3: We show that personalization is a key component in achieving higher accuracies because aspects such as mood that are subjective are hard to be generalized using one model that fits all. In addition, we propose an approach of Community-Based personalization, building on similarities of users to create a unique community for each target user (Section 6.4.3) for personalized inferences. We show that our approach achieves an accuracy of 80.7% for the Mexico dataset and 65.9% for the

multi-country dataset, surpassing the accuracies obtained using the traditional hybrid model by margins over 15%.

The chapter is organized as follows. In Section 6.2, we describe the background and related work. Then, we describe the dataset and features used in Section 6.3. In Section 6.4, we present tasks and approaches of RQs. Then in Section 6.5, we evaluate the inference tasks of each RQ and conclude the chapter with the Discussion in Section 6.6, and the Conclusion in Section 6.7.

6.2 Definitions, Background and Related Work

6.2.1 Defining Mood-While-Eating

Previous research has employed various methods for capturing mental wellbeing-related attributes, including mood, using continuous scales [468], two-point scales [502], and seven-point scales [60], and has applied them in two/three-class inferences and regression tasks. In this chapter, the used datasets reported mood on five-point scales, similar to prior work [592, 285]. The responses were: (1) in a very negative mood; (2) in a slightly negative mood; (3) in a neutral mood; (4) in a slightly positive mood; and (5) in a very positive mood. For the proposed three-class classification task, responses 1 and 2 were classified as negative, 3 as neutral and 4 and 5 as positive. As such, throughout this chapter, we consider these three levels as a fundamental construct in relation to self-reported mood-while-eating, in line with previous studies [540, 482, 324]. Moreover, the limitation with the number of data points in the very negative and slightly negative mood classes was taken into consideration while classifying the 3 classes. Additionally, there is literature that suggests inferring self-reported mood [495] and normalized mood (i.e., since users report their mood differently, users have individual mood distributions and the user is said to be in a positive mood if a user reports a value higher than the median value of the past distribution.) [468]. While both of these approaches have distinct advantages, in this chapter, we infer self-reported mood without normalization, as the use of a Likert scale makes it difficult to normalize based on individual users or populations. As previous studies have demonstrated [131, 459], valence may explain a greater portion of a consumer's emotional response than arousal. For example, as highlighted in [459], the authors attempted to predict the food preferences of children correlated with arousal and valence, gathered using an emoji-based data collection method, and were able to show a high correlation between food selection and emojis, which describe valence levels (positive, negative, and neutral mood). However, even though we acknowledge that both valence and arousal could affect eating behavior, in this chapter, we only consider valence due to the unavailability of arousal labels in the dataset. Considering all these aspects, *the mood-while-eating corresponds to the instantaneous valence during eating events as reported by study participants, converted to a three-point scale corresponding to positive, negative, and neutral classes.*

6.2.2 Mood and Food Consumption

Many prior studies associate mental or emotional state with eating behaviors [253]. A plethora of studies have proposed similar ideas saying mood could influence food choice [75, 171, 306]. There have been repeated findings that negative mood or stress increases 'unhealthy' food choices rich in sugar or fat [10, 49, 549] and decreases the 'healthy' or fresh food choices [49]. While relatively fewer studies have reported the influence of positive mood on food intake, the association between positive mood and caloric intake has also been reported. The association between positive mood

and overeating has been observed in regression analysis [64] and controlled laboratory experiments [148]. In connection to possible effects of mood as well as the bi-directional and multi-dimensional association between mood and eating behavior, there has been a series of studies on mediators around emotional eating, ranging from the comforting effect of palatable food for high emotional eaters [522] to the role of social context in inducing emotional eating behaviors [16, 426]. In the mental health and wellness domain, the bi-directional relation between mental disorders and unhealthy eating behavior has been richly investigated, using eating behavior as the indicative measure of mental health. For example, Johnson et al. [232] and Pyle et al. [414] have observed the mood changes associated with bulimic patients. Moreover, some studies were conducted to get feedback and analyze how episodes of ravenous overeating with their daily unhealthy eating behavior symptoms [6]. Two other studies were carried out to monitor the association between binge eating with depression, stress and dietary restraints [188, 189]. All these studies show how widely the association between mood and eating behavior has been investigated, in the medical or clinical domain. However, it is also observable that the majority of the studies focus on classical methods, such as laboratory experiments and mood charting, or clinically diagnosed target groups, and less attention to the everyday association between food and mood. Although there have been recent ideas that relate mood and eating practice via online platforms [84] or mobile technologies [23], a research gap lies in investigating the relationship between mood and food in mobile sensing.

6.2.3 Mood and Smartphone Technologies

Prior studies in mental health have used mobile technologies ranging from text messaging [11] to multi-modal sensors [83] for mood tracking. In its early days, the use of mobile apps as a mood chart for the clinical purpose has been addressed [317]. Regarding users of mood-tracking apps, Schueller et al. [460] mentioned that users were primarily motivated by negative events or moods that prompted them to engage in tracking moods to improve the situation. Aligned with their work, we can find a plethora of research that proposed mood-tracking smartphone applications as a proxy to depressive symptoms [77, 361] for various demographic groups, including adolescents [245], university students [278, 544], clinically diagnosed or high-risk populations [543, 149]. In all these studies, moods are tracked via self-reports that involve user engagement in mobile systems. Regarding the efficacy of self-reported mood tracking, Bakker et al. [32] and Birney et al. [56] showed through randomized controlled trials that engaging in mobile mood self-monitoring increases emotional self-awareness that results in reducing depressive and anxious symptoms for the clinically depressed population. Besides clinical research, moods in mobile technologies are often associated with more than one behavioral mediator. In one study, Glasgow et al. [176] showed how travel choices, social ambiance, and destinations are related to moods using mobile experience sampling methodology and GPS location tracking. In another study, Carroll et al. [80] captured how individuals' moods are associated with emotional eating behaviors. In these studies, moods and mediator behaviors are captured via self-reported, ecological momentary assessment (EMA) responses. Recent works in mood and smartphone technology leverage EMA responses with unobtrusive smartphone sensing. One approach is to associate mobile sensor data with mediators. Here, mobile data serves as a descriptive proxy to concurrent activities or environments. Some of them includes physical movements [273] smartphone usage [78, 600], social contexts [586, 94], and sound [418]. Another approach is to leverage passive sensor data and machine learning to recognize negative moods [149, 468] or provide real-time intervention for depressive states [73]. Such advances in mobile technologies in mood monitoring provide more integrated information, opening the possibility of understanding various human behaviors as holistic events.

6.2.4 Eating as a Holistic Event

Bisogni et al. [57] proposed a contextualized framework for eating and drinking events with eight interconnected dimensions after studying why and how people consume food and drinks. Prior studies in mobile sensing have used this framework to model eating, and drinking behaviors using sensor data [454, 55, 330]. This framework is backed by the idea that eating is guided by behavioral and situational factors. The eight dimensions are: (1) food or drink: type, source, amount, and how it is consumed; (2) time: chronological, relative experienced; (3) location: general/specific, food access, weather/temperature; (4) activities: nature, salience, active or sedentary; (5) social setting: people present, social processes; (6) mental processes: goals, emotions, and moods; (7) physical condition: nourishment, other status; and (8) recurrence: commonness, frequency, what recurs. Many studies have proposed similar ideas saying psychological aspects [477, 205, 210], activity levels and types [556, 477], social context [389, 235], food availability [235, 244], and location [235] could affect eating behavior. As highlighted by these studies, mood is a component that could affect eating behavior to a larger extent.

Borrowing from such studies, by considering eating as a holistic event with interconnected dimensions, mobile sensing studies have used behaviors and contexts sensed passively to infer various eating behavior-related attributes. Compared to detecting eating episodes (that a user is eating) [323], additional characterizations help understand eating better, hence allowing more personalized and context-aware interventions and feedback [354, 325]. Biel et al. [55] used mobile and wearable sensing modalities to capture everyday eating episodes of a group of 122 Swiss university students and showed that it is possible to infer meal vs. snack eating episodes with an accuracy of around 84% with sensor features. Meegahapola et al. [330] showed that the self-perceived food consumption level could be inferred with accuracies over 83% using passive smartphone sensing features. Another study [328] emphasized the importance of identifying the social context of eating and showed that passive smartphone sensing could be used to infer lonely eating episodes for two student populations in Mexico and Switzerland with accuracies of the range 77% to 81%. All these studies leverage the idea that eating is a holistic event, not limited to the food type and amount, as captured in typical mobile food diaries. Moreover, these studies suggest that such sensor-based inferences are useful to reduce user burden in capturing mobile food diaries (because there is no need to capture them from users if they can be inferred) [567], providing interventions (because these inferences recognize moments that can be used to provide interventions) [354, 325], and on a larger scale, to understand population-level eating behaviors with passive mobile sensing (could be useful for social scientists to conduct large-scale studies with less self-reporting burden on participants) [323]. However, even though mood is a key component in better characterizing eating episodes, to the best of our knowledge, mood-while-eating has not been studied in depth in mobile sensing.

6.2.5 Smartphone Sensing Inference Personalization

Prior studies in smartphone sensing have used population-level models for inferences regarding health and well-being-related aspects [325]. Even though working with sparse and heterogeneous sensing modalities is a challenge, prior work has shown that inferences can be made with reasonable accuracies with such population-level models [468]. However, people are diverse in nature, and so are their behavioral patterns [152]. While capturing such diverse behaviors using population-level models has shown promise, prior work has highlighted that more complex tasks need personalization to achieve greater accuracies [502]. In addition, Kao et al. [240] and Jiang et al. [231] have shown that even though population-level models work well in experimental settings, they struggle to generalize well for larger

populations due to the heterogeneous nature of individuals. Hence, many commercial mobile health applications have looked into personalization by using user-level models [329, 285, 323]. Such models capture the behavioral diversity of people to a greater extent. However, since data from a single user alone is not enough to train a model for that user, hybrid models are used (refer Table 6.1 for the explanation of model types). Both user-level and hybrid models offer personalization to varying extents and have inherent advantages and drawbacks. In this context, there are two main ways of personalizing models [533, 568]:

1. Instance-Specific Models. These models would use data and features from the specific instance (a user) in building a model. Re-training or fine-tuning a population-level model by considering individual data (hybrid model) and building user-level models by training models on previously collected data of a user (user-level model) fall under this category. LiKamWa et al. [285] showed that building fully personalized models could increase binary mood inference accuracies compared to population-level models. To personalize, they trained user-level models for participants by training a model using participants' own data and testing on the rest of the data. Similar techniques have been used in ubicomp research in the past [330, 308, 106]. However, such fully personalized models require a single user to provide large amounts of data for model training, and prior work has shown that it is difficult to capture high volumes of data from participants during in-the-wild deployments [240, 231]. LiKamwa et al. showed that for binary mood inference, several days of data were required for user-level models to outperform population-level models. In addition to building user-level models, another study [330] used neural network model re-training for personalization. They showed that food consumption level inferences could be made with reasonable accuracies of the range 70%-80% using population-level models. However, they reported that personalizing by re-training population-level models using a part of the user's data would increase accuracies by around 1%-4%. In this case, since the population-level models already performed well, there is little value in increasing accuracies by small margins with hybrid modeling. It is unknown whether the same technique would work for a different or more complex task where population-level models do not provide reasonable accuracies.

2. Similarity-Based Models. The models under consideration employ similarity distances between instances to form a community of instances for model training, resulting in a hybrid model. Kao et al. [240] proposed a methodology that utilized collaborative filtering for imputation and clustering techniques to establish fixed user cohorts based on their similarity in sensor data and self-reported health information. For each new user, assignment to a pre-defined cluster is made based on similarity scores, and model training is performed using data from users in the cluster, resulting in improved accuracy for various health-related inferences. A similar technique was used to identify fixed groups of users for mental health severity inferences [383]. Additionally, Abdullah et al. [3] demonstrated that user-clusters can be identified, and models can be trained for clusters instead of individuals for superior performance. Suo et al. [498] employed latent distributions obtained from convolutional neural networks (CNNs) to find similarities between users, which were then used to cluster individuals into groups. They showed that their CNN-based approach outperformed other similarity metrics in clustering similar users for disease prediction. In most of the aforementioned studies, a single model is utilized for multiple users if they are assigned to the same cluster. In contrast, our approach does not utilize a set of pre-defined clusters, instead, it identifies a unique community of users for the target user in real-time using cosine similarity of dataset features, resulting in a more personalized outcome for each user.

In addition, previous techniques, such as those discussed by Abdullah et al. [3] assume a large amount of data is available for each user for personalization, which may not be feasible in tasks such as mood inference. To address this issue, Xu et al. [568] proposed a collaborative filtering-based approach to

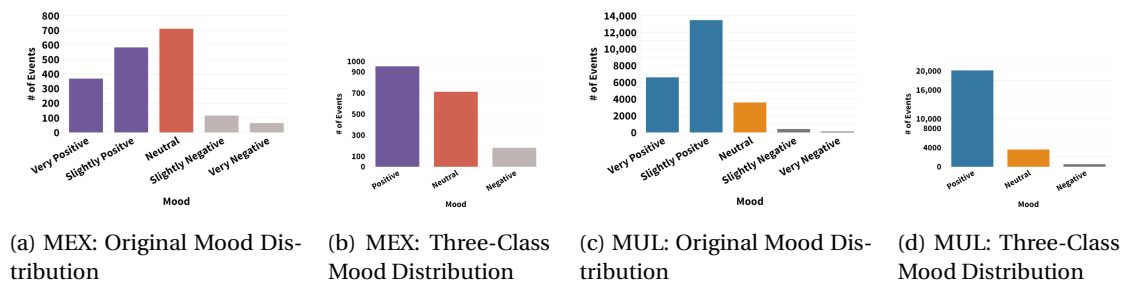


Figure 6.1: Original and Three-Class Mood Distributions of datasets

detect depression in college students, using majority voting to classify by making a prediction for each dataset feature based on its similarity with a target label. However, it is not clear whether this technique can be used to train more complex models or combined with other machine learning techniques, as the technique is not model-agnostic. Additionally, their technique requires longitudinal data from users and aims to infer a person-level attribute, whereas our approach aims to capture an event-level aspect that may change over time for the same user.

In summary, our technique is model-agnostic and does not assign a user to a pre-defined community. Instead, it finds a unique community for each target user, the size of which can be adjusted using a tunable threshold parameter. This accounts for both inter-personal and intra-personal variability that could affect model performance.

6.3 Dataset Description

Not a lot of datasets exist to study the intersection of mood and eating behavior with mobile sensing data. Hence, we used the two datasets from our previous work [330, 324], that were mentioned in Chapter 2.

6.3.1 Mexico Dataset (MEX)

This dataset has passive smartphone sensing and self-report data regarding the holistic context of eating episodes from 84 college students (56% women, mean age of 23.4) in Mexico. After removing self-reports with missing information and participants with poor sensor data coverage, the final dataset contains a total of 1843 eating episode reports that have matching and sensor data from 61 participants, with an average of around 30 eating reports per participant. It is also worth noting that sensor data and eating episodes in this study are non-overlapping because reported eating episodes are at least four hours apart. We used features described in Table 2.2 for the analysis.

Figure 6.1a shows a distribution of 1843 self-reports. The responses were distributed as, very positive (N=369), slightly positive (N=583), neutral (N=711), slightly negative (N=116) and very negative (N=64). After regrouping the responses into three classes, we got the distribution as depicted in Figure 6.1b. The number of responses considered as negative (N=180) is considerably fewer than the other two moods. As past studies regarding mood have shown [280, 139], these self-reports can be skewed towards positive responses. Hence, this skew was expected, specially for a group of young adults.

6.3.2 Multi-Country Dataset (MUL)

The objective of this dataset was to collect data from a diverse population of students, which helps to analyze human behavior and context connected with smartphone passive sensing features and self-reports. A total of 329K self-reports were collected from participants, where 24k self-reports were associated with eating events. We used features described in Table 2.4 for the analysis.

Figure 6.1c shows the distribution of 24207 self-reports. The distribution consists of very positive (N=6610), slightly positive (N=13475), neutral (N=3597), slightly negative (N=412), and very negative (N=113) mood responses. Figure 6.1d shows the three-class mood distribution after regrouping the five-class mood responses. The data distribution of the MUL dataset is highly skewed towards positive responses compared to the MEX dataset distribution.

6.4 Methods

In this section, we discuss the experimental setup for the inference of mood while eating. The inference was carried out using several model types such as Random Forest (RF), Naive Bayes (NB), Support Vector Classification (SVC), XGBoost (XGB), and AdaBoost (AB) [117, 432, 365, 87, 457, 373]. These models were chosen by considering the characteristics of the dataset, such as the size, tabular nature, and interpretability. We used the Synthetic Minority Over-sampling Technique (SMOTE) [86] for training sets and down-sampled testing sets to obtain balanced datasets for testing, hence having a baseline accuracy of 33.3% for all experiments. Hyperparameter tuning was done with GridSearch. We used scikit-learn [392] along with Python to carry out the experiments.

The three target classes of the experiment are: negative, neutral, and positive. We made inferences using many models, with different techniques for different degrees of personalization, to obtain a clear understanding of how the inference accuracy for mood inferences during eating episodes can be improved. The study was carried out using the following three approaches.

- **Population-Level Model (PLM):** We followed a population-level, leave-one-user-out strategy [468, 324]. This approach involved utilizing the data of each target user as the testing set, while the remaining dataset was utilized for training the model. The objective of this strategy was to simulate a scenario in which models were created using data collected from a population utilizing a mobile health (mHealth) application with mobile sensing, and to evaluate the performance of these models on a new target user. Therefore, for each selected user, the model was trained using data from the remaining participants and then tested using the data of the target user. To enhance the robustness of the training process, for each target user, a randomly sampled subset of data from the population (90%) was utilized for training, and a 30% random sample of the target user's data was used for testing. This process was repeated five times for each user, and the results were subsequently averaged across all users.
- **Hybrid Model (HM):** This case is somewhat similar to PLM, but with a percentage of data from the target user included in the training set, which leads to a partially personalized model for that target user [285, 324]. The technique illustrates a model where the user has already used the sensing app for some time, and the model has been already personalized to some degree after getting some ground truth data via self-reports from the target user. In the testing phase, new data gathered from the same user will be used in the testing set. When conducting experiments, for each target user, the model training was done using 70% of that user's randomly sampled data and a randomly sampled portion of data from other users (90%) combined, and the rest of the data of that target user (30%) was used to

test the model. The process was iterated five times for each user. Finally, the results across all users were averaged.

- **Community-Based Model (CBM):** community-Based model is based on the similarity-based technique proposed in Section 6.4.3, where for each target user, we extracted the community of that user empirically. To detect the community, different th values were used, and different communities were tested for each user to find the best community. The training split was created using a randomly sampled set of data from the community of that user (90%), together with 70% of the target user's data. Then, the model was tested using the remaining 30% of the data of the target user. The process was iterated five times for each user. Finally, the results across all users were averaged.

6.4.1 Generic Models and Context-specific Models Analysis (RQ1)

In order to assess the performance of a generic mood inference model within specific contexts, we utilized the Multi-Country dataset (MUL) which provided information on the concurrent activities being performed by the participants at the time of self-reporting their mood. This dataset included a list of 12 broad activities, such as resting, walking, sports, eating, drinking, studying, and working. By utilizing this dataset, we were able to examine the performance of the mood inference model in different situated contexts, as the participants were engaged in various activities. This allows for an examination of the model's performance across a diverse range of settings.

First, we trained a mood inference model using the population-level approach on the MUL dataset. Then, we obtained a breakdown of the mood inference performance for the testing test, for each activity performed by users during the mood report. By examining the mood inference performance across different activities, we aimed to determine whether generic, one-size-fits-all models for mood inference are similarly effective across all situated contexts and whether context-specific models are necessary for eating behavior, as previously studied in the workplace context. Experiments were carried out in both population-level and hybrid modeling approaches, hence examining the partial personalization effect. Experiments were done five times with random sampling, and the results were averaged. Second, considering the total dataset of the size 329K mood reports including the 24K mood-while-eating reports, we conducted an analysis in which we kept the number of training data points from all the activities constant (280K mood reports) and increased the number of mood-while-eating reports in training set from 0 to 15000 (in steps 0, 2500, 5000, 7500, 10000, 12500, 15000). The testing set contained around 20K mood reports captured during other contexts and 2000 mood-while-eating reports. For each step, experiments were done five times with random sampling for training and testing sets, and the results were averaged. Hence, by increasing the number of mood-while-eating reports in the training set and observing the performance of the overall testing set, and in addition, to the 2000 mood-while-eating reports, we examined how the inclusion of data from a specific context affects model performance. Finally, while we reported results for this setup in later sections (considering page limitations and brevity), we also experimented with other different setups, as in, changing the size of the other mood reports in training (e.g., 100K, 150K, 200K, 250K instead of 280K, which is the maximum number of data points we could use for training) and testing sets (e.g., 2.5K, 5K, 10K, 15K instead of 20K, which is closer to the maximum number of data points we could use for testing), which did not yield contrasting conclusions or results.

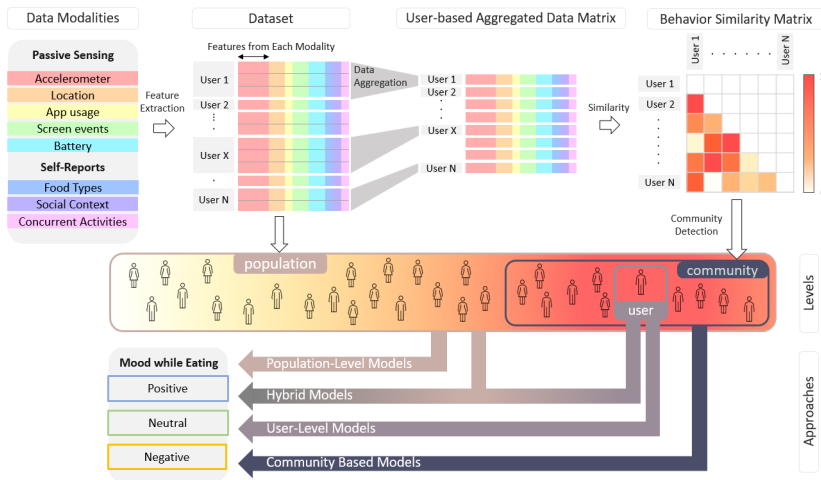


Figure 6.2: High-Level Architectural View of the Community-based Personalization Approach

6.4.2 Mood-While-Eating Inference (RQ2)

For this section, we used both datasets. In MEX, we used the whole dataset as all mood reports corresponded to eating events, hence providing the mood-while-eating. However, in MUL, since there are over 329K mood reports, out of which a majority corresponded to different activities, we only used mood reports corresponding to eating events to examine this research question. We first trained population-level models and hybrid models to examine the effectiveness of mood inference. We conducted experiments with both approaches with multiple model types (RF, NB, SVC, XGB, AB) across both datasets. Then, we also discuss the reasons for not being able to examine and train user-level models for mood-while-eating inference in limited data settings from each individual. Hence here, we identify the need for a better personalization strategy for limited data setting.

6.4.3 Community-Based Model Personalization Approach (RQ3)

When a large population of users is considered, there can be users who are different from each other as well as users with certain similarities. With the proposed technique, we aim to discover similar users for a target user and categorize them into a group (i.e., the community of the target user) in order to increase the number of data points in the training dataset when compared to training a user-level model. This is because gathering a large amount of data from a single user can be time-consuming, and it also allows higher accuracies as compared to population-level models [240, 231]. (The usage limitations of the user-level models because of the number of data points in the negative class and why this personalization technique is proposed for this specific inference task will be discussed in more detail in Section 6.5.2 and Section 6.5.3). A high-level architectural view of the approach we used to infer mood-while-eating is shown in Figure 6.2. The approach consists of a user-level data aggregation method and user similarity matrix calculation as described in Section 6.4.3 and a community detection mechanism described as described in Section 6.4.3. The purpose of the community-based personalization approach is to discover similar users for a given target user based on their behavioral and contextual data.

Obtaining the User Similarity Matrix

User-based Data Aggregation. First, to quantify the similarity between users, we use a mean-based data aggregation method similar to [342]. As given in Algorithm 1, for each user, we first filter out the data points of that user from the dataset (D_u) using the user id. Then we calculate the mean for each feature and generate a feature vector (u_{aggr}) for each user, by considering all the available data points. Hence, the feature vector of each user would have one value for all features in the dataset. Then, the feature vectors were included in the user-based aggregation matrix (U_{aggr}). This matrix provides an overview of all the users in the dataset and a summary of their features. In addition, this representation allows comparing users based on their feature vector.

User Similarity Matrix. By using U_{aggr} matrix, the next step is to calculate similarities between users. When we consider two vectors, the similarity between them can be calculated with the cosine similarity [281]. Consider two aggregated user vectors $u1_{aggr}$ and $u2_{aggr}$ of randomly selected user1 and user2, respectively. Then we can find the similarity between those two users with Formula 6.1 [282], where i is the i^{th} column index of the u_{aggr} (i.e. i^{th} feature of F).

$$Sim_{cosine} = \cos(\Theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i^2)} \times \sqrt{\sum_{i=1}^n (B_i^2)}} = \frac{\sum_{i=1}^{|F|} u1_{aggr(i)} \times u2_{aggr(i)}}{\sqrt{\sum_{i=1}^{|F|} (u1_{aggr(i)})^2} \times \sqrt{\sum_{i=1}^{|F|} (u2_{aggr(i)})^2}} \quad (6.1)$$

This similarity value between the two users suggests how close they are based on their smartphone sensing features and self-reports. We calculate this metric for all user pairs and represent it in a matrix of dimensions $|U|^2$. For each user pair, this value suggests how similar they are, where similarity increases when going from 0 to 1 (this is after normalizing the original -1 to +1 scale). Similar users to a target user can be selected using this metric.

Community Detection

In this section, we discuss the approach of threshold selection (Section 6.4.3), where we select different community sizes based on a tunable threshold value and community detection (Section 6.4.3).

Threshold Selection. We use a threshold (th) to filter users similar to the target user. Depending on the value of the threshold, the size of the community would differ. For example, if $th = 0.85$, we take all the users who have similarity values equal to or greater than 0.85 with the target user as the community. Hence having the threshold at zero keeps all the other users in the community (similar to a population-level model), and we can increase the threshold to reduce the number of users in the community. While this makes the community much smaller, it also makes it much more similar to the target user. On the other hand, increasing the threshold helps to remove some users from the training set, who could have been noisy otherwise. Increasing the th would affect the resulting dataset in two ways. First, it would reduce the number of users in the community, hence leading to comparatively smaller datasets for training models. Second, it is made sure that the selected data points are coming from similar users. Hence, depending on the chosen th , it could increase or decrease the size of the resulting dataset, hence affecting the accuracies of any model, trained on the dataset. In our experiment, we used a range of th values to obtain accuracies for all users.

Community Detection. Depending on th , for each user, the number of similar users ($|U_{u_t}|$) might vary.

As stated in the Algorithm 2, Sim_{u_t} for each user is a vector with the dimensions of $1 \times (|U| - 1)$, which contains the similarity matrix calculated for target user u_t , with other users in the dataset (i.e. $U \setminus \{u_t\}$). When iterating through similarity values for the target user, $Corr_u$ which is a similarity value of the target user with another user, is compared against th to decide whether the other user is included in the community of the target user. After the community is detected, training and testing of personalized models can be done using the community dataset.

6.5 Results

6.5.1 Generic and Context-specific Model Performance (RQ1)

First, Figure 6.3 shows that the generic mood inference model (best performing random forest classifier) trained with data from a wide range of everyday life occasions displayed contrasting performance across activities performed while the self-reports were provided. On the one hand, the results indicate that in both population-level and hybrid approaches, activities such as Resting, Walking, and Sports (which correspond to Mood-While-Resting/Walking/Engaging in Sports) showed higher performance than the overall accuracies (PLM: 43.2%, HM 50.1%) against a baseline of 33%. On the other hand, the model performed the worst across Eating, Studying, and Working (which correspond to Mood-While-Eating/Studying/Working). Interestingly, people tend to use phones while resting and walking, while it is not the case while eating, studying, or working, which are exigent activities. Sport is an exception, where even though it is an exigent activity for which there were not many mood reports, the model still performed decently well. Along this line, prior work has shown that mental well-being during work is an interesting aspect that is worth further investigation [353]. However, as we have discussed previously, such studies do not exist for mood-while-eating.

Next, Figure 6.3 shows the effect of having mood reports during eating events for the overall performance on the testing set and also the performance of the model only for eating reports in the testing set (Mood-While-Eating). Results indicate that when no mood-while-eating reports are present in the training set, mood-while-eating performance is 45.6% and overall performance is 53.3%. However, when increasing the number of mood-while-eating reports in the training set, performance for eating reports in the testing set increases to 47.8%. This could be because more representation of mood-while-eating reports in the mood inference model training helped the model generalize better to eating events in the testing set. Interestingly, when adding more mood-while-eating reports to the training set, the overall performance declined by around 3% to 50.1% before the curve plateaued. This means that even though the number of data points for training the model increased, it did not work well for the overall performance because all the data points came from a similar context (i.e., mood-while-eating). This, in turn, would make the tree-based model learn representations for such contexts better than other contexts, leading to declining performance for other contexts at the expense of increasing performance for mood-while-eating events in the testing set.

In summary, in response to **RQ1**, our findings indicate that generic models for mood inference do not exhibit consistent performance across various situated contexts. Similar scenarios where models do not perform for a sub-group of the distribution is known as sub-population shifts [575]. Hence, this result could be attributed to the fact that self-report data may not be equally captured across all daily situations and that the ease of learning mood may vary depending on the context, such as during eating, studying, or working. Furthermore, our analysis revealed that an increase in the representation of mood-while-eating reports in the training set for the mood inference model led to

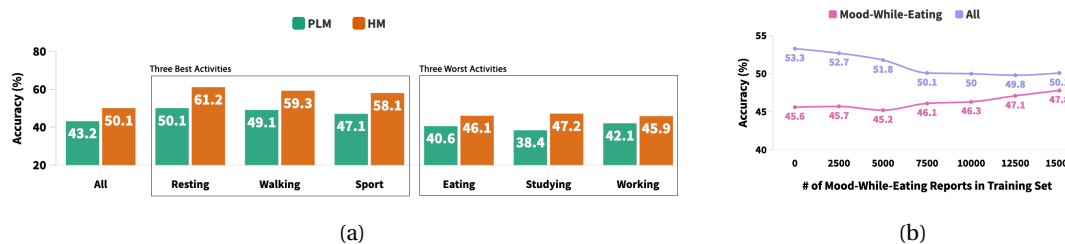


Figure 6.3: (a) This shows the context-specific accuracies of the generic mood inference model trained for MUL for approaches PLM and HM. (b) This shows the overall mood inference accuracy (All) and accuracy for eating events (Mood-While-Eating) when the number of mood-while-eating reports in the training set changes, with HMs for MUL.

Table 6.2: Mean (\bar{A}) and Standard Deviation (A_σ) of inference accuracies, calculated using five models for the PLM and HM of mood inference task: \bar{A} (A_σ), F1.

Feature Group	PLM				HM
	RF	SVC	XGB	AB	RF
(# of Features)	\bar{A} (A_σ)	\bar{A} (A_σ)	\bar{A} (A_σ)	\bar{A} (A_σ)	\bar{A} (A_σ)
Baseline	33.3 (0.0)	33.3 (0.0)	33.3 (0.0)	33.3 (0.0)	33.3 (0.0)
MEX (40)	36.9 (11.9)	20.4 (4.1)	30.4 (11.5)	31.2 (13.4)	64.0 (20.8)
MUL (114)	43.6 (15.6)	32.5 (14.8)	41.0 (15.5)	37.9 (17.6)	50.3 (20.6)

improved performance on mood-while-eating reports in the testing set. However, overall performance declined, highlighting the challenge of generalizing to different situated contexts. These results suggest the need for context-specific models or better domain generalization techniques for mood inference in order to improve the accuracy of continuous mood tracking in mobile food diaries and health applications.

6.5.2 Mood-While-Eating Inference (RQ2)

Population-Level and Hybrid Model Results

The results of the PLM and HM techniques are summarized in Table 6.2. PLM accuracies were not sufficiently high. The highest accuracy obtained for the MEX dataset was 36.9% using the RF, which is 3.6% higher than the baseline accuracy. And for the MUL dataset, the highest was 43.6% with the RF model type, which was 10.3% higher than baseline accuracy. Model types such as NB, and SVC did not perform well with both datasets and the accuracies were lower than the baseline. The authors of [411, 240, 568] have mentioned the averaging effect in PLMs. Hence, In summary, the PLM approach did not perform well for the three-class inference task regardless of the data or model type. In addition, the HM results in both datasets show an increment in performance with the hybrid model type. The best accuracy from the HM approach is 64.0%, which is an increment of nearly 27% than PLM.

User-Level Results

Even though we did not report results for user-level models (ULM), we conducted experiments with them for both datasets. ULMs are models that only use the target users' data in both training and testing with a 70:30 split (refer Table 6.1 for explanation). However, the number of users for whom experiments

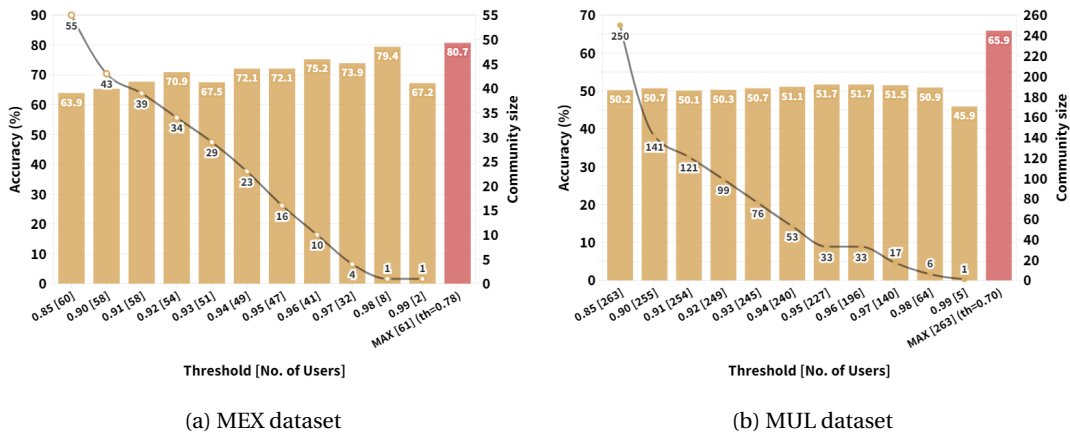


Figure 6.4: Mean accuracy values (column values in the graphs) calculated using the random forest for CBM for multiple threshold values [No. of Users] and averaged community sizes (line values in the graphs)

could be carried out was fairly low because there was a low number of users with enough data to be split into training and testing sets. For example, in the MEX dataset, if we consider the number of classes represented in the dataset for each user, there are four users with data from only one class (only positive or only neutral), 32 users with data from only two classes (only two classes out of positive, negative, and neutral), and 25 users with data from all three classes (all positive, negative, and neutral). Hence, a generic ULM can only be built for these 25 users. However, as shown in Figure 6.5, the number of users with a high number of negative class labels decreases significantly when the number of labels is increased. For example, there are 16 users with at least two negative class labels, and it goes down to 4 users with at-least eight negative class labels. However, when data is split into training and testing, there might not be enough labels. Moreover, in the MUL dataset, the same issue can be identified, where there was a lesser percentage of negative class labels compared to the MEX dataset, which makes it more difficult to build user-level models using MUL dataset. Due to these reasons, ULMs can not be trained for a majority of users under reasonable assumptions. For the purpose of understanding, with an arbitrary negative label count of 6, we could train ULMs with the MEX dataset for six users with 70:30 training and testing splits and achieve an accuracy of 73.5% with passive sensing modalities.

Overall, these results of RQ2 suggest that, even though model types such as population-level, hybrid, and user-level can be used to infer mood-while-eating, either they lack performance or the model cannot be used with a large number of users due to lack of data.

6.5.3 Mood-While-Eating Inference Using Community Based Personalization Approach (RQ3)

Community-Based Model Results with Changing Thresholds

Figure 6.4 summarized the results of the CBM approach with MEX and MUL datasets. Since RF models performed relatively well with the CBM approach compared to other classification model types and considering space limitations, we only discuss results obtained from RFs from here onwards. Both Figure 6.4a and Figure 6.4b show results of models when the threshold (*th*) value is increased from 0.85 to 0.99. Even though we obtained results for other threshold values ranging from 0 to 0.99, we only included threshold values that showed high variations in terms of accuracies, average community size,

and the number of people to which the inference can be applied [No. of Users]. The point to note is that in these cases, the same threshold has been used on all users, which is not ideal. For example, for the MEX dataset, accuracy increases to a maximum of 79.4% (at $th = 0.98$) and decreases afterward. This is mainly because as the threshold increases, the number of people in the community decreases, leading to a low number of samples for training models for a target user. This could decrease model performance. On the other hand, as th increases, the similarity of users increases, hence leading to more similar data samples for a target user. This could increase model performance. Hence, the threshold value performs a trade-off between these two aspects. Even though there is no noticeable accuracy increase in MUL dataset with the increase in threshold values as the MEX dataset, similar behavior can be identified. In addition, we mention the number of users for which models can be trained using the approach for different thresholds. While a $th = 0.85$ allows creating models for 60 users in the MEX dataset, $th = 0.99$ only allows models for two users. The reason for this again is that for many users, there is no community at such high thresholds.

In the rightmost column (MAX), we use different thresholds for different users, and the average threshold values for all the users, are given within brackets (ex: $th=0.70$). We empirically found the ideal threshold for each target user that yields the highest accuracy for the inference. Since the community size decreases when we increase the threshold, in order to build a model for a target user, there should be users with similarity values greater than the given threshold. MAX column summarizes the average results for such different thresholds. These are the best accuracies that we could expect when CBMs are deployed. For example, the accuracy value 80.7% obtained for the MEX dataset (rightmost column) represents an average of maximum accuracies obtained for all the users. The 0.78 threshold value represents the average of threshold values associated with those accuracy values for each user. For the MUL dataset, the maximum accuracy obtained was 65.9% with an average threshold of 0.70. Therefore, with CBMs, sensors performed reasonably well showing the potential of making mood inferences for eating occasions just using sensor data. Moreover, while such adaptive thresholds increased the overall accuracies, they also allowed the creation CBMs for all 61 users in the MEX dataset and 263 users in the MUL dataset.

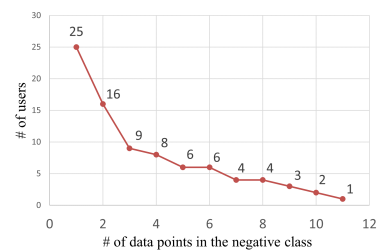


Figure 6.5: MEX dataset: Distribution change in no. of users with the increase of no. of data points in the negative class.

Comparison of Approaches

The results of population-level, hybrid, and community-based approaches obtained with RFs for both MEX and MUL datasets are shown in Figure 6.6. Compared to PLM, HM shows higher accuracies, supporting the idea that personalization gives better results than classic population-level leave-one-out methods.

In real-world mHealth applications, it is difficult to capture large amounts of data from users [240, 231]. This is the main drawback of user-level models (i.e. the model is trained and tested by splitting the target user's data into two splits), even though they provide reasonably high accuracies. This MEX dataset was collected from the users for a reasonably long time period (close to 30 days). However, the number of collected data points was not sufficient to build a user-level model for many users because of the lack of negative class labels. The same goes with MUL dataset where the number of negative class

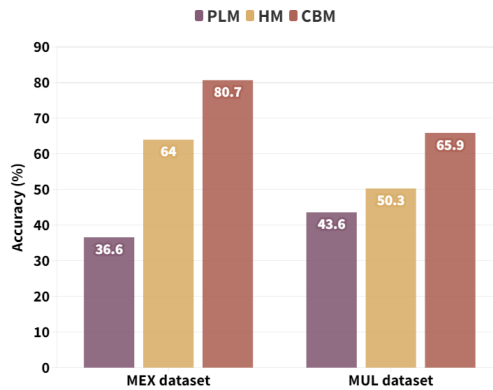


Figure 6.6: Mean accuracies calculated using the random forest for the PLM and HM of mood detection task for all the users and Maximum mean accuracies for CBM calculated for all the users.

labels in the dataset is not sufficient to build a 3-class model for most of the users. Additionally, there is a possibility of over-fitting user-level models to the target user when training the model using only the user’s data because the model does not even account for intra-user variability, let alone inter-user variability. Hence, if the user’s behavior changes with time, there is a chance of not being able to predict the mood correctly with the previously trained user-level models.

According to Figure 6.6, the CBM shows better performance compared to the other two approaches for both datasets. Specifically, compared to HM, CBM accuracy was increased by over 15%. In contrast to the user-level models we mentioned earlier, where each user needs to have a reasonable amount of data points in order to build a user-level model, this similarity-based approach can be used to build models for a high number of users, because the selected community of each user contains a comparatively higher amount of data which can be used to build the model. Furthermore, unlike user-level models, the machine learning model that had been built for each user shows robustness against behavioral fluctuations because the training split contains data from different users, which helps to avoid over-fitting. Hence, the CBM approach has the ability to capture both the inter-personal and intra-personal diversity of a target user. Moreover, with CBM approach the community of the selected user would be adjusted to the user’s behavior change with the context around them. For the same set of users in MEX dataset, PLM and HM performed worse than CBM, again showing the benefit of selecting a similar community to train a model.

As shown in Figure 6.4, we can observe that the average community sizes decrease with the increase of threshold value. And it can be observed that each user has different threshold values and community sizes. To get an overall idea of the community size variability with multiple threshold values, a heat map generated using the MEX dataset is given in Figure 6.7a. It shows how the community size is reduced when the threshold is increased. Figure 6.7b shows the average of accuracies obtained for all the users. It clearly depicts that the accuracy slightly increases with the threshold. This concludes that for any user, when the threshold is increased, the community size of that user decreases and the accuracy might increase. However, after high thresholds, the accuracy drops off again because of lack of members in user communities. A cumulative density function (CDF) figure for maximum accuracies is included in Figure 6.7c. Please note that considering the space limitation we have only included the distributions of MEX dataset.

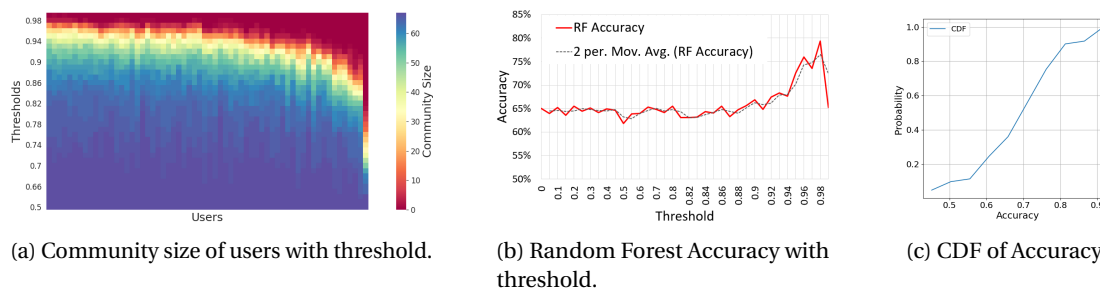


Figure 6.7: Distributions of CBM using MEX dataset: (a) How the community size of each user changes with threshold th ; (b) How the mean accuracy of the Random Forest model changes with threshold th ; (c) Cumulative Distribution Function (CDF) of Accuracy.

In summary, these results of **RQ3** suggests that personalization is a key component in achieving better performance in mood-while-eating inference task. The proposed community-based technique helps to overcome the challenges such as the lack of individual data and the cold-start problem.

6.6 Discussion

6.6.1 Summary of Results

In this chapter, we did a mood-while-eating inference using two datasets with passive smartphone sensing and self-report data. Our objective was first to check whether the generic mood predictions work well when predicting context-specific moods, such as mood-while-eating. Secondly, to carry out a three-class mood inference task, and finally, on ways to improve the performance of the mood inference model by proposing a personalization approach. The summary of the main findings is as follows:

RQ1: The results of our study indicate that the use of generic mood inference models may not be sufficient for accurate mood tracking across diverse contexts. An analysis of the performance of the model on mood-while-eating reports, specifically, revealed that increasing the representation of such reports in the training set led to improved performance on the corresponding subset of the testing set. However, this improvement was accompanied by a decline in overall performance, suggesting that the model struggled to generalize its predictions to other situated contexts. These findings suggest the necessity for the development of context-specific models for mood-while-eating inference in order to enhance the accuracy of mobile food diaries and mobile health applications.

RQ2: The performance of mood-while-eating inference models was evaluated using both population-level and hybrid modeling approaches. The population-level approach resulted in an accuracy of 36.9% for the MEX dataset and 43.6% for the MUL dataset when utilizing passive sensing data. In contrast, the hybrid modeling approach demonstrated a notable improvement in accuracy, with an increase of 7.4% for the MEX dataset and 6.7% for the MUL dataset. These results suggest that the hybrid modeling approach may be a more effective method for mood-while-eating inference. In addition, we also described how it becomes difficult to train fully personalized user-level models across both datasets, due to the lack of data from some classes.

RQ3: We proposed a community-based personalization technique as a solution to the challenges faced by traditional mood inference models, such as a lack of individual data from certain classes and the cold-

start problem. Through the implementation of this technique, we observed notable improvements in performance in the mood inference task, with accuracies of 80.7% and 65.9% achieved for the MEX and MUL datasets, respectively. These results demonstrate the potential of community-based personalization as an effective approach for addressing the limitations of conventional models in mood inference.

6.6.2 Implications

Implications for Modeling. In this chapter, we presented an investigation of personalization as a technique to improve the performance of models in a three-class mood inference task that uses only smartphone sensing data to infer mood-while-eating. Our proposed personalization technique addresses the challenges of limited individual data in a dataset and the cold-start problem. Our results indicate that the use of personalization techniques is crucial for the improvement of performance in mood inference tasks. However, we acknowledge that the collection of individual data is a challenging task. Thus, future studies should focus on capturing more data from individual users to investigate further various personalization techniques that can improve the performance of mood inference models. Additionally, our study highlights the importance of context-specific models for mood inference. In this chapter, we focused on the specific context of eating events, and future research should explore other context-specific mood inference tasks. Furthermore, the application of domain adaptation techniques for multimodal sensor data has yet to be studied in depth [569, 324]. While domain adaptation has been extensively researched in the fields of computer vision, NLP, and speech, its application to multimodal sensor data is an area that warrants further investigation.

Implications for Applications. The results presented in this chapter provide valuable insights into the potential of utilizing passive sensing data to infer context-specific moods, specifically in the context of eating events. This has important implications for the design and development of mobile health applications and mobile food diaries, which could leverage this information to provide tailored interventions and feedback to users. For example, by identifying negative moods associated with eating events, mobile health applications could send notifications or suggestions to interrupt users who have a tendency to overeat in such states, thereby preventing unhealthy eating patterns. Additionally, the use of smartphone-based sensing data in this chapter highlights the potential for mobile application developers to create cost-effective solutions that do not require expensive wearable devices. As such, further research should be conducted to investigate other context-specific mood inference tasks, and to explore the potential of domain adaptation techniques in multimodal sensor data analysis. We also open up another question for the research community: while capturing sensor data throughout the whole day and inferring generic mood could be useful, with resource-limited devices, whether doing this throughout the day is feasible is questionable. However, adding to the previous body of literature that attempted to infer mental well-being-related outcomes in specific contexts, what we could understand is, tracking mood in certain contexts (i.e., mood-while-eating, mood-while-at-work, mood-while-driving) that have specific applications and practical use cases built around it, could be more energy efficient for phones and adds value to user's everyday life.

6.6.3 Limitations and Future Work

The field of mood inference using passive sensing has been an active area of research for over a decade [325]. Our study, however, highlights the need for further experimentation in order to ensure that inferences work effectively in a diverse range of real-life contexts. In this particular case study, we

focused on the context of eating occasions, yet there are numerous other contexts, such as drinking occasions and social gatherings, that could be explored. Additionally, cultural diversity represents another important aspect to consider. The datasets used in this study were collected in Mexico and eight other countries, and it is well-known that individual, and food consumption behaviors can vary significantly across cultures [254, 125, 529]. Despite the second dataset being collected in eight countries, the limited number of labels in the negative class prevented us from building models for each country separately. Thus, it is imperative to investigate whether the inference and modeling techniques proposed in this chapter would generalize well to other countries, cultures, age groups, and contexts, such as eating, drinking, commuting, and being at home or at school. Conducting such experiments would greatly enhance the applicability of these models to real-world scenarios.

This chapter presents several limitations that could be addressed in future research. One such limitation is the use of cosine similarity as the similarity metric. While it is appropriate for the datasets used in this chapter, other similarity measures such as Mahalanobis distance [270] may be more suitable depending on the nature of the dataset. Additionally, the time window used for sensor data aggregation around a self-report, which is one hour for the MEX dataset and 10 minutes for the MUL dataset, may not accurately reflect the users' mood during the entire time period. It is important to note that this time window is used for the purpose of aggregating sensor data to infer the mood at the time of eating. Furthermore, while the chapter's inference pipeline is based on a well-established framework for eating behavior [57], the mood reports captured are based on a single answer regarding valence, similar to prior studies in the field of ubicomp [285]. This approach may have less validity compared to more elaborate instruments used in clinical psychology to capture affective states (e.g., PANAS). While the single-question approach was chosen to minimize user burden, future research could explore the use of additional instruments to study mood-while-eating at the episode level. Finally, it is worth noting that this chapter focuses on inferring valence, but not arousal [445]. As a future direction, studies could be conducted to infer both arousal and valence at the time of eating. Moreover, future studies can focus on carrying out more specific mood inference tasks, for example, five class mood inference tasks and seven class mood inference tasks using large datasets.

6.7 Conclusion

In this chapter, we aimed to investigate the relationship between mood and eating with mobile sensing data. We acknowledged the prior literature on mobile sensing that explored mood inference for generic moments and the need for more context-specific models. Additionally, we emphasized the significance of automatically inferring the mood-while-eating, as it could provide valuable feedback and interventions for mobile food diary users. Through examination of two datasets pertaining to the everyday eating behavior and mood of college students, we revealed that passive mobile sensing does not effectively infer the mood-while-eating using population-level modeling techniques. Furthermore, we demonstrated that user-level modeling techniques are limited in their applicability due to the scarcity of data. To address these limitations, we proposed a community-based personalization approach that allows for the training of partially personalized models even for target users with limited data. With this approach, we achieved an accuracy of 80.7% with the Mexico dataset and 65.9% with the multi-country dataset using passive sensing data. These results are promising, as they demonstrate the effectiveness of our technique in attempting a three-class inference task. We also highlighted the importance of considering the application of mood inference in specific domains, such as mobile food diaries, where a comprehensive understanding of the user's mood is crucial to gaining a holistic understanding of eating behavior.

7 Generalization and Personalization of Mobile Sensing-Based Mood Inference Models

Mood inference with mobile sensing data has been studied in ubicomp literature over the last decade. This inference enables context-aware and personalized user experiences in general mobile apps and valuable feedback and interventions in mobile health apps. However, even though model generalization issues have been highlighted in many studies, the focus has always been on improving the accuracies of models using different sensing modalities and machine learning techniques, with datasets collected in homogeneous populations. In contrast, less attention has been given to studying the performance of mood inference models to assess whether models generalize to new countries. In this chapter, we used the MUL dataset from Chapter 2, with 329K self-reports from 678 participants in eight countries (China, Denmark, India, Italy, Mexico, Mongolia, Paraguay, UK) to assess the effect of geographical diversity on mood inference models. We defined and evaluated country-specific (trained and tested within a country), continent-specific (trained and tested within a continent), country-agnostic (tested on a country not seen on training data), and multi-country (trained and tested with multiple countries) approaches trained on sensor data for two mood inference tasks with population-level (non-personalized) and hybrid (partially personalized) models. We show that partially personalized country-specific models perform the best yielding area under the receiver operating characteristic curve (AUC) scores of the range 0.78-0.98 for two-class (negative vs. positive valence) and 0.76-0.94 for three-class (negative vs. neutral vs. positive valence) inference. Further, with the country-agnostic approach, we show that models do not perform well compared to country-specific settings, even when models are partially personalized. We also show that continent-specific models outperform multi-country models in the case of Europe. Overall, we uncover generalization issues of mood inference models to new countries and how the geographical similarity of countries might impact mood inference. The material of this chapter was originally published in [324].

7.1 Introduction

Mental well-being-related issues are common among young adults due to a plethora of personal and societal reasons such as leaving home, study workload, poor financial stability, and complex social relationships [390, 431]. These issues are even more prominent in the post-pandemic world, where social relationships have taken a toll due to more emphasis on remote work/study settings. Some studies have shown that this emerging lifestyle has affected phone usage behavior as well [491, 594, 427, 446, 283]. Further, declining mental well-being conditions could lead to adverse outcomes such as substance abuse and suicidal thoughts [441, 113, 155]. In this context, prior research has discussed

the potential of timely and accurate mood tracking for both personal and clinical care [485, 543, 317, 149]. Ecological momentary assessments (EMAs) and survey questionnaires are commonly used for mood tracking. However, such techniques are burdensome to users, and prior work has shown that it is difficult to sustain the practice of reporting for long periods unless there is a strong motivation [40, 424, 460]. As a possible alternative, multi-modal sensors in smartphones could be used to infer mood unobtrusively with reasonable accuracies [411, 285, 468].

According to prior work in psychology and social sciences, physiological aspects, including mood, are perceived and expressed differently in different countries, cultures, and societies [303]¹. According to a cross-country study by Becht et al. [42], mood and related behaviors could vary based on a person's culture, and perceptions and beliefs regarding different moods stemming from one's culture. However, prior work in mobile sensing does not study the effect of the geographical diversity of users (e.g., country of residence) on smartphone sensing-based mood inference models.

Issues of generalization and fairness with regard to the geographical diversity of data sources have been discussed extensively in domains such as computer vision, speech, and natural language processing [193, 599, 548, 82, 310]. For example, gender classification models trained with data predominantly from the USA have performed poorly on people of African and Asian descent [82]. Many geographical-related biases (e.g., Indian brides being recognized as dancers, etc.) have been shown in models trained with the imagenet dataset, in which a majority of data is from western countries [599]. Such findings have uncovered issues in data collection practices and helped shape research directions to address issues related to diversity and biases. In this context, many prior mobile sensing studies that attempt inferences regarding well-being related aspects highlighted that models are trained in specific countries, and the generalization of techniques for other countries or regions should be explored further [90, 358, 327, 330]. However, mood inference studies have focused on only one or two countries for data collection [285] or have not considered the diversity of data sources in terms of the country, even when data were collected from multiple countries [468].

Bardram et al. [36] emphasized the need for generalization and reproducibility of sensing-based models for mental well-being-related outcomes. However, even though examining gender, age, and occupation-related diversity is feasible even within the same country, examining geographical diversity requires a considerable effort in conducting the same study, with the same protocol, in several geographic regions because studies are time-consuming and expensive; and logistical difficulties in conducting experiments such as language barriers, technology barriers, differences in motivating use cases and required incentives. Hence, studies that examine the geographical diversity of mobile sensing-based inferences are rare [402, 247]. In this chapter, we study and compare the performance of country-specific, country-agnostic, and multi-country approaches for mood inference. In addition, we also examine the effects of model personalization and generalization to new geographically diverse countries. To our knowledge, this is one of the first studies to examine the effect of the geographical diversity of users on smartphone sensing-based mood inference models, hence shedding light on distributional shift-related issues. Considering these aspects, we ask three research questions.

RQ1: What behavioral and contextual characteristics around mood reports of college students (from eight countries spanning Europe, Asia, and Latin America) can be extracted from the analysis of smartphone sensing and self-report data?

¹For pragmatic reasons, we are equating the geographical location (country) of our participants with a specific culture that is distinct to this particular country. We acknowledge that cultures can be multidimensional and exist in tension with each other and in plurality within the same geographic boundary [584]. However, throughout the chapter, we use country, culture, and geographic region interchangeably.

RQ2: How do smartphone sensing-based mood inference models perform in different countries (country-specific)? Can a model trained in one/more countries be deployed in another country not seen on training data to achieve reasonable accuracies, hence generalizing well (country-agnostic)?

RQ3: How do country-specific or continent-specific models perform as compared to a multi-country model?

By addressing the above research questions, this chapter provides the following contributions:

Contribution 1: As described in Chapter 2, we conducted a new smartphone-based data collection campaign among 678 participants in eight countries (China, Denmark, India, Italy, Mexico, Mongolia, Paraguay, UK) representing Europe, Asia, and Latin America to study their everyday mood and behavior. During the study, we collected 329,974 fully complete self-reports. In addition, we also collected rich passive sensing data with continuous sensing (activity type, step count, location, cellular, wifi, bluetooth, proximity, etc.) and interaction sensing (app usage, touch events, user presence, screen-on/off episodes, notifications, etc.) throughout the deployment. First, we found that negative mood reports in all countries would increase from morning to night. Moreover, with statistical analysis, we found that the features that help infer mood are different across countries. However, the best features included both continuous and interaction sensing modalities in all countries.

Contribution 2: We found that the country-specific approach performs reasonably for both two-class and three-class mood inferences with AUC scores in the range of 0.76-0.98 with hybrid (i.e., partially personalized) models. However, we noticed that across both two-class and three-class inferences, models do not generalize well to other countries, where AUC scores drop to the range of 0.46-0.55 on average in the population-level (i.e., non-personalized) setting and 0.66-0.73 in the hybrid setting. These findings raise the significance of discussing issues of generalization of mobile sensing-based models to different world regions.

Contribution 3: In the hybrid setting, we found that multi-country models do not perform as well as country-specific models even though they achieved an AUC of 0.81. However, they performed better than continent-specific models built for Asia and worse than the one built for Europe. Even though the performance differences were not high, this again highlights that building a model within European countries leads to higher performance and better generalization for those countries than using multi-country or even some country-specific models. A possible explanation is that the European countries under study (Italy, Denmark, UK) might share some daily behavioral patterns. In contrast, the three countries in Asia under study (China, India, Mongolia) have less similarity regarding daily patterns. Hence, these findings point toward the benefit of considering the geographical/cultural diversity of data collection on smartphone sensing-based mood inference models.

7.2 Background and Related Work

7.2.1 Definitions and Terminology

What is Mood?

There is no single way to define mood [134]. However, in prior work in mobile sensing, some operationalizations have been commonly used. Positive Negative Affect Schedule (PANAS) is a widely used validated questionnaire that can be used to capture the positive and negative affect of individuals

Table 7.1: Terminology and description regarding different model types and approaches.

Terminology	Description
Country-Specific	This approach uses training and testing data from the same country. Each country has its own model, without leveraging data from other countries. As the name indicates, these models are specific to each country (e.g., a model trained in Italy and tested in Italy). Both population-level and hybrid model types can be trained in the country-specific approach.
Continent-Specific	This approach uses training and testing data from the same continent. Each continent has its own model, without leveraging data from other continents. As the name indicates, these models are specific to each continent (e.g., a model trained in Europe and tested in Europe). Continent specific approach can be trained with population-level and hybrid models.
Country-Agnostic	This approach assumes that data and models are agnostic to the country. Hence, a trained model can be deployed to any geographical region regardless of the country of training. Country-agnostic approach too can be trained with population-level and hybrid models. There are two types of country-agnostic settings: (1) Country-Agnostic I: The first setting uses training data from one country, and testing data from another country. This corresponds to the scenario where a model trained in a country already exists, and we need to understand how it would generalize to a new country (e.g. a model trained in Italy and tested in Mongolia). (2) Country-Agnostic II: The second setting uses training data from four countries, and testing data from the remaining country. This corresponds to a scenario where the model was already trained with data from several countries, and we need to understand how it would generalize to a new country (e.g. a model trained with data from Italy, Denmark, UK, and Paraguay, and tested in Mongolia).
Multi-Country	This one-size-fits-all approach uses training data from all eight countries and tests the learned model in all countries. This corresponds to the setting in which multi-country data is aggregated to build a single model. However, this is also how models are typically built without considering aspects such as geographical diversity. Multi-Country models too can be trained with population-level and hybrid approaches.

[238]. In addition, the Patient Health Questionnaire (PHQ-9) has been used in the past to quantify depressive mood with mobile sensing [536]. However, these questionnaires are long and could be cumbersome to users [285]. Further, they can capture mood over the past week (or two), and might not be suitable to measure the in-situ mood for long time periods. Hence, prior work has also used an affect grid based on the circumplex mood model [468, 285] that would capture the *valence* and *arousal*. As described in later sections, due to pragmatic reasons, the data collection in this study does not focus on arousal because positive and negative affects of the circumplex model are important in determining negative moods that could be useful for adverse mental well-being related outcome detection, feedback, and interventions [40, 460]. Hence, only *valence* has been captured in a five-point scale: very positive (😊), positive (🙂), neutral (😐), negative (😞), very negative (😡). This five-point scale is similar to LiKamwa et al. [285] and Horlings et al. [216]. For inference, we reduce the five-point scale to two-point and three-point scales similar to prior work [540, 134, 71]. This is usually done based on the idea that in mood inference, the more important aspect is to detect extreme moods (i.e., negative, positive) rather than to identify all fine-grained intermediate mood levels in the middle of the spectrum [216]. First, obtaining a three-point scale using the five-point scale was obvious by combining very positive and positive to positive; neutral as it is; and negative and very negative to negative, hence having three classes [540, 482]. However, for two-class inference, the categorization is not as obvious. Some prior studies have removed the class in the middle (i.e., neutral), hence obtaining positive and negative labels [591, 216]. Even though it is possible to do it with the available classes in the dataset, we believe it would lead to a biased classifier that would not perform reasonably well when exposed to data corresponding to neutral mood labels. Hence, we followed prior work that binned very positive, positive, and neutral moods as positive; and negative and very negative moods as negative [591, 71]. This two-class inference also allows for detecting negative moods, which is useful in mobile health apps for feedback and interventions [40, 460] because it is such negative moods, along with other aspects

like stress that could be harmful to individuals on the long term. Hence, in the scope of this chapter, mood can be defined as *the instantaneous valence reported by study participants on a five-point scale (from very positive (😊) to very negative (😞)), reduced to either a two-point scale corresponding to positive and negative classes or a three-point scale corresponding to positive, neutral, and negative classes, for inference using smartphone sensing data.*

Model Types and Approaches

This section introduces the definitions and terminology used in this chapter, as summarized in Table 7.1. In terms of model types, we use population-level (subject-independent) and hybrid models [152, 323, 285]. While population-level models are not personalized, hybrid models are partially personalized. The operationalization of models is described in Section 7.4. Second, in terms of approaches, we consider the country-specific approach that is trained and tested within each country; the continent-specific approach that is trained and tested within each continent; the country-agnostic approach in which models are trained in one or more countries, and tested in an unseen country; and the multi-country approach that would ignore the diversity in terms of countries, and train a one-size-fits-all model considering data from all countries. As an important note, all these approaches can be evaluated with both population-level and hybrid model types. For example, in a country-specific setting, imagine a model trained with a certain population in Italy and tested with some new users in Italy, hence examining the model performance on new users from the same country. This is equivalent to a population-level model of the country-specific approach. Then, imagine the set of unseen users producing data for model training after using a mobile app for some time, and these data points being used to update the model. This would then lead to a hybrid model of the country-specific approach. Similarly, for the country-agnostic approach, a model trained in Italy deployed to unseen users in Paraguay is similar to evaluating a population-level model. Then, imagine the users in Paraguay providing some data for model personalization. This leads to a hybrid model created with a mix of data from Italy and Paraguay that can be evaluated on new data points from users in Paraguay, whose data were used in model training. While this model too can be called a multi-country model, for ease of understanding in the scope of this chapter, we would still call it a hybrid model with the country-agnostic approach. Using the combination of model types and approaches, we can examine the effect of personalization (with model types) and model generalization to new countries (with the four approaches), hence uncovering distributional shift-related issues of multi-modal mobile sensing datasets for mood inference.

7.2.2 Considerations for Research in Mobile Sensing Involving Geographic Diversity

Mood and Geographical Diversity

Across different geographical regions and cultures, behavior is mediated by inherent beliefs, presses, and affordances of physical and/or socio-cultural environments [402]. Even for behaviors that are similar across cultures, the psychological meaning of those behaviors might not be the same due to [402]: (a) Certain behaviors that are acceptable in certain countries/cultures are not perceived as normative or appropriate in other countries [532]; (b) The same behavior might be indicative of different outcomes/functions. For example, while cycling is everyday behavior in certain regions (e.g., Aalborg, Denmark), it might only be used for exercise in other areas (e.g., Ulanbaator, Mongolia); and (c) Different behaviors might be indicative of a similar outcome/function. For example, while

people in some countries might perform cycling for exercise, people in other countries might prefer going to the gym for exercise. Why people cycle will depend on many contextual and cultural factors such as road safety, availability of public transport, alternative exercise options, weather conditions, and perceptions about cycling in a specific geographical region. Given that smartphone sensors can capture such physical activities (e.g., Google Activity Recognition API [183] and other activity engines built by researchers [544]) and are used to infer more complex variables [77, 544], invariably, such behavioral differences across geographical areas could affect mood inference models that leverage activity data from accelerometers and location [402]. In addition, device-mediated behavior or phone usage behavior could also vary between geographical areas depending on cultural norms, weather conditions (e.g., the phone usage behavior while walking outside in a cold vs. a hot country), network coverage, and subscription plans (e.g., people in countries where internet plans are expensive might turn off internet frequently, people in countries where the used phones are old might turn off Wifi and location sensors often to save battery of the phone, etc.), and availability of alternative equipment that could serve similar functionality (e.g., using a laptop for zoom calls instead of the phone, hence showing differences in the sensed app usage behavior). Given that mood inference models in prior work have used both continuous (activity types, step counts, location, proximity, wifi, etc.) and interaction (typing and touch events, user presence, application usage, screen on and off events, etc.) sensing modalities to examine/infer mood and related psychological constructs, how behaviors and contexts captured with smartphones affect mood inference in different countries is worth investigating.

Studies about Psychological Constructs and Geographical Diversity

According to Khwaja et al. [247], psychological mobile sensing research aims to quantify and measure constructs related to mood, stress, depression, and user personality over the last decade due to the advancement of sensing technologies. Even though there is a myriad of studies about such psychological aspects, ranging from clinical to non-clinical studies, many have focused on a population within a single country [402]. In addition, even when the construct of analysis used in studies is the same (e.g., circumplex mood model, positive-negative affect schedule, etc.), comparing different studies across countries is complicated because data have been collected using different protocols and sensing modalities [7]. Furthermore, Phan et al. [402] have discussed how prior psychology studies in mobile sensing have collected data focusing on WEIRD samples (Western, Educated, Industrialized, Rich, and Democratic) and paid less attention to the global south. This has also been highlighted in a review study on smartphone sensing by Meegahapola et al. [328]. For these reasons, prior work has emphasized the need for studies that examine the generalization of models across countries/cultures by building diversity-aware approaches to machine learning-based modeling of sensor data [325, 36]. According to a recent review by Phan et al. [402], only Khwaja et al. [247] have considered the cultural diversity of smartphone sensing-based models on psychological aspects, where they studied personality traits based on Big-Five model. In that study, the authors collected data from 166 participants from five countries (UK, Spain, Colombia, Peru, and Chile). They showed that country-specific models perform the best, regardless of the gender or age balance, for the prediction of Extraversion, Agreeableness, and Conscientiousness. Compared to that study, we also collected data from multiple countries. However, our primary focus is on studying mood inference models that could vary from time to time, even within the same person (more dynamic), instead of stable personality traits. In addition, Muller et al. [358] used mobile GPS data to predict depression in socio-demographically homogeneous sub-samples within the USA. They trained algorithms for the whole sample and homogeneous sub-samples (e.g., highly educated men, women residing in rural regions, etc.) and tested within and across sub-samples. They found that the technique that led to high AUC scores for student populations (0.82), did not

generalize well to the USA-wide population-level model (AUC of 0.57). In contrast, our work focuses on valence instead of depressive mood. In addition, rather than concentrating on socio-demographic differences within a particular country, we focus on cross-country differences.

7.2.3 Mood and Smartphone Technologies

Mood Tracking with Self-Reports

In the early days, mobile phone-based mood charts were used to track the mood of individuals. These are based on self-reported questionnaires and ecological momentary assessment (EMA) responses [83, 325]. Similar to how mobile food diaries were designed for people who wanted to control their diet [330], mood charts were designed to support people who wanted to control negative moods and increase self-awareness, allowing for monitoring and feedback [40, 460]. With randomized controlled trials, some studies explored the usefulness and efficacy of self-report-based mood tracking and showed that engaging in mood tracking tools increases self-awareness, hence reducing the possibility of having anxiety, even within clinically depressed populations [32, 56]. Going beyond applications related to health and well-being, Glasgow et al. discussed how aspects like destinations, travel choices, and social ambiance are related to mood [176]. Further, in this context, prior work that uses mood tracking has focused on different populations such as college students [278, 544], adolescents [245] and clinically diagnosed, high-risk populations with mental well-being related issues [543, 149, 317]. Hence, most prior studies relied on user engagement to keep track of mood. This could be a burden to users in the long run, and it is known that apps that require many self-reports do not have high adoption rates. In our work, even though we captured self-reports about mood, they were captured as ground-truth labels to train classifiers with sensor data for mood inference. Such inferences could be used to update mood-tracking applications that could be used to provide context-aware interventions, and feedback to users, with less user burden [468].

Mood Tracking with Sensing

Mobile phone sensors allowed researchers to build context-aware systems that could infer various aspects regarding the health and well-being of people [268]. Most of such studies rely on using features captured from sensors in smartphones as proxies to personal attributes (mood, stress, etc.), behavior (eating, drinking, running, walking, etc.), and context (social context, semantic location, ambiance, etc.) [325]. Hence, there are studies that infer aspects like mood [468, 285], stress [300, 453], depression [77, 150], eating behavior [330, 55], drinking behavior [454], activity types [352], and social contexts [327, 328], among many others. If we specifically focus on mood-related studies, LiKamwa et al. [285] showed that the mood of individuals captured with the circumplex mood model could be inferred with an accuracy of 66% with all user models (population-level), which can be increased up to 93% using personalization (user-level) with a dataset collected from 32 individuals. They suggested that building hybrid models (partially personalized) would help overcome the drawbacks of both population-level and user-level models. Servia-Rodríguez et al. [468] collected a large-scale dataset of mood self-reports and passive sensing data from multiple countries. They also showed that binary mood captured with the circumplex mood model could be inferred with an accuracy of 70% with population-level models. Some studies examined mood instability derived using mood reports, with phone sensor data [356, 593]. In our work, we look into inferring mood valence with population-level and hybrid models. However, we are more interested in examining (a) the similarities and differences in mood models for different countries; and (b) the generalization of models to unseen countries, both of which have not been

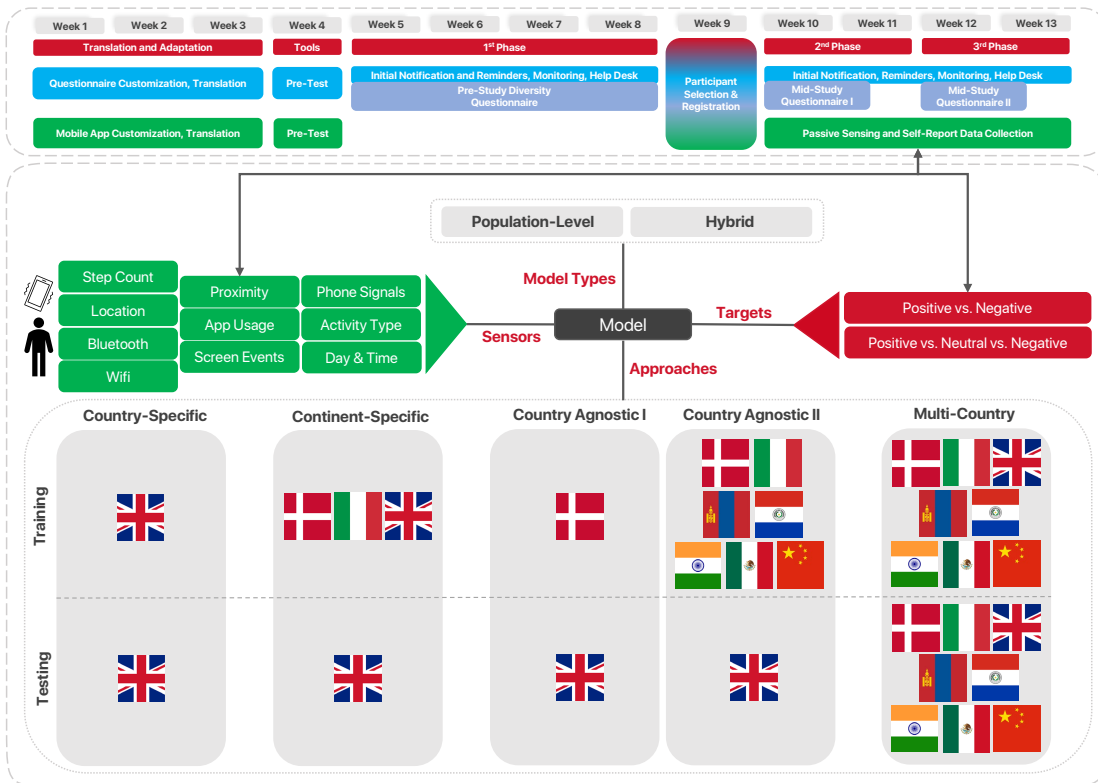


Figure 7.1: High-level overview of the study.

examined in prior work. Further, as Bardram et al. [36] highlighted, there is a lack of reproducibility and generalization of machine learning models across studies in this domain. We believe the results presented in this study would be a step in the right direction for better awareness of these issues in examining the characteristics and generalization of smartphone sensing-based mood inference models across different geographical regions.

7.3 Behavioral and Contextual Characteristics Around Mood Reports Extracted from Sensor Data and Self-Reports (RQ1)

In this chapter, we used the 8-country dataset (MUL) described in Section 2. The study overview is shown in Figure 7.1.

7.3.1 Descriptive Analysis

Figure 7.2 shows the distribution of mood labels for the eight countries. We observed fewer labels for the ‘negative’ and ‘very negative’ classes compared to the ‘neutral’, ‘positive’, and ‘very positive’ classes. As shown in Figure 7.2a, except for China, where there were more ‘very positive’ reports than ‘positive’ or ‘neutral’ reports, all other countries had ‘positive’ as the majority label. This behavior of skewed reporting is common in studies about valence [468, 285]. Furthermore, we plot the hourly distribution of mood reports in Figure 7.3. According to Figure 7.3a, across all countries, we could see

7.3 Behavioral and Contextual Characteristics Around Mood Reports Extracted from Sensor Data and Self-Reports (RQ1)

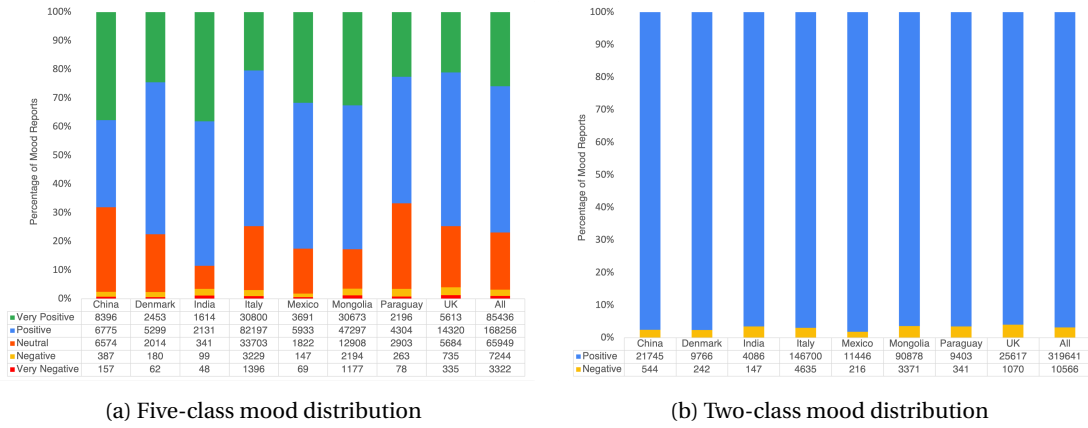


Figure 7.2: Summary of self-reported mood distributions.

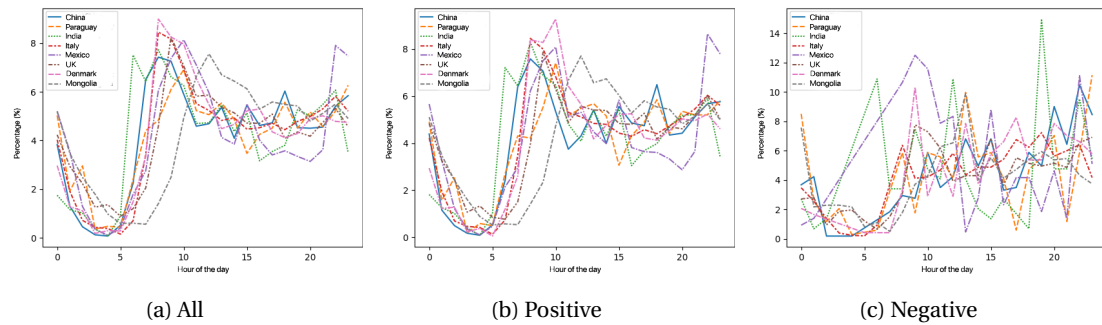


Figure 7.3: Distribution of self-reported moods for 24 hours of the day.

more self-reports in the morning compared to the afternoon or evening. However, Figure 7.3b shows that most self-reports around morning are for the positive valence. This means that users had a more positive mood after waking up and around the morning. Interestingly, we also observed that the curve for Mongolia indicates late sleep and late wake-up, according to reports, which the partner institution later confirmed to be consistent with the routines of students in the country. As shown in Figure 7.3c, we also noticed that negative valence reports increase with time in most countries. This is in line with prior studies about mood and stress levels increasing with the time of the day [341].

As mentioned in Section 2.2.2, participants' social context and semantic location labels were captured with time diaries, in addition to mood. So, in the sub-figures of Figure 7.4, we show the distributions of social context (alone or not) and location context (home or away) for positive and negative moods. These two aspects were chosen because prior work has shown that being alone and being away from home could affect mental well-being and behavior [390, 431, 472, 327]. In the figure, on the X-axis, the eight countries are shown. On the Y-axis, the percentage of self-reports is shown. Regarding location, except in China, in all other countries, most mood reports were captured when participants were home. Please note that the data was collected in the Fall of 2020, during the covid pandemic—so participants spent a significant amount of time at home. The more interesting aspect is the difference in the percentages for Positive and Negative moods: that is when comparing Figure 7.4a and Figure 7.4b. The highest difference was in Mongolia, where 67% of negative moods were reported at home out of all negative reports. In contrast, 90% of positive moods were reported when at home, out of all positive reports. This means that in Mongolia, participants reported a higher proportion of negative reports

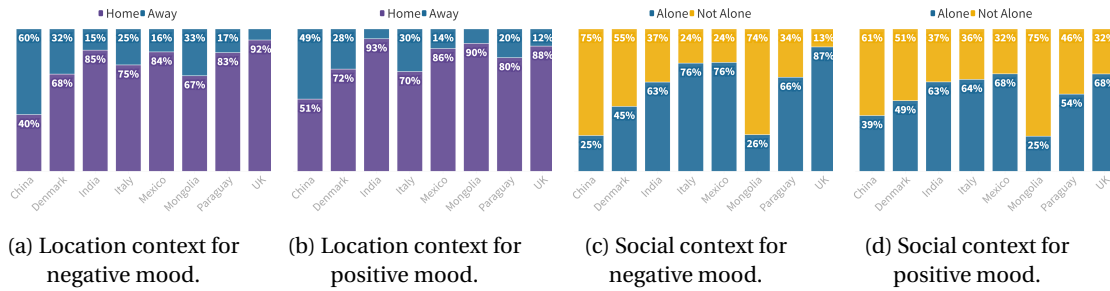


Figure 7.4: Location and social context distributions for negative and positive mood.

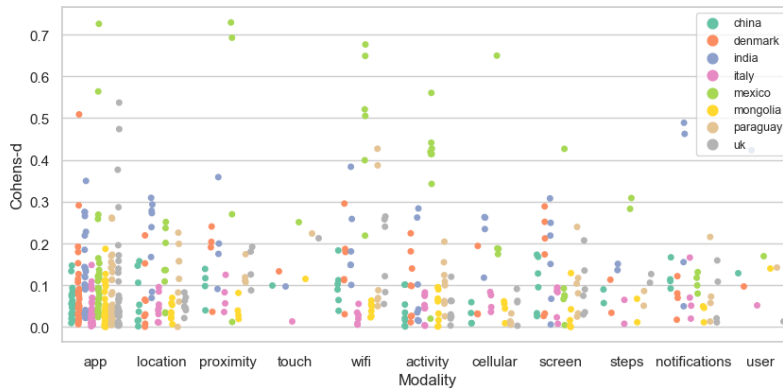


Figure 7.5: Cohen's-d (effect Size) distribution of features for negative and positive classes, grouped by countries and modalities.

when away from home. This is a difference of 23%. The difference is the lowest in Mexico. For social context, the highest difference was found in the UK, where 87% of negative reports were done when alone. In contrast, only 68% of positive reports were done when alone, indicating that in the UK, people tend to report more negatively when alone. The trend is similar in all other countries except China and Denmark, where proportionally more people reported that they are alone when having positive moods.

7.3.2 Statistical Analysis

In this section, we seek to understand features with high statistical significance in discriminating either positive, neutral, or negative classes from the other two. Therefore, in Table 7.2, we show the t-statistic [249] and p-values [190] (p-values higher than 0.05 after Bonferroni correction for multiple hypothesis testing [555] are marked with *). In addition, since p-values are limited in determining statistical significance [276], we also report Cohen's-d [429] (all features have 95% confidence interval not crossing zero [267]) for positive, neutral, and negative classes for each country. The rule of thumb to evaluate Cohen's-d is 0.2 = small effect size, 0.5 = medium effect size, and 0.8 = large effect size. For the positive mood, across all the European countries and Mongolia, proximity sensor-related features were among the top five features, indicating that phone usage and/or location of the phone could reveal positive moods. However, except for Denmark, where there was a small effect size, the proximity feature had less than small effect sizes in all other countries. In addition, in Denmark, cycling activity was indicative of positive moods. Interestingly, Copenhagen in Denmark is a city widely known for

7.3 Behavioral and Contextual Characteristics Around Mood Reports Extracted from Sensor Data and Self-Reports (RQ1)

Table 7.2: t-statistic (TS) (p-value > 0.05 : *) and Cohen's-d (CD) (all features reported here had 95% confidence intervals not overlapping with zero) for positive, neutral, and negative moods for each country.

	Positive			Neutral			Negative		
		TS	CD		TS	CD		TS	CD
China	location altitude min	10.43	0.16	proximity std	11.51	0.17	app health & fitness	5.03	0.08
	wifi connected	7.98	0.11	location speed mean	7.81	0.11	app music & audio	4.52	0.13
	app communication	7.95	0.11	app health & fitness	7.61	0.09	location speed min	3.71	0.15
	screen # of episodes	5.56	0.08	app tools	7.31	0.10	location speed mean	3.34	0.15
	app lifestyle	5.15	0.08	app personalization	6.87	0.10	proximity mean	2.81	0.12
Denmark	app not found	24.17	0.62	touch # of events	28.91	0.56	app puzzle	8.76	0.19
	activity onbicycle	12.88	0.35	app video players/editors	28.91	0.11	app music & audio	7.32	0.29
	wifi # of devices	9.93	0.24	app weather	9.73	0.16	screen std episode	4.07	0.25
	proximity mean	9.89	0.27	app personalization	9.35	0.23	app lifestyle	3.84	0.11
	wifi std rssi	9.72	0.24	activity still	8.83	0.22	app social	3.21	0.18
India	noti posted w/o duplicates	6.65	0.34	wifi min rssi	8.12	0.44	app business	4.17	0.22
	wifi # of devices	6.59	0.34	wifi mean rssi	6.31	0.38	activity tilting	4.02	0.28
	noti removed w/o duplicates	5.99	0.28	location radius of gyration	5.08	0.27	app tools	3.47	0.27
	app strategy	5.72	0.36	app books and reference	5.02	0.11	wifi min rssi	3.17*	0.26
	screen # of episodes	9.10	0.27	screen min episode	4.65	0.24	app communication	3.12*	0.27
Italy	proximity max	26.20	0.16	wifi # num of devices	12.96	0.07	app news & magazine	11.47	0.11
	proximity std	15.19	0.09	app video players/editors	10.45	0.06	app action	8.55	0.09
	location speed min	13.21	0.08	activity still	6.80	0.04	activity still	4.98	0.07
	proximity mean	12.61	0.08	app adventure	6.52	0.03	app video players/editors	4.62	0.07
	wifi min rssi	11.75	0.07	app lifestyle	6.16	0.03	app social	4.17	0.06
Mexico	wifi max rssi	24.57	0.68	proximity std	41.39	0.98	cellular lte min	10.29	0.65
	wifi mean rssi	23.99	0.69	proximity max	0.93	0.06	wifi # of devices	9.74	0.65
	wifi std rssi	22.28	0.63	app communication	20.98	0.49	proximity max	9.34	0.73
	screen # of episodes	13.74	0.35	cellular lte std	18.23	0.32	app tools	8.79	0.73
	location altitude max	12.39	0.34	app music & audio	18.03	0.36	activity still	7.92	0.56
Mongolia	app not found	13.76	0.12	wifi # of devices	16.46	0.14	app personalization	10.99	0.19
	wifi std rssi	13.04	0.12	location altitude max	16.25	0.15	app music & audio	10.09	0.11
	proximity	7.25	0.06	app role playing	15.24	0.09	app educational	9.79	0.05
	wifi connected	6.66	0.05	wifi min rssi	12.26	0.11	app sports	8.10	0.08
	user presence time	6.07	0.05	app tools	11.98	0.11	app communication	6.67	0.11
Paraguay	wifi min rssi	24.32	0.53	touch # of events	24.29	0.49	app role playing	6.99	0.15
	wifi mean rssi	20.07	0.43	location speed min	22.49	0.48	app productivity	5.71	0.17
	noti posted w/o duplicates	13.60	0.31	app tools	12.30	0.27	app tools	5.02	0.26
	activity running	8.08	0.19	wifi # of devices	11.53	0.25	screen std episode	4.71	0.23
	user presence time	7.69	0.17	app strategy	8.72	0.16	touch # of events	4.49	0.22
UK	proximity std	10.52	0.18	wifi mean rssi	17.93	0.24	app role playing	39.48	0.53
	wifi # of devices	10.51	0.15	wifi min rssi	15.67	0.21	app board	21.19	0.29
	app business	9.83	0.16	wifi max rssi	11.89	0.17	app personalization	17.32	0.47
	app tools	9.82	0.14	app role playing	10.34	0.13	touch # of events	8.27	0.21
	proximity max	9.53	0.16	cellular lte mean	9.08	0.13	screen # of episodes	6.78	0.20

cycling [412], which might explain this finding. Further, running activity could discriminate positive mood with a small effect size in Paraguay. Prior work has also shown that high physical activity could lead to positive moods and less stress [54, 239].

Regarding the negative class, app features were predominant in most countries. For some apps, high usage indicated negative moods (e.g., puzzle in Denmark, news & magazine in Italy, etc.). In contrast, for some apps, low usage indicated negative moods (e.g., health & fitness in China, music & audio in China and Mongolia, role-playing games in UK and Paraguay, etc.). In addition, for both UK and Paraguay, a high number of touch events on the phone was indicative of negative moods. This finding is generally in line with prior studies that examined fine-grained smartphone usage and mental well-being [223, 285]. In summary, features from modalities such as app usage, screen and phone usage events (episodes, touch events, user presence, proximity, etc.), WiFi, activity types, and location were among the ones that helped discriminate between different moods. Further, except for the ‘proximity std’ feature in Mexico for neutral mood, none of the features had a larger effect size. For a few country-mood pairs, there were cases of features having above medium effect sizes (e.g., number of touch events in Denmark for Neutral, many features from modalities such as cellular, WiFi, proximity, etc. in Mexico, minimum RSSI value for WiFi in Paraguay for Positive, and role-playing apps in the UK for Negative). Figure 7.5 shows the distribution of Cohen’s-d values for all features grouped by sensing modalities for the two classes studied in this chapter (i.e., negative vs. positive). Results indicate that depending on the country, the expressiveness of different sensing modalities in discriminating negative classes from other classes is different. For example, for ‘app’ features, effect sizes are small for countries such as China, Italy, and Mongolia. In contrast, more informative features with larger effect sizes are present for Denmark, Mexico, and the UK.

7.4 Mood Inference (RQ2 & RQ3)

7.4.1 Experimental Setup

The primary goal of this chapter is to investigate aspects related to mood inference, personalization, and generalization to different countries using smartphone sensing data. As described and defined in Figure 7.1 and Table 7.1, we use two model types: population-level and hybrid, to examine personalization to individuals, and four modeling approaches: country-specific, continent-specific, country-agnostic, and multi-country, to examine generalization and country-wise performance. Hence, this section will describe the operationalization of the experimental protocol.

We used python with scikit-learn [391] and Keras [91] frameworks to conduct all experiments. Initially, we conducted country-specific experiments with different model types such as random forest (RF), gradient boosting, support vector classification, XGBoost, AdaBoost, and multi-layer perceptron neural networks [117, 432, 365, 87, 457, 373]. We obtained the best results for a larger majority of inferences with RFs. In addition, these models allow interpreting results better because they provide Gini feature importance values for trained models. Because of these reasons and space limitations, we will only report results for RF models with default parameters in this chapter². Further, to fill in missing values of the dataset, we used k-nearest-neighbor (kNN) imputation [50, 596]. In addition, we report all the results with the area under the receiver operating characteristic curve (AUC) [69] because they

²Note that we also tried out GridSearch for parameters in the random forest (for `n_estimators`: 50, 100-2000 with intervals of 100, `max_depth`: 2-16 with intervals of 2, `min_samples_split`: 2-10) that did not yield better performance than the default parameters (`n_estimators`: 100, `max_depth`: NA, `min_samples_split`: 2), except in a few cases. Hence, we used default parameters for all experiments for consistency.

provide a better assessment of performance when dealing with imbalanced data (when used with macro averaging which gives equal emphasis to all classes in an inference). While we provided a basic description of model types in Table 7.1, the operationalization of models is given below.

- **Population-Level Models (PLM):** Since this represents a scenario where models are deployed to a set of users unseen in model training, we use the leave- n -participants-out strategy when testing models. This is an extension of leave-one-out cross-validation, where we consider n users in testing instead of one. Hence, if the number of users in the considered population is N , we pick n such that it is roughly 20% of N (can be obtained with group- k -fold cross-validation with $k = 5$ in scikit-learn). So, for each n user in the testing split, 50% of their data would be used for testing to be coherent with hybrid models (stratified based on users and mood labels), and data from the rest of the $N - n$ users would be used for the training split. Then, experiments were repeated ten times by randomly sampling n users, and the results were averaged.
- **Hybrid Models (HM):** Since this represents a scenario where models are deployed to a set of users already seen in model training (hence partially personalized models), we first use the leave- n -participants-out strategy similar to PLM. So, for each n user in the testing split, data from the rest of the $N - n$ users would be used for the training split. In addition, 50% of the data from the testing split (stratified based on users and mood labels) would be included in the training set to represent partial personalization. In addition, an equal number of data points to the number of data points added to the training set from the testing set would be removed randomly to make the number of data points in the training and testing sets for HM and PLM equal making them more comparable. Finally, experiments were repeated ten times by randomly sampling n users, and the results were averaged.

Using the above two model types, we conducted the experiments using four approaches. The country-specific approach examines how models trained within a country perform. We examine both PLM and HM types for this approach, hence examining the personalization within countries. The country-agnostic approach examines how models trained in one or a few countries generalize to a new country. With PLM and HM model types, we examine how personalization affects model performance when models are deployed to countries unseen on training data. The multi-country approach is similar to a one-size-fits-all model trained with data from all available countries. This is similar to a model in which country diversity is ignored. Both PLM and HM model types were used to examine the effects of personalization on model performance.

7.4.2 Results

Country-Specific Models

In Table 7.3, we show country-specific results with PLM and HM. In addition, we also show the aggregate results from country-specific (as ‘Aggregate’) and multi-country models. Under ‘Multi-Country (Balanced)’, we use an equal number of data points from each country (equal to the country with the minimum number of data points, which is India) by randomly sampling when training and testing models. The results show that PLMs do not perform well for two and three-class inferences. Models in Mexico performed better than in other countries. These results are reasonable because many features in Mexico had medium to large effect sizes, as shown in Figure 7.5. However, HM results show that they perform better than PLMs, showing the usefulness of personalization within each

Table 7.3: Country-Specific and Multi-Country results with PLM and HM: Mean (\bar{S}) and Standard Deviation (S_σ) AUC scores computed from ten iterations. Results are presented as $\bar{S}(S_\sigma)$, where S is AUC.

	PLM		HM	
	Two-Class	Three-Class	Two-Class	Three-Class
Baseline	.50 (.00)	.50 (.00)	.50 (.00)	.50 (.00)
China	.51 (.04)	.45 (.04)	.78 (.02)	.79 (.01)
Denmark	.41 (.10)	.56 (.03)	.83 (.03)	.86 (.01)
India	.46 (.15)	.45 (.04)	.79 (.03)	.76 (.02)
Italy	.55 (.05)	.52 (.01)	.82 (.01)	.81 (.00)
Mexico	.62 (.21)	.62 (.13)	.98 (.01)	.94 (.01)
Mongolia	.49 (.08)	.49 (.02)	.85 (.01)	.83 (.00)
Paraguay	.48 (.08)	.53 (.01)	.84 (.01)	.84 (.01)
UK	.56 (.05)	.52 (.05)	.91 (.01)	.87 (.00)
Aggregate	.51 (.10)	.52 (.04)	.85 (.02)	.84 (.01)
Multi-Country	.52 (.03)	.53 (.02)	.83 (.01)	.79 (.00)
Multi-Country (Balanced)	.53 (.02)	.52 (.03)	.81 (.03)	.78 (.02)

country. With HMs, the performance for two-class inference almost doubled for Denmark, and even for other countries, the AUC bump was above 30%. These results suggest that for both two-class and three-class inferences, partial personalization within each country leads to significant improvements in performance. When the aggregate results of country-specific models are compared with multi-country models, PLMs do not show a significant difference. However, with HMs, it is clear that country-specific models outperform multi-country models by 2% for two-class and 5% for three-class. This suggests that model personalization within countries leads to better performance when compared to the personalization of one-size-fits-all models. This is reasonable given that we are reducing the distributional shift by only considering data within a country and adding an effect of personalization by being geographically diversity-aware. In addition, the ‘Multi-Country’ approach performed slightly better than the ‘Multi-Country (Balanced)’ case. This could be because, in the imbalanced case, models favor countries with more data points, such as Italy and Mongolia, leading to a slight increase in performance for those countries that occupy a majority of the dataset. Furthermore, regardless of whether it is a two/three-class inference, the performance of models did not degrade much.

Country-Agnostic I Models

Next, we examine the country-agnostic approach. Table 7.4 and Table 7.5 show the results for two-class and three-class inferences, respectively. In both tables, we first show results for models trained in specific countries when tested on an unseen country in the form of a matrix with an empty diagonal. Then, under ‘Aggregate’, we show the aggregate value of those results for each training country (e.g., PLM performance for models trained in China when deployed to other countries). In addition, we calculated AUC scores for the same set of models with partial personalization (all the results are not shown here due to space limitations), and similar to the aggregate of PM, we show the aggregate values under HM. Results show that PLMs do not generalize well to new countries with AUCs of 0.47 - 0.52. However, these results are on par with PLM accuracies in country-specific and multi-country approaches. This suggests that regardless of the country from where sensing data were obtained to train models for mood inference, PLMs performed similarly. However, HM results convey an opposite conclusion for two and three-class inferences. For the two-class inference, the country-specific approach had AUC scores in the

Table 7.4: Country-Agnostic I PLM & HM: Two-Class Inference – Mean (\bar{S}) and Standard Deviation (S_σ) of AUC scores obtained by testing each Country-Specific model (rows) on a new country. Results are presented as $\bar{S}(S_\sigma)$, where S is AUC score. Aggregate of the reported population-level results and results from hybrid models indicated under ‘Aggregate’.

Training	Testing (PLM)								Aggregate	
	China	Denmark	India	Italy	Mexico	Mongolia	Paraguay	UK	PLM	HM
China		.53 (.02)	.44 (.03)	.49 (.01)	.58 (.05)	.50 (.01)	.42 (.03)	.51 (.02)	.55 (.02)	.67 (.04)
Denmark	.51 (.00)		.47 (.01)	.51 (.00)	.58 (.02)	.50 (.00)	.58 (.01)	.46 (.00)	.52 (.01)	.69 (.03)
India	.48 (.00)	.37 (.00)		.50 (.00)	.40 (.02)	.50 (.00)	.44 (.01)	.52 (.00)	.46 (.00)	.70 (.02)
Italy	.49 (.00)	.45 (.00)	.51 (.01)		.40 (.02)	.51 (.01)	.48 (.00)	.50 (.00)	.48 (.01)	.69 (.02)
Mexico	.49 (.00)	.58 (.01)	.44 (.01)	.49 (.00)		.49 (.01)	.56 (.01)	.47 (.01)	.50 (.01)	.73 (.03)
Mongolia	.49 (.00)	.48 (.01)	.52 (.00)	.50 (.00)	.51 (.00)		.48 (.00)	.51 (.00)	.50 (.00)	.71 (.03)
Paraguay	.51 (.00)	.53 (.01)	.49 (.01)	.50 (.00)	.55 (.02)	.53 (.01)		.50 (.01)	.52 (.01)	.70 (.02)
UK	.48 (.01)	.43 (.02)	.57 (.00)	.50 (.01)	.32 (.01)	.50 (.01)	.49 (.01)		.47 (.01)	.66 (.02)

Table 7.5: Country-Agnostic I PLM & HM: Three-Class Inference – Mean (\bar{S}) and Standard Deviation (S_σ) of AUC scores obtained by testing each Country-Specific model (rows) on a new country. Results are presented as $\bar{S}(S_\sigma)$, where S is AUC score. Aggregate of the reported population-level results and results from hybrid models indicated under ‘Aggregate’.

Training	Testing (PLM)								Aggregate	
	China	Denmark	India	Italy	Mexico	Mongolia	Paraguay	UK	PLM	HM
China		.48 (.01)	.54 (.01)	.48 (.01)	.47 (.01)	.50 (.01)	.51 (.01)	.50 (.00)	.50 (.01)	.68 (.02)
Denmark	.52 (.01)		.41 (.02)	.56 (.01)	.54 (.04)	.51 (.01)	.50 (.02)	.58 (.01)	.52 (.02)	.66 (.04)
India	.52 (.01)	.42 (.02)		.52 (.01)	.38 (.02)	.52 (.01)	.52 (.01)	.38 (.01)	.47 (.01)	.68 (.03)
Italy	.52 (.01)	.49 (.01)	.47 (.02)		.32 (.02)	.51 (.01)	.51 (.00)	.54 (.00)	.48 (.01)	.69 (.02)
Mexico	.49 (.00)	.59 (.00)	.44 (.00)	.47 (.00)		.50 (.00)	.61 (.00)	.54 (.00)	.52 (.00)	.71 (.02)
Mongolia	.49 (.00)	.50 (.00)	.43 (.00)	.51 (.00)	.55 (.00)		.54 (.00)	.53 (.00)	.51 (.00)	.67 (.02)
Paraguay	.44 (.01)	.51 (.02)	.48 (.03)	.52 (.01)	.58 (.05)	.53 (.01)		.55 (.01)	.52 (.02)	.65 (.04)
UK	.53 (.01)	.51 (.01)	.51 (.03)	.53 (.01)	.40 (.06)	.52 (.01)	.53 (.02)		.50 (.02)	.67 (.03)

range of 0.78-0.98, whereas the country-agnostic approach yielded scores in the range of 0.66-0.73. A similar pattern can be seen for three-class inference, where scores dropped from 0.76-0.94 to 0.65-0.71. This shows that the effect of personalization achieved with HMs is strong for the country-specific approach, whereas country-agnostic models still did not generalize well. However, we also noticed that with HMs for both two-class and three-class inferences, models trained in European countries consistently performed better in other European countries than the rest. For example, in the two-class inference, the Italian model had AUC scores of 0.76 and 0.78 in Denmark and the UK, respectively. In contrast, the next best score for the Italian model was 0.70 in India. Finally, for three-class inference, the UK model had AUC scores of 0.73 and 0.75 for Italy and Denmark, respectively, whereas the next best score was 0.69 for Paraguay. These results could be partly justified given that European countries have somewhat closer everyday patterns that could get captured in the models.

Country-Agnostic II Models

In Table 7.6, we show results for country-agnostic models that were trained in seven countries and tested in the shown country. Compared to the previous setting, where the models were trained in only one country and tested in another, these models capture a more considerable intra-subject variability in model training. Moreover, HM results were not included here because, technically, it is similar to the HM of multi-country models. PLM results show that the performance is not high for both two-class and three-class inferences. For some countries, performance slightly increased compared to country-specific (e.g., China, Paraguay in two-class). For some, the performance declined (e.g., India,

Table 7.6: Country-Agnostic II PLM: Mean (\bar{S}) and Standard Deviation (S_σ) of AUC scores obtained by testing each a seven-country model on data from a new country. Results are presented as $\bar{S}(S_\sigma)$, where S is the AUC.

	Two-Class	Three-Class
Baseline	.50 (.00)	.50 (.00)
China	.54 (.01)	.48 (.01)
Denmark	.51 (.02)	.48 (.01)
India	.53 (.03)	.47 (.01)
Italy	.54 (.01)	.50 (.01)
Mexico	.41 (.02)	.54 (.01)
Mongolia	.49 (.01)	.49 (.01)
Paraguay	.56 (.01)	.55 (.01)
UK	.48 (.01)	.51 (.01)

Denmark, Italy, Mexico, and the UK in two-class). Hence, there is no clear evidence that having more data from multiple countries would help to generalize better for an unseen country, even in this case.

Multi-Country and Continent-Specific Models

Finally, in Table 7.7, we show the results for the multi-country approach and also the continent-specific approach that is similar to the country-specific; however, instead of countries, we considered two continents: Europe (Italy, Denmark, UK) and Asia (China, Mongolia, India)³. The primary motivation for examining these models is the result we obtained in the country-agnostic approach, where for HM, models trained in European countries performed better in other European countries with HMs. Results for the continent-specific approach show that models performed similarly to any other approach for both two-class and three-class inferences for PLM. However, the Europe model for two-class inference had an AUC score of 0.58, which is second only to the Mexican model (0.62) in the country-specific approach.

Furthermore, results show that the continent-specific model for Europe with an AUC of 0.89 for two-class inference, performed better than the multi-country (0.83) and even country-specific approach for Italy (0.82) and Denmark (0.83) and closer to the country-specific UK model with an AUC of 0.91. Similar results can be seen for three-class HM inference. This suggests that for western Europe, where everyday patterns might be somewhat similar across countries, continent-specific models could perform reasonably. However, for the continent-specific Asian model, it is not the same. For example, for the two-class inference, the Asia model had an AUC score of 0.79, which is similar to country-specific China (0.78) and India (0.79) results but significantly lower than the result for Mongolia (0.85). On the other hand, for the three-class HM, the Asia approach reached an AUC of 0.74, whereas China, India, and Mongolia models reached 0.79, 0.76, and 0.84, respectively. Hence, continent-specific models did not perform as well as country-specific or multi-country models for Asia. This could be because even though China, India, and Mongolia are geographically on the same continent, the behaviors and cultures of students are different. In addition, ‘balanced’ models decreased performance for Europe and Multi-Country, whereas for Asia, it is not the same, where three-class HM performance increased in the balanced case. Again, this is because India and China get more representation in training, leading to better performance in testing.

³There are arguments for and against on whether North and South America are a single continent or two [366, 565, 550]. In the Anglo-Saxon world, it is often stated that there are seven continents, with North and South America being separate. In contrast, it is taught otherwise in Latin America [550]. Hence, we did not include ‘America’ results by combining Mexico and Paraguay.

7.4 Mood Inference (RQ2 & RQ3)

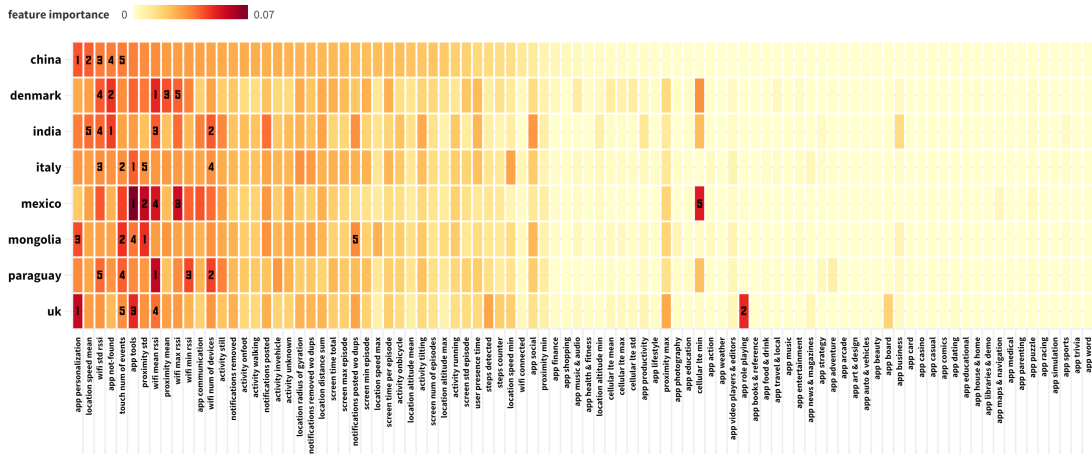


Figure 7.6: Country-Specific HM: Gini feature importance values from RF models for two-class inference.

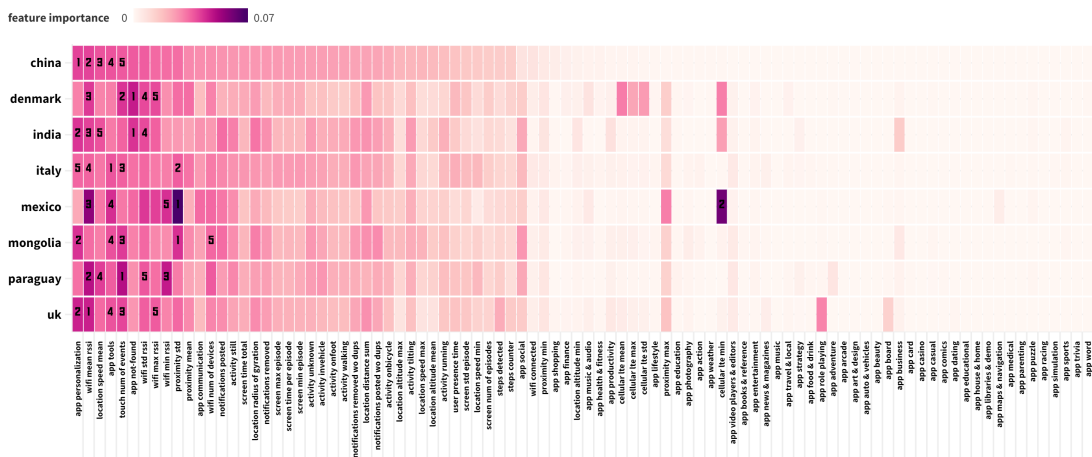


Figure 7.7: Country-Specific HM: Gini feature importance values from RF models for three-class inference.

Table 7.7: Multi-Country and Continent-Specific with PLM and HM: Mean (\bar{S}) and Standard Deviation (S_σ) of F1-scores and AUC scores obtained by testing the "worldwide" model. Results are presented as $\bar{S}(S_\sigma)$, where S is any of the two metrics.

	PLM		HM	
	Two-Class	Three-Class	Two-Class	Three-Class
Baseline	.50 (.00)	.50 (.00)	.50 (.00)	.50 (.00)
Europe	.58 (.03)	.50 (.03)	.89 (.03)	.86 (.02)
Asia	.51 (.02)	.52 (.05)	.79 (.03)	.74 (.01)
Multi-Country	.52 (.03)	.53 (.02)	.83 (.01)	.86 (.03)
Europe (Balanced)	.53 (.02)	.50 (.05)	.86 (.04)	.82 (.03)
Asia (Balanced)	.52 (.04)	.54 (.03)	.79 (.02)	.76 (.02)
Multi-Country (Balanced)	.53 (.02)	.52 (.03)	.81 (.03)	.78 (.02)

Gini Feature Importance Values

Figure 7.6 and Figure 7.7 show the Gini feature importance values for each country for two-class and three-class mood inferences with HMs. We report diagrams for HMs because they provide the highest performance. Further, the top five features within each country are marked with numbers from one to five. Moreover, in both diagrams, values are arranged in the decreasing order of values in China, from left to right. For both inferences, many apps had very low feature importance values. On the other hand, ‘app personalization’ and ‘app tools’ were among the top five features for many countries. For the UK, personalization apps were highly important in two and three-class inferences. However, for Mexico, the importance of the feature was relatively lower in both inferences. In addition, the number of touch events on the phone was within the top five features for Italy, Mongolia, Paraguay, and the UK in the two-class inference and all countries except India and Mexico in the three-class inference. This aligns with previous literature that presented findings of typing and touch events indicative of aspects such as mood and stress [285]. Another feature discussed in the literature on psychological aspects and mobile sensing [77], which appeared again in the diagrams is speed, calculated using location sensors (‘location speed mean’). Diagrams indicate that the feature was in the top five in two-class inference for India and China and three-class inference for India and Paraguay. In addition to these features, multiple features captured using Wifi signals were among the top five in all countries. Wifi-related features (i.e., ‘wifi std rssi’, ‘wifi mean rssi’, ‘wifi min rssi’, ‘wifi max rssi’ - The standard deviation/mean/minimum/maximum of RSSI signal strengths captured with unique devices within the time window) were present with high importance values for all countries across both inferences. Prior work highlights that the number of wifi devices and signal strengths could be indicative of user context, including the location [454], and location-related features have shown to be closely tied to the mood of individuals [77]. In summary, the top five features for mood inference, regardless of whether it is two-class or three-class, were not the same across all countries. Certain features are unique to individual countries. At the same time, we can also observe a specific set of features (shown in the left quarter of both figures) that consistently appeared on the top list in all countries.

7.5 Discussion

In this section, we discuss the main findings of the chapter, and highlight limitations and future work.

7.5.1 What do the Results Suggest?

In the country-specific setting, PLMs did not perform well across countries, with the highest performance for both two-class and three-class inferences coming from Mexico, with an AUC of 0.62. However, performance increased significantly, with HMs showing the effect of personalization within countries. Comparable performance gains were observed for the multi-country setting as well. However, country-specific models (AUC scores of 0.78-0.98 for two-class and 0.76-0.94 for three-class) would be preferred over multi-country models (0.83 and 0.79 for two and three classes, respectively). Then, in the country-agnostic setting, we observed that even HMs performed poorly compared to the country-specific setting. This means that if a model is trained in a different country, even if it is personalized to a person in another country, the model might not perform as well as a country-specific model that is personalized to a person in the same country. However, we also observed that models perform relatively better in culturally similar countries (i.e., within Europe). Within Asia, even though countries are in the same geographic region, cultural differences (i.e., India and China have different cultures and behaviors) could be one reason that did not allow models to perform better. Finally, building continent-specific models for Europe worked reasonably better than for Asia or a multi-country setting. Please note that the number of participants in several of these countries remained small, and so we cannot make any strong assertions.

7.5.2 Comparison of Results to Previous Studies

First, it should be noted that mood inference with smartphone sensing data is inherently a difficult task because of the task's subjectiveness. In this context, if we consider the results we obtained compared to some prior work, LiKamWa et al. [285] showed that they could achieve a 66% accuracy with population-level models, around 75% accuracy with hybrid models, and 94% accuracy with user-level models. However, comparing the results in their paper to ours is difficult because we reported results with AUC, which is a more holistic performance metric, especially in an imbalanced class scenario. However, purely in terms of numbers, the performance gain from PLM to HLM is greater in our case (from around 50% to 80%). This could be because of our dataset's more extensive set of features compared to their dataset, which only has phone usage-related features such as messages, calls, websites visited, and app usage. In addition, they modeled the inference as a regression task using multi-linear regression and provided model performance as a percentage using an error bound of 0.25 around the predicted value.

Another paper that used a similar dataset was by Servia-Rodriguez et al. [468]. It is also worth noting that this dataset contains data from multiple countries, even though the analysis did not explicitly focus on that aspect. Furthermore, they only showed results for PLMs, obtaining an accuracy of around 70% for weekends. Again, purely in terms of numbers, this is a good performance compared to what we obtained (AUC scores of about 0.5). However, it is worth noting that they only reported results for weekends, for which inference performance was high, and we do not separate weekdays and weekends. In addition, the feature sets used for inference are again different. Another potential reason for the lack of performance in our PLMs could be participants' lack of movement during the pandemic when data were collected. This could result in sensors such as location (used in both the discussed papers) not being highly informative of different moods. Hence, this could lower the performance of our models. Interestingly, one common result across all three studies was that fewer negative labels were reported, which could make the development of fully personalized models more challenging due to the lack of data for negative classes from certain individuals. Hence, future studies could look into ways of capturing negative mood labels accurately and more often using different techniques. In addition, model personalization in situations where some users lack data for certain classes is a potential problem

that could be explored further (a similar skewed labels-related scenario for depression detection has been discussed in a recent study [568]).

7.5.3 Diversity-Aware Research in Mobile Sensing

According to Gong et al. [181], diversity and diversity awareness are topics in machine learning that have gained importance in the recent past, and increasing generalization and decreasing biases in models for different populations are two fundamental goals discussed in this domain [325]. According to them, diversity is achieved in machine learning with data diversification (maximizing the informativeness in training data such that the model fits data better), model diversification (increased diversity in model parameters leading to better learning), and inference diversification (model provides choices/information with more complementary information). Our study examined diversity awareness, primarily with data diversification. Since the whole data collection was done to emphasize the need for diversity awareness in machine learning-based mobile sensing systems, we defined diversity based on social practice theory [198, 458, 173]. Accordingly, diversity is a complex and multi-layered construct that does not exist within individuals but surfaces when two or more individuals interact. Considering these conceptions, data and model diversity can be achieved by considering various types of diversity attributes ranging from country of residence, gender, and age, to personality, values, etc. [198, 458]. In this chapter, we focused on ‘country of residence’ as an attribute for analysis because of the way mood is perceived and expressed, as well as phone usage and everyday behavior are different in countries around the world. In future work, other diversity attributes could be used to study mood (e.g., studying personality and mood with mobile sensing). Furthermore, other constructs collected in the study (e.g., social context, activity, food consumption) could be examined with mobile sensing, using country as a diversity attribute.

7.5.4 Diversity-Awareness: Countries or Cultures?

In this chapter, we considered the geographical diversity of users when building smartphone sensing-based mood inference models. Hence, our primary construct of diversity is the ‘country of residence’. However, depending on the city, even though it is within the same country, the cultural composition of students could vary significantly. For example, our specific university in London, UK, is considered more diverse and has a high international student population compared to our specific university in India. These differences could also affect inference performance. In addition, our study also leaves the open question of whether the geographical region affects mobile sensing inference performance, or whether it is the culture of study participants that mediates their everyday life and phone usage behavior. Section 7.4 presented some initial results about these aspects. Future work could investigate these aspects further.

7.5.5 Ethical Considerations

Mood is a self-reported internal state and thus constitutes sensitive information. Ethical implications related to inference of affective states have been discussed in previous literature in affective computing [110, 388, 348], ubicomp [214, 372], and other disciplines [66, 346]. From the perspective of possible applications beyond supporting research on youth well-being, as we do here, it is fundamental that human-centered principles are followed and limit their use to cases that benefit individuals and avoid potential harm.

7.5.6 The Effect of the Pandemic and Weather on Mood Inference Models

In this chapter, we showed how mood inferences could be done in the context of a mobile sensing application. In addition, we also showed how models lack generalization to unseen countries and the need for personalization. However, a limitation of this study is that the study was conducted during the pandemic. During the data collection time period in 2020, many countries have imposed different measures to curb the coronavirus. However, it is worth noting that, except for China, where strict lockdown measures were not present, universities have been in remote work/study mode in all the other countries. Hence, most students engaged in their studies from home. This could be the reason why there are many app usage, touch event, proximity, and wifi related features informative about mood according to Figure 7.6, Figure 7.7, and Table 7.2. It is also worth noting that the seasons in each country during the data collection period were different. On the positive side, none of the countries were in extreme winter or summer seasons. The September-November time period in European countries is the fall season, and none of those countries faced extreme cold weather conditions during that period. At this time, the season in Mongolia was comparable to European countries like Denmark or UK. All the other countries had comparatively higher temperatures. However, given that students in all the sites were affected by movement restriction measures and were stuck at home, we believe that weather conditions might not have affected the study as much compared to a time period when student behavior in outdoor environments would significantly change based on weather conditions. However, the results should be understood and interpreted with this limitation in mind. Future work could explore the effects of seasons and weather conditions on mobile sensing-based inferences.

7.5.7 Domain Adaptation for Multi-Modal Mobile Sensing

In this chapter, we highlighted the issue of generalization and the possible distributional shifts in a mobile sensing dataset collected with the same protocol in different countries. Even though issues of generalization, biases, and domain shifts have been discussed extensively in other domains such as computer vision [301], natural language processing [144], and speech [496], smartphone/mobile sensing studies have not focused on those aspects extensively thus far [180]. Even though we provide evidence of the fundamental issue, we did not go into depth about finding a potential solution for that issue, as it is not within the scope of this chapter (especially given page limits and extensive work that would be needed). Further, even though we showed that model personalization (hybrid setting) could minimize domain shift to an extent, other advanced techniques inspired by the work related to domain shift/adaptation in other domains could provide cues for solving such problems in mobile sensing. Recent studies also suggest that domain adaptation techniques for time series data are limited [562]. For example, a longstanding problem in the human activity recognition (HAR) domain is the wearing diversity of wearables in different body positions. The wearing diversity hinders the performance of HAR models. A few recent studies suggested that unsupervised domain adaptation could be a solution for wearing diversity issues [85, 315]. Further, Wilson et al. [562] explored domain adaptation for similar datasets captured from people from two age groups. However, the above studies focused on time series accelerometer data, which are more straightforward than the multi-modal datasets we are working with within this study. Hence, to the best of our knowledge, a research gap lies in solving domain adaptation for multi-modal sensing data coming from smartphones and wearables. In fact, in a recent study, Adler et al. [7] discussed the issue of generalization in multi-modal mobile sensing data and showed that lack of similarity across datasets collected in different time periods does not allow studying generalization of techniques to a greater depth. Therefore, with the dataset discussed in this chapter, we believe solutions to domain adaptation and generalization could be explored further

(not regarding generalization across time, but across geographically/culturally distinct areas), hence pushing the boundaries of multi-modal mobile sensing systems towards more real-world utility.

7.5.8 Other Limitations and Future Work

This work has several limitations and areas that could be improved in future work. First, the dataset used in this study is highly imbalanced, where there are fewer negative and very negative mood labels than neutral, positive, and very positive mood labels. However, this distribution is in a way similar to previous studies about valence [468, 285]. Inherently, this also makes both inference tasks much harder. On the other hand, there is an imbalance in the dataset regarding data per country, where Italy and Mongolia had a significantly higher number of self-reports. In addition to the experimental results that we reported with imbalanced datasets, we conducted experiments with stratified down-sampled datasets for each country (each country having samples equal to the number of India, which had the lowest number of self-reports). While we reported some results for balanced cases in multi-country and continent-specific cases, more extensive analysis could be done to explore that aspect further. Hence, diversity-aware sampling strategies could be explored in future work to mitigate biases in mobile sensing-based inference models. Further, we only considered valence in the circumplex mood model in this study. Other time diary questions were used to capture other behaviors and contexts, and we did not want to overburden users with multiple questions or lengthy questionnaires. However, we agree that collecting the arousal and understanding the geographical diversity of arousal inference could be studied in future work. In addition, the clinical validity of the valence in the circumplex mood models might be questionable. Future work could look into conducting studies with more clinically valid instruments for mood inference. In addition, in this chapter, we did not use a 'wrapper' feature selection technique before training models because tree-based models, such as random forest, inherently use 'embedded' feature selection with Gini impurity to find a set of good features to build the trees with [530], especially when the feature space is small (i.e., around 100 in this dataset). However, if the feature space was larger, the dataset size was smaller, or if another non-tree-based model was used, using feature selection is highly preferred. Therefore, future work could also look into improving models based on feature selection and finding solutions to the issue of generalization using careful feature selection.

7.6 Conclusion

In this exploratory study, we used a mobile sensing dataset with around 329K self-reports from 678 participants in eight countries (China, Denmark, India, Italy, Mexico, Mongolia, Paraguay, UK) for over three weeks to assess the effect of geographical diversity on mood inference models. We evaluated country-specific, continent-specific, country-agnostic, and multi-country approaches trained on sensor data for two mood inference tasks with population-level (non-personalized) and hybrid (partially personalized) models. We showed that partially personalized country-specific models perform the best yielding AUC scores in the range of 0.78-0.98 for two-class (negative vs. positive) and 0.76-0.94 for three-class (negative vs. neutral vs. positive) inference. Further, with the country-agnostic approach, we showed that models do not perform well compared to country-specific settings, even when models are partially personalized. We also uncovered generalization issues of sensing-based mood inference models to new countries. We hope that these findings will be of benefit to ubicomp researchers towards building future mobile sensing applications with an awareness of geographical diversity.

8 Complex Daily Activities, Country-Level Diversity, and Smartphone Sensing

Smartphones enable understanding human behavior with activity recognition to support people's daily lives. Prior studies focused on using inertial sensors to detect simple activities (sitting, walking, running, etc.) and were mostly conducted in homogeneous populations within a country. However, people are more sedentary in the post-pandemic world with the prevalence of remote/hybrid work/study settings, making detecting simple activities less meaningful for context-aware applications. Hence, the understanding of (i) how multimodal smartphone sensors and machine learning models could be used to detect complex daily activities that can better inform about people's daily lives, and (ii) how models generalize to unseen countries, is limited. We analyzed a subset of MUL dataset from Chapter 2—smartphone data and ~216K self-reports from 637 college students in five countries (Italy, Mongolia, the UK, Denmark, and Paraguay). Then, we defined a 12-class complex daily activity recognition task and evaluated the performance with different approaches. We found that even though the generic multi-country approach provided an AUC of 0.70, the country-specific approach performed better with AUC scores in [0.79-0.89]. We believe that research along the lines of diversity awareness is fundamental for advancing human behavior understanding through smartphones and machine learning, for more real-world utility across countries. The material of this chapter was originally published in [24].

8.1 Introduction

The field of activity recognition has gained substantial attention in recent years due to its usefulness in various domains, including healthcare [492], sports [598], transportation [364], and human well-being [325]. For instance, fitness-tracking mobile health applications enable users to access activity-specific metrics [598, 476]. Similarly, smart home systems can make changes to the environment (e.g., lighting, temperature) based on the information gathered about people's activities [309, 367]. Context awareness, a key aspect of mobile phone user experience, is enabled with the integration of activity recognition [547, 376].

Traditionally, sensor-based activity recognition relied on custom sensors attached to the body [93]. While this approach is effective for small-scale studies, it is often challenging to scale up. The cost and maintenance required for these sensors can make them both expensive and obtrusive, reducing the motivation to use them. The alternative approach of using commercial wearables is not immune to these challenges, and these devices are often perceived as niche or abandoned after a short period of usage [332, 107]. This is where the presence of smartphones comes in handy. In the United States, 85% of adults and 96% of young adults own a smartphone, making it easier to target a broader audience

[343]. Research in mobile sensing has revealed the potential of smartphone data for activity recognition [492, 325]. The widespread ownership and unobtrusive nature of smartphones make them an attractive solution to traditional sensor-based activity recognition. However, there is still a need to understand how multiple sensing modalities in smartphones can be utilized for complex daily activity recognition. Additionally, the generalization of complex daily activity recognition models across different countries remains an under-explored area of research.

Recognizing complex daily activities is important. In the activity recognition literature, multiple types of activities have been considered, each at different granularity levels [130, 447]. Coarse-grained or simple activities like walking, sitting, or cycling are repeated *unitary* actions directly measurable from a proxy (e.g., inertial sensor unit). Fine-grained complex activities, or activities of daily living (ADL), are built on top of simple activities, but convey more specific contextual information [560, 421, 447]. For example, eating, studying, working, and movie watching entail participants sitting. Such activities can not be measured by inertial sensor units alone [68, 329, 55] and need a more holistic multimodal sensing approach that captures a wide range of contexts and behaviors that build on top of simple activities [447]. Further, recognizing such complex daily activities could: (i) allow tracking the digital well-being of individuals in a more fine-grained manner (e.g., providing a breakdown of time spent eating, resting, attending a lecture, and studying, instead of just sitting [470, 68]); (ii) provide context-aware user experiences and notifications by understanding user behavior better (e.g., not sending phone notifications when a person is studying or attending a lecture, suggesting products while a user is shopping [331]); and (iii) allow better content recommendation (e.g., recommending music based on the current daily activity such as working, studying, or shopping [547]), where complex activities can be more informative and valuable than simpler ones. However, even though inertial, location, or WiFi/Bluetooth data have been used separately for activity recognition [421, 447], prior work has not exhaustively studied complex daily activities by using multimodal smartphone sensing data.

The use of multimodal smartphone sensing data in machine learning models could provide a more comprehensive picture of complex daily activities when compared to using single modalities. This is especially relevant in light of the Covid-19 pandemic, which has brought about a significant shift in daily habits and activities [491, 594]. The lockdown measures enforced to slow the spread of the virus resulted in a decrease in physical activity and an increase in sedentary behavior, particularly among young adults. This shift is evident in changes to smartphone use patterns [427, 446, 283], which can impact the effectiveness of location-based activity recognition methods in a remote/hybrid work/study setting where individuals tend to remain sedentary for extended periods of time. Hence, the importance of inertial and location sensors as predictive features could diminish due to sedentary behavior. This underscores the importance of incorporating fine-grained multimodal sensing features to accurately characterize the complex daily activities of these emerging lifestyles through smartphones. However, there is currently little understanding of which smartphone sensing features are systematically useful in characterizing different complex daily activities.

Taking a few steps back, we can also consider the “country” dimension and its influence on smartphone usage. Country differences can affect smartphone usage in different world regions [316]. For example, it could be socially frowned upon to take a call at a formal restaurant in Japan, while people in Europe could leave a movie theater to check their phone [76]. It has been shown that people in Japan tend to be more reticent than in Sweden about talking on the phone in public transportation or, more generally, about being loud in public [38]. Another study about smartphone addiction among young adults in 24 countries found that the rigidity of social norms and obligations highly influenced smartphone usage [378]. In addition to how people use the phone, prior work also discussed how passively sensed behavioral data about people differ in many countries [17]. These differences across countries constitute

a form of diversity, which is a growing area of interest in computing and AI research [115]¹. From a machine learning point-of-view, a diversified system contains more information and can better fit various environments [181]. More generally, diversity-aware machine learning aims to improve the model's representational ability through various components such as input data, parameters, and outputs [181]. Concretely, country-level, diversity-aware activity recognition should try to understand the effect of the country diversity of smartphone users, on inference model performance. However, the understanding of how country diversity affects the smartphone sensing pipeline (from collected data to model performance) is limited, as previous work aimed at quantifying such effects has been scarce [247, 325, 402], due to reasons including, but not limited to, logistical difficulties in conducting longitudinal smartphone sensing studies with the same protocol in diverse countries.

This chapter looks into mimicking the experimental approach proposed in Chapter 7 for complex daily activity recognition. Hence, this chapter uses a set of experimental approaches (country-specific, country-agnostic, and multi-country, described in Table 7.1), and model types (population-level and hybrid, described in Table 7.1 and operationalized in Section 8.5). With the support of rich multimodal smartphone sensing data collected in multiple countries under the same experimental protocol, we address three research questions:

RQ1: How are complex daily activities expressed in different countries, and what smartphone sensing features are the most useful in discriminating different activities?

RQ2: Is a generic multi-country approach well-suited for complex daily activity recognition? To which extent can country differences be accurately modeled by country-specific approaches?

RQ3: Can complex daily activity recognition models be country-agnostic? In other words, how well do models trained in one or more countries generalize to unseen countries?

In addressing the above research questions, we provide the following contributions:

Contribution 1: We examined a subset of MUL dataset, with over 216K self-reports (including complex daily activities) collected from 637 college students in five countries (Denmark, Italy, Mongolia, Paraguay, and the United Kingdom) for over four weeks. We defined 12 complex daily activity classes based on participant responses, prevalence, and prior work. The list includes sleeping, studying, eating, watching something, online communication and social media use, attending classes, working, resting, reading, walking, sports, and shopping. On the one hand, we found that similar features are most informative for all countries for specific activities (e.g., sleep, shopping, walking). On the other hand, for some other activities, the most informative features vary across countries. Interestingly, however, they remain approximately similar across geographically closer countries. For example, the "sport" activity has the use of "health & fitness apps" as a top feature across European countries. However, the feature was not prominent in Mongolia and Paraguay, where such physical activity-related app usage is lower. This divide is also visible in the "watching something" activity, which is influenced by the use of entertainment apps in European countries, and not in the other two countries.

¹While we acknowledge that cultures can be multidimensional and exist in tension with each other and in plurality within the same country [584], some prior studies in mobile sensing, psychology, and sociology have used "culture" as a proxy to refer to the country of data collection [402, 247, 532, 201]. However, in this study, for consistency, we use "country" (a more specific geographic region) as the unit of analysis that could affect phone usage behavior and sensing data. We also used the term "geographic" rarely, when appropriate and when referring to regions (i.e., Europe).

Contribution 2: We defined and evaluated a 12-class complex daily activity inference task with country-specific, country-agnostic, and multi-country approaches (similar to Table 7.1 in Chapter 7). We also used population-level (not personalized) and hybrid (partially personalized) models to evaluate how model personalization affects performance within and across countries. We show that the generic multi-country approach, which directly pools data from all countries (a typical approach in many studies), achieved an AUC of 0.70 with hybrid models. Country-specific models perform the best for the five countries, with AUC scores in the range of 0.79-0.89. These results suggest that even though multi-country models are trained with more data, models could not encapsulate all the information towards better performance, possibly due to the averaging effect of diverse behaviors across countries. The country-specific approach consistently worked better.

Contribution 3: With the country-agnostic approach, we found that models do not generalize well to other countries, with all AUCs being below 0.7 in the population-level setting. With hybrid models, personalization increased the generalization of models, reaching AUC scores above 0.8, but not up to the same level as country-specific hybrid models. Moreover, even after partial personalization, we observed that models trained in European countries performed better when deployed in other European countries than in Mongolia or Paraguay. This shows that in addition to country diversity, behavior, and technology usage habits could be what mediates the performance of models in different countries. In light of these findings, we believe that human-computer interaction and ubiquitous computing researchers should be aware of machine learning models' geographic sensitivities when training, testing, and deploying systems to understand real-life human behavior and complex daily activities. We also highlight the need for more work to address the domain shift challenge in multimodal mobile sensing datasets across countries.

To the best of our knowledge, this is the first study that focuses on the use of multimodal smartphone sensing data for complex daily activity recognition, while examining the effect of country-level diversity of data on complex activity recognition models with a large-scale multi-country dataset, and highlighting domain shift-related issues in daily activity recognition, even when the same experimental protocols are used to collect data in different countries.

The chapter is organized as follows. In Section 8.2, we describe the related work and background. Then, we describe the dataset in Section 8.3. In Section 8.4, we present the descriptive and statistical analysis regarding important features. We define and evaluate inference tasks in Section 8.5 and Section 8.6. Finally, we end the chapter with the Discussion in Section 8.7 and the Conclusion in Section 8.8.

8.2 Background and Related Work

8.2.1 Mobile Sensing

In prior work, researchers have collected and analyzed mobile sensing data to understand various attributes of a particular population. Depending on the study, that goal can be put under coarse categories such as behavior, context, and person-aspect recognition [325]. Behavior recognition is aimed at understanding user activities broadly. Person aspect recognition looks into understanding demographic attributes (e.g., sex, age, etc.), psychology-related attributes (e.g., mood, stress, depression, etc.), and personality. Finally, context recognition identifies different contexts (e.g., social context, location, environmental factors, etc.) in which mobile users operate.

Regarding behavior recognition, there are studies that aimed to capture binary (sometimes three) states

of a single complex activity/behavior such as eating (e.g., eating meals vs. snacks [55], overeating vs. undereating vs. as usual eating [330]), smoking (e.g., smoking or not [321]) or drinking alcohol (e.g., drinking level [404, 30], drinking or not [454]). Another study used the action logs of an audio-based navigation app to predict its usage and understand what drives user engagement [294]. Then, regarding person aspects, the MoodScope system [285] inferred the mood of smartphone users with a multi-linear regression based on interactions with email, phone, and SMS, as well as phone location and app usage. Servia-Rodriguez et al. [468] observed a correlation between participants' routines and some psychological variables. They trained a deep neural network that could predict participants' moods using smartphone sensor data. Additionally, Khwaja et al. [247] developed personality models based on random forests using smartphone sensor data. Finally, context recognition is aimed at detecting the context around behaviors and activities. [328] used sensing data from Switzerland and Mexico to understand its relation to the social context of college students when performing eating activities. More specifically, they built an inference model to detect whether a participant eats alone or with others. Similarly, [327] examined smartphone data from young adults to infer the social context of drinking episodes using features from modalities such as the accelerometer, app usage, location, Bluetooth, and proximity. In this case, context detection is two-fold: it's based on the number of people in a group, and on their relationship to the participant (e.g., alone, with another person, with friends, with colleagues). Similarly, mobile sensing studies attempted to infer other contexts, psychological traits, and activities by taking behavior and contexts sensed using smartphone sensors as proxies [325, 108, 217].

One common aspect regarding most of these studies is that they were done in the wild, focused on two or three-class state inference, and sensing is not fine-grained (i.e., using behavior and context as proxies to the dependent variable). This chapter follows a similar approach with a dataset captured in the wild, using multimodal smartphone sensor data, and taking behavior and context as proxies for our dependent variable. However, in this study, the target attribute entails a 12-class daily activity recognition problem that is complex and novel compared to prior work. In addition, we are interested in examining model performance within and across five countries, with and without partial personalization.

8.2.2 Activity Recognition

Human activity recognition (HAR) aims to understand what people are doing at a given time. Large-scale datasets issued from the activity of smartphone users have a lot of potential in solving that task. This "digital footprint" has been used to re-identify individuals using credit-card metadata [350]: it has been shown that only 4 data points are required to re-identify 90% of individuals. While the same approach could be followed using smartphone sensing data, our main focus is activity recognition at a single point in time rather than using time series for re-identification. We will focus on two types of activity recognition techniques: wearable-based and smartphone-based [492].

Wearable-based HAR

In wearable-based activity recognition, the users wear sensors such as wearable accelerometers from which the data is analyzed and classified to detect activities. For example, in healthcare, wearable-based HAR can be used to analyze gait and prevent falling or monitor physical activity and observe health outcomes [293]. The wearable-sensing trend emerged two decades ago and relied on custom-designed wearable sensors [151, 387], which were backed by encouraging findings in health research. With time, custom sensors were replaced by commercial fitness or activity trackers. Unfortunately, applying these findings to real-world settings was rare due to the high cost of producing custom sensors, the

difficulty distributing devices to a broad audience, and their unpopularity among some users [107]. This restricted most studies using wearables to performing experiments in a controlled environment or in the wild with smaller populations. However, wearable-based HAR models that could recognize simple activities are currently deployed across many commercial wearable devices.

Smartphone-based HAR

With the popularity of smartphones in the past two decades, the problems of wearable-based HAR were solved. Reality Mining [143] is a pioneering study in the field of mobile sensing: it showed the utility of mobile sensing data in a free-living setting. In smartphone-based activity recognition, people do not need to use wearable sensors. Instead, the system relies on a smartphone that is always on and stays closer to its user. Smartphones replace wearable devices as the former contains multiple sensors such as an accelerometer, gyroscope, GPS, proximity, or thermometer. Nevertheless, smartphones capture data at multiple positions (e.g., a pocket, hand, or handbag), which introduces a bias in sensor measurements as they are position-dependent [574].

Regardless of the device used, most prior activity recognition tasks have been done in lab-based/controlled settings where accurate ground truth capture is possible [492]. The prime goal of such studies is to increase the accuracy of activity recognition models with precise ground truth and sensor data collection (e.g., by placing sensors on fixed body positions, recording ground truth with videos, etc.). However, these studies are hard to scale and do not capture the real behavior of participants, and this is especially true for complex daily activities [447]. For example, a person's behavior when studying, working, or shopping in an unconstrained environment can not be replicated in a lab. On the other hand, some studies are done in the wild [271, 447], where the ground truth and sensor data collection might not be that precise but allow capturing complex daily activities in a naturalistic setting. Our study is similar, where our intention was to take a more exploratory stance, build country-level diversity-aware models, and compare their performance within and across different countries.

8.2.3 Activity Types

One crucial difference across existing studies is in the selection of activities. A majority of studies work towards the recognition of simple activities. For example, Straczekiewicz et al. [492] classified activities into groups such as posture (lying, sitting, standing), mobility (walking, stair claiming, running, cycling), and locomotion (motorized activities). Laput and Harrison [271] called such activities coarse or whole-body. Activities belonging to these groups are directly measurable from one or more proxies (e.g., inertial sensor unit, location). For example, when considering the accelerometer, each activity has a distinct pattern on the different axes [130]. However, they constitute a small subset of activities performed by people in daily lives [447, 421, 100].

Notice that some of the simple activities described above are usually part of more complex activities (e.g., sitting while eating, walking while shopping). Dernbach et al. [130] defined complex activities as a series of multiple actions, often overlapping. Along with Bao et al. [35], they used the same techniques to recognize both simple and complex activities. This results in weaker performances for complex activities since their structure is more complicated. Another approach is considering complex activities hierarchically by using combinations of simple activities to predict more complex ones. Huynh et al. [224] characterized user routines as a probabilistic combination of simple activities. Blanke et al. [58] used a top-down method that first identifies simple activities to recognize complex ones. However,

this requires pre-defining simple activities and mappings to complex activities. Some studies focus on detecting binary episodes of a single complex activity or a specific action. For example, the Bites'n'Bits study [55] examined the contextual differences between eating a meal and a snack, and presented a classifier able to discriminate eating episodes among students. Likewise, DrinkSense [454] aimed at detecting alcohol consumption events among young adults on weekend nights. Unfortunately, such task-specific classifiers will perform poorly when exposed to situations they were not trained on.

In this study, we focus on a majority of complex daily activities (11 out of 12 and one simple activity: walking) derived by considering over 216K self-reports from college students in five countries. In this context, drawing from prior studies that looked into activities of daily living [447, 421], for the scope of this chapter, we define complex activities as *"activities that punctuate one's daily routine; that are complex in nature and occur over a non-instantaneous time window; and that have a semantic meaning and an intent, around which context-aware applications could be built"*. While it is impossible to create a classifier that could recognize all complex human activities, we believe the classifier we propose captures a wide range of prevalent activities/behaviors, especially among young adults.

8.2.4 Diversity-Awareness in Smartphone Sensing

Research in the field of smartphone sensing, including the studies mentioned above, lacks diversity in their study populations [325]. Regarding country diversity, with a few exceptions [468, 247], most experiments were conducted in a single country or rarely two. This can be problematic with respect to the generalization of findings since smartphone usage differs across geographic regions, which can lead to different patterns being observed in, for example, two populations of different genders or age range [126]. Khwaja et al. stressed the importance of diversity awareness in mobile sensing [247]. Moreover, experiments performed in a controlled setting usually can not accommodate many participants. While this makes the whole process lighter and more manageable, it also restricts the generalization of results to a broader free-living audience [456, 206]. According to Phan et al. [402], cross-country generalizability is the extent to which findings apply to different groups other than those under investigation.

Diversity awareness and model generalization are two essential aspects, as they will allow an activity recognition system to be deployed and to perform well across different user groups and countries [324, 458]. In computer vision research, the lack of diversity has been repeatedly shown for specific attributes such as gender and skin color [422, 116, 243]. In natural language processing and speech research, not accounting for dialects in different countries could marginalize groups of people from certain countries [410]. Hence, ignoring country diversity when developing AI systems could harm users in the long run by marginalizing certain groups of people [410]. In this context, smartphone sensing studies that consider country-level diversity are still scarce [402]. This could be due to the lack of large-scale datasets, logistical difficulties in data collection in different countries, and studies being time and resource-consuming. Khwaja et al. [247] built personality inference models using smartphone sensor data from five countries and showed that such models perform well when tested in new countries. To the best of our knowledge, their study is one of the first to investigate the generalization of smartphone sensing-based inference models across different countries. In our work, we focus on complex daily activity recognition with smartphone sensing and aim to uncover and examine model behavior in multi-country settings.

8.2.5 Human-Centered Aspects in Smartphone Usage

Our literature review has so far focused on the technical aspects such as data collection or target variables. We now discuss the impact of smartphone usage on individuals and society, which is studied by various disciplines in the social sciences. Previous work includes the study of smartphone dependence among young adults, where it was found that problematic smartphone use varies by country and gender [296, 521], and those specific activities such as social networking, video games, and online shopping contribute to the addiction [296, 378]. Another study [428] summarized findings on correlations between smartphone usage and psychological morbidities among teens and young adults. Excessive smartphone usage could lead to emotional difficulties, impulsivity, shyness, low self-esteem, and some medical issues such as insomnia, anxiety, or depression. From a sociological standpoint, Henriksen et al. [208] studied how smartphones impact interactions in cafés and defined three concepts of social smartphone practices. *Interaction suspension* (e.g., your friend goes to the bathroom), which can lead to using the smartphone to appear occupied or to avoid uncomfortable situations while being alone. *Deliberate interaction shielding* corresponds to situations where one suspends an ongoing interaction to answer a phone call or a text message, whether it is an emergency or just in fear of missing out. *Accessing shareables*, which leads to a collective focus on shared content (e.g., pictures or short videos), giving the smartphone a role of enhancing face-to-face social interactions rather than obstructing them. Nelson and Pieper [368] showed that smartphone attachment “inadvertently exacerbates feelings of despair while simultaneously promises to resolve them”, thus trapping users in negative cycles.

According to Van Deursen et al. [521], older populations are less likely to develop addictive smartphone behaviors. While they are often associated with younger generations, smartphones are slowly gaining popularity among older generations as they are coming up with creative ways to integrate them into their habits. Miller et al. [336] investigated the role that smartphones play in different communities across nine countries. Through 16-month-long ethnographies, they showed that various groups of people have specific ways of taking ownership of their smartphones through apps, customization, and communication. For example, in Ireland, smartphones are used by the elderly in many of their daily activities, and in Brazil, the usage of messaging applications for health has led to the creation of a manual of best practices for health through such applications. More globally, smartphones can help users stay in touch with their extended families or distant friends, a feature that has been particularly important during the 2020 global pandemic. In this chapter, we attempt to uncover country-specific smartphone usage patterns through multimodal sensing data. While these insights may not have the depth that field observations provide, they represent a starting point for future research to draw upon.

Hence, all while considering these factors, we aim to examine smartphone sensing-based inference models for complex daily activity recognition with country-specific, country-agnostic, and multi-country approaches, as described in Figure 8.1.

8.3 Data, Features, and Target Classes

8.3.1 Dataset Information

To address our research questions, we used a subset of the 8-country dataset presented in Chapter 2. We considered only five countries (Italy, Denmark, UK, Mongolia, and Paraguay) because data preparation activities in other countries were not finalized by the time of conducting experiments related to this chapter.

8.4 How are activities expressed in different countries, and what smartphone features are most discriminant? (RQ1)

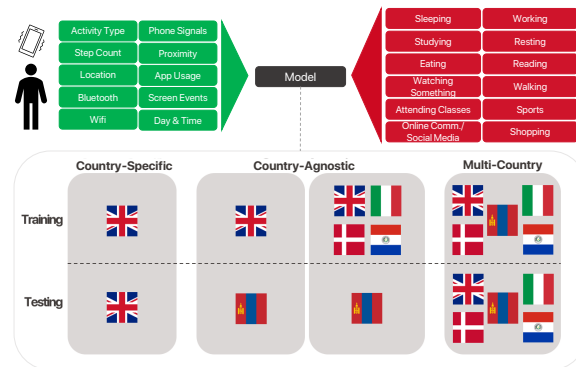


Figure 8.1: High-level overview of the study. The study uses continuous and interaction sensing modalities and different approaches (country-specific, country-agnostic, and multi-country) to infer complex daily activities.

8.3.2 Determining Target Classes

Hourly self-reports required participants to log what they were doing at the time by selecting an activity from a predefined list of thirty-four items. These items were derived based on prior work [174, 592]. By looking at their distribution in different countries (Figure 8.2), one can quickly notice that they are highly unbalanced. The remote work/study constraints during the time of data collection were one of the causes behind this imbalance, because activities such as traveling, walking, or shopping would have been more popular if mobility was not restricted. A closer look at the list of activities shows that some classes are too broad in terms of semantic meaning. Hence, similar to prior work that narrowed down activity lists based on various aspects [271], we narrowed down the original list of activities into 12 categories to capture complex daily activities that are common enough in the daily lives of people, especially in a remote work/study setting. For example, under “hobbies”, one can be playing the piano or painting, and the two do not entail the same smartphone usage and are not common enough. Similarly, “social life” is too broad, as one could be in a bar, a restaurant, or a park. Moreover, to mitigate the class imbalance problem, we decided to filter the target classes. First, classes that had similar semantic meanings were merged: this is the case of eating and cooking, and social media and internet chatting. Classes representing a broad activity were removed, such as personal care, household care, games, and hobbies. Finally, classes that did not have enough data in all countries were removed, such as listening to music, movie, theatre, concert, and free-time study. Filler classes such as “nothing special” or “other” were also removed. This filtering reduced the number of target classes to twelve, and their updated distribution is shown in Figure 8.3. These classes entail activities performed during daily life that are complex in nature and have a semantic meaning around which context-aware applications could be built. Moreover, the selected activities also align with prior work that looked into complex daily activity recognition [447].

8.4 How are activities expressed in different countries, and what smartphone features are most discriminant? (RQ1)

To understand the distribution of activities in each country and to determine the influence of features on the target, we provide a descriptive and statistical analysis of the dataset in this section, hence shedding light on RQ1.

Table 8.1: A summary of participants of the data collection. Countries are sorted based on the number of participants.

University	Country	Participants	μ Age (σ)	% Women	# of Self-Reports
University of Trento	Italy	259	24.1 (3.3)	58	116,170
National University of Mongolia	Mongolia	224	22.0 (3.1)	65	65,387
London School of Economics & Political Science	UK	86	26.6 (5.0)	66	20,238
Universidad Católica "Nuestra Señora de la Asunción"	Paraguay	42	25.3 (5.1)	60	6,998
Aalborg University	Denmark	26	30.2 (6.3)	58	7,461
Total/Mean		637	24.0 (4.3)	62	216,254

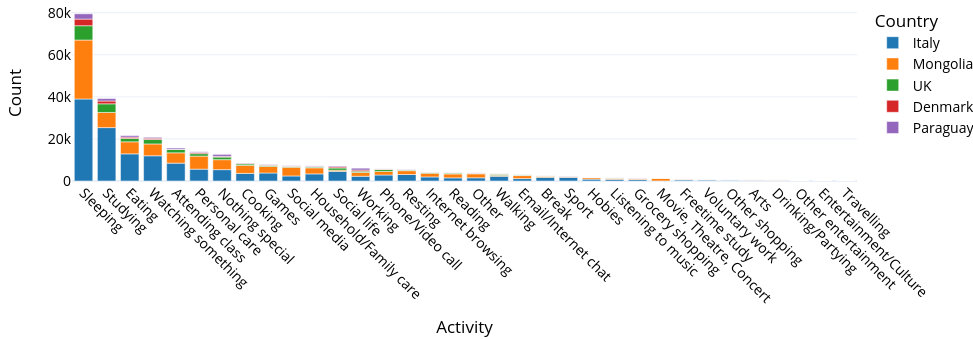


Figure 8.2: The original distribution of target classes before any filtering or merging was done.

8.4.1 Hourly Distribution of Activities

The activities we consider all seem to occur at different times: people tend to sleep at night, work during the day, and eat around noon and in the evening. However, not all schedules are the same, especially not across different countries [154, 147]. We reported the density function of each target class at different hours of the day in Figure 8.4. In each diagram, the x-axis refers to the hour of the day, and the y-axis refers to the density of each activity. On an important note, while most activities were reported as they were being performed, in the case of sleeping, participants reported the activity after they woke up and still in bed, meaning that peaks for that activity could also be interpreted as “waking up”. This was later confirmed with many participants in all countries during post-study interviews. This also makes the time of the day less informative when inferring the sleeping activity.

A first look at the distribution shows the “expected” patterns, such as a peak of sleeping during the night or peaks around eating times for lunch and dinner. Notice that participants from Paraguay tend to sleep less than others, reflecting that they start working and resting earlier in the day. Online communication and social media usage happen around noon, coinciding with a break from classes and lunchtime, followed by a high peak towards the end of the day. This is in line with prior studies that showed that depending on the location context and hour of the day, the use of certain social media applications (i.e., Twitter) could differ [127]. Moreover, we also observe country differences in hourly social media and online communication app usage patterns as reported by users. For example, between noon and 6 pm, there is a dip in the usage of these types of apps in Italy, Paraguay, and Denmark, whereas that pattern is not visible in the UK. Prior work has also studied social media app usage and adoption-related differences, especially across countries. As per those studies, such usage differences could result from

8.4 How are activities expressed in different countries, and what smartphone features are most discriminant? (RQ1)

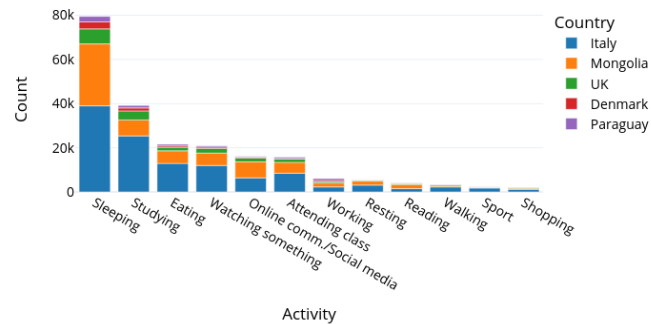


Figure 8.3: Distribution of target classes after removing classes that are semantically broad or lack data.

cultural characteristics within countries and from motives of people for using different apps [14, 286]. Most leisure activities (reading, shopping, sport, watching something) happen towards the end of the day, right when students have finished their classes.

Another activity that showed clear cross-country differences is “Eating”. We can observe that Italians tend to eat later than others, which hints at their Mediterranean customs [513]. Italy also showed two clear peaks for lunch and dinner with a sharp dip in between the two meals. The dip is less visible in other countries, indicating that meals are more spread out across different times. Moreover, the dinner peaks for all countries except Mongolia were peaking on or after 6 pm, whereas in Mongolia, it was before 6 pm. These findings suggest that the hour of the day could indicate whether people are eating or not—slightly differently in Italy, Mongolia, and other countries. In fact, prior studies that used mobile sensors for studies regarding eating behavior showed that the hour of the day is an important feature in predicting aspects related to eating [55, 330]. To add to that, prior studies have also pointed out that meal times, frequency, and sizes could differ between countries [89], even within Europe [484]. Finally, the activity “walking” had more or less similar distributions across countries. In fact, a smartphone-based activity tracking study by Althoff et al. [17] mentioned that the average number of steps walked by people across Italy, the UK, Denmark, and Mongolia were in the same ballpark (i.e., around 5000-6000 daily steps).

8.4.2 Statistical Analysis of Features

To understand the importance of each smartphone sensing feature in discriminating each target activity from others, we reported in Table 8.2 the top three features and their ANOVA (Analysis of variance) F-values [248] for each activity and each country. The goal is to identify features that define an activity and how those differ across countries. We consider each country-activity pair alone to find features that influence the classification task in a binary setting (i.e., determining whether the participant is sleeping or not, studying or not, eating or not, etc.).

The resulting features across countries for the same activity are different in most cases, highlighting the dataset’s diversity and each country’s cultural differences or habits. For example, when studying, features regarding screen episodes dominate in the UK, Italy, and Denmark, while the day period appears in Italy, Mongolia, and Paraguay. This could mean that European students tend to use their phones when studying more (or less) than students from Paraguay or Mongolia. This divide is also visible when “watching something”, which is influenced by the use of entertainment applications in

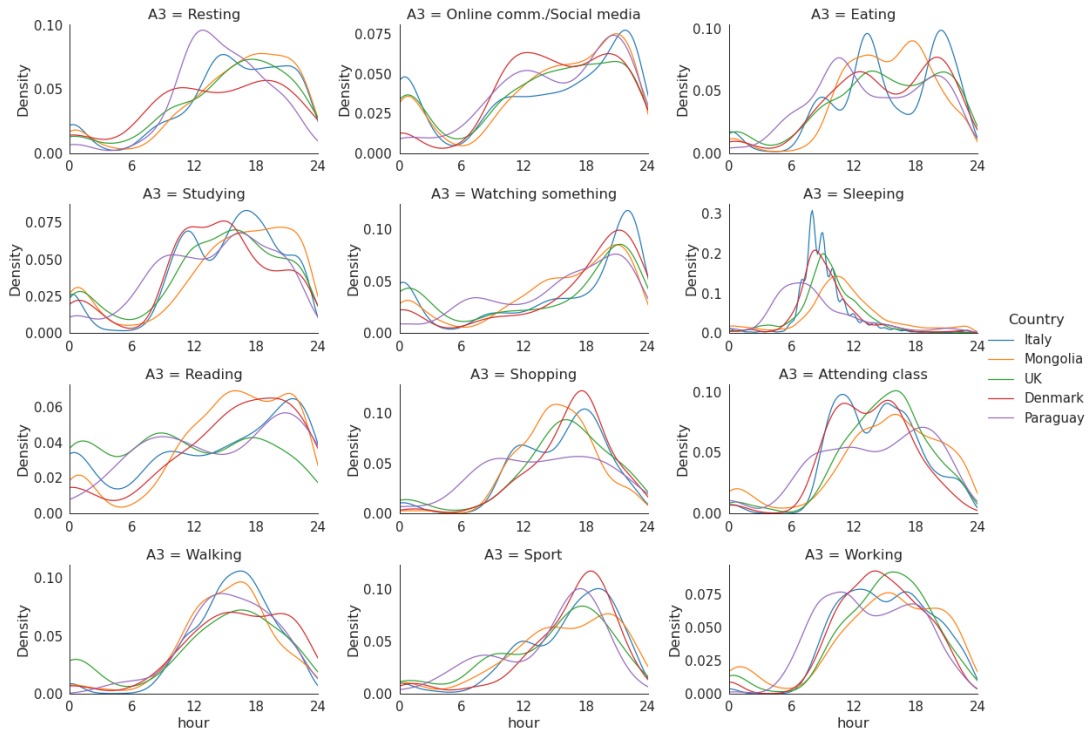


Figure 8.4: Density functions of target classes as a function of the hour of day in each country.

Table 8.2: ANOVA F-values (F) with p-value < 0.05 for each target activity and each country. The best feature is the first in the list. Comparing F-values are only valid locally within the same activity and country.

	Italy		Mongolia		UK		Denmark		Paraguay	
	Feature	F	Feature	F	Feature	F	Feature	F	Feature	F
Sleeping	app_tools	5423	day_period	6623	day_period	1632	day_period	354	app_not-found	510
	day_period	4439	app_not-found	3595	screen_max_episode	603	screen_max_episode	249	noti_removed_wo_dups	348
	screen_max_episode	2498	noti_removed_wo_dups	1052	screen_time_per_episode	534	screen_time_per_episode	156	notifications_posted_wo_dups	289
	screen_time_total	1378	day_period	683	screen_time_total	446	screen_max_episode	241	app_video players & editors	147
Studying	day_period	1146	noti_removed_wo_dups	220	screen_max_episode	396	screen_time_total	225	app_not-found	84
	day_period	271	app_photography	178	screen_time_per_episode	247	weekend	154	day_period	43
	app_tools	98	day_period	518	day_period	61	proximity_std	38	app_not-found	37
	app_not-found	61	app_not-found	180	app_not-found	26	proximity_max	29	wifi_mean-rssi	23
Eating	app_entertainment	715	activity_still	72	app_video players & editors	23	app_communication	18	wifi_max-rssi	21
	app_not-found	426	day_period	326	app_video players & editors	397	app_entertainment	151	wifi_mean-rssi	51
	weekend	334	app_not-found	325	wifi_std_rssi	85	app_not-found	59	app_lifestyle	38
	app_social	1381	wifi_num_of_devices	217	app_entertainment	66	notifications_posted	58	weekend	29
Watching something	screen_time_total	565	touch_events	503	wifi_num_of_devices	112	app_tools	64	app_tools	95
	screen_max_episode	473	screen_time_total	355	wifi_connected	93	app_causal	58	proximity_max	58
	weekend	3167	app_not-found	354	screen_time_total	92	screen_time_total	42	proximity_mean	48
	screen_num_of_episodes	745	day_period	455	weekend	357	app_not-found	119	notifications_posted_wo_dups	148
Attending class	app_tools	476	weekend	289	day_period	260	notifications_posted	104	weekend	112
	steps_detected	271	app_not-found	251	screen_max_episode	70	screen_max_episode	37	screen_time_total	87
	screen_time_per_episode	210	wifi_mean-rssi	1049	screen_time_per_episode	143	proximity_mean	305	activity_invehicle	441
	screen_num_of_episodes	206	wifi_max_rssi	848	proximity_mean	129	proximity_max	304	wifi_num_of_devices	226
Working	day_period	337	wifi_min_rssi	633	screen_max_episode	124	proximity_std	292	activity_walking	163
	app_tools	117	day_period	191	app_medical	374	notifications_posted	22	app_photography	145
	app_educational	66	screen_time_total	89	app_arcade	72	app_not-found	16	app_trivia	64
	app_books & reference	955	screen_max_episode	75	day_period	55	touch_events	14	app_maps & navigation	23
Reading	app_comics	93	app_not-found	167	app_not-found	215	cellular_lte_min	252	app_adventure	21
	app_news & magazines	93	touch_events	122	wifi_std_rssi	109	app_tools	83	app_comics	16
	activity_onfoot	3518	day_period	121	wifi_max_rssi	77	location_altitude	76	location_altitude	6
	activity_walking	3497	activity_onfoot	1582	steps_detected	376	steps_detected	285	activity_walking	25
Walking	steps_detected	3374	activity_walking	1579	steps_counter	314	activity_walking	101	activity_onfoot	25
	app_health & fitness	502	steps_detected	1009	activity_walking	232	activity_onfoot	101	location_radius_of gyration	23
	day_period	233	app_health & fitness	33	app_health & fitness	931	app_health & fitness	1248	wifi_max_rssi	50
	notifications_posted	132	wifi_num_of_devices	32	proximity_min	52	noti_removed	72	proximity_std	41
Sport	steps_detected	283	wifi_min_rssi	23	day_period	48	day_period	34	wifi_mean-rssi	41
	activity_onfoot	267	activity_onfoot	1270	day_period	74	activity_walking	132	app_weather	86
	activity_walking	265	activity_walking	1269	user_presence_time	41	activity_onfoot	131	app_auto & vehicles	84
	steps_detected	265	steps_detected	504	screen_num_of_episodes	38	steps_detected	55	activity_walking	79

Europe, but not in Paraguay or Mongolia. This effect could be due to the unpopularity of streaming services classified as entertainment applications in the latter two countries, where participants might rely on alternatives. In fact, differences in using streaming services across countries have been studied in prior work, highlighting differences in usage percentages [298] and the relations to income level [371]. On the other hand, it could also be that students watch something on a medium that is not their smartphone. In fact, research shows that young adults aged 18-29 use more online media streaming services as compared to television in the USA [401]. However, whether similar percentages hold across different countries with contrasting cultures, income levels, and internet quality remains a question. While not conclusive, these could be the reason for entertainment apps not being indicative of “watching something” in Mongolia and Paraguay, which are the non-European countries in this study.

For some activities, the top three features are inherent to the nature of the activity. For example, “reading” in Italy has features corresponding to reading applications such as books, comics, newspapers, and magazines. Other countries do not show this. The same observation can be made for the “sports” activity: health and fitness apps are one of the determining features in European countries. This effect could correspond to participants tracking their workouts using a smartphone app.

The “walking” activity has almost the same features in all five countries: steps detected and an on-foot or walking activity detected by the Google Activity Recognition API. This homogeneity is due to the nature of the activity—walking is considered a simple activity. This is also why shopping has some of the same features as walking since participants also walk when they shop. To summarize, in most cases, each country has different defining features when looking at the same activity. For some activities, the features found are inherent to the activity and are usually app categories. Finally, it is worth mentioning that the period of the day is an important feature, which matches what has been observed in Figure 8.4 — all activities do not occur at the same frequency throughout the day.

Finally, it is worth noting that we could expect some of the highly informative features to change over time, with changes to technology use and habits of people, in different countries [569, 7]. For example, a reason for the lack of use of streaming services in certain countries is the lack of laws surrounding the usage of illegally downloaded content (e.g., Germany has strict laws about not using illegal downloads [443]). Changes in the laws of countries could change the behavior of young adults. Further, internet prices could also affect the use of streaming services. While bandwidth-based and cheap internet is common in developed countries, it is not the same in developing nations in Asia, Africa, and South America, where internet usage is expensive, hence demotivating streaming. In addition, income levels could influence captured features a lot. For example, with increasing income levels (usually happens when a country’s GDP changes), young adults may use more wearables for fitness tracking, leading to the usage of health and fitness apps on mobile phones. Another aspect that could affect the captured behaviors is the weather conditions. All five countries mentioned in this study go through different seasons, as all are somewhat far from the equator. Hence, we could expect changes in features in different seasons. More about this is discussed in the limitations section.

8.5 Machine Learning-based Inference: Experimental Setup, Models, and Performance Measures

This study aims to perform a multi-class inference of smartphone sensing data to predict what participants do at a particular time. The input space consists of the features in the tabular dataset previously

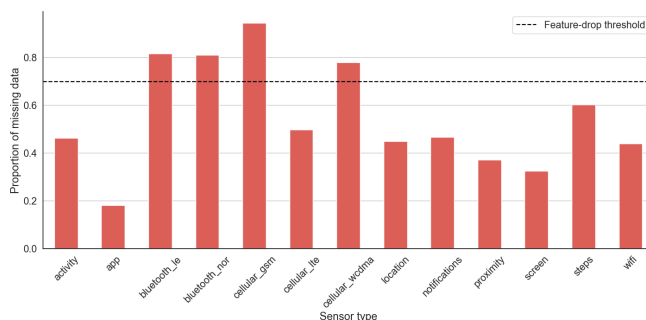


Figure 8.5: Proportion of missing data per sensor type.

mentioned. We study the three approaches to the problem as summarized in Figure 8.1, going from country-specific to multi-country.

8.5.1 Data Imputation

The first step in preparing the dataset for inference was data imputation. Missing data in the context of smartphone sensing can occur for multiple reasons [468, 30, 323]: the device being on low-consumption mode, the failure of a sensor, or insufficient permissions from the participants. In the dataset we used, we noticed that most sensors have some missing values (see Figure 8.5). For example, more than 90% of GSM cellular sensor values were unavailable, possibly due to devices being put in airplane mode, sensor failure, or the phone mostly operating with LTE signals. To deal with missing values, we decided to drop features from sensors that were missing more than 70% of their data (refer to the dotted line on Figure 8.5) similar to prior work [454]. For the remaining features, and each country individually, we used k-Nearest Neighbour (kNN) imputation [590] to infer missing information from neighboring samples².

8.5.2 Models and Performance Measures

To conduct all experiments, we used the scikit-learn [391] and Keras [91] frameworks, with Python. We first trained the following two baseline models: one that always predicts the most frequent label and another that randomly predicts targets by considering the class distribution. This will allow us to understand if the trained models perform better than a randomized guess. The experiments were carried out with the following model types: Random Forest Classifier [70] (RF), AdaBoost with Decision Tree Classifier [199], and Multi-Layer Perceptron neural networks (MLP) [546]³. The first two inherently leverage class imbalance, and RFs also facilitate the interpretability of results. Each experiment was carried out ten times to account for the effect of randomness. For each experimental setup, we reported the mean and standard deviation across the ten runs for the following metrics: F1 score [465], and the area under the Receiver Operating Characteristic curve (AUC) [464]. Even though we calculated

²We also tried mean imputation, user-based mean imputation, most frequent value imputation, last observation carried forward (LOCF) imputation, in addition to kNN. However, we obtained the best results for inferences with kNN. In addition, using kNN is common in studies that used passive sensing [587, 425, 596, 570]. Hence, we only reported results obtained with kNN.

³We initially tried out other model types such as Gradient Boosting and XGBoost in addition to the reported models. Results for these models were not reported, considering their performance. All these model types are commonly used in small mobile sensing datasets that are in tabular format [334, 55, 330]

the accuracies of models, and while the accuracy is easy to interpret, it might not present a realistic picture in an imbalanced data setting. Hence, we did not include it in the results. The weighted macro F1 score computes metrics for each class and averages them following their support, resulting in a metric that considers label imbalance. Moreover, it takes a significant hit if one of the classes has a lot of false positives. A low F1 score could imply that the classifier has difficulty with rare target classes. The AUC score measures how well the model can distinguish each activity. It can be understood as an average of F1 scores at different thresholds. We also used a weighted macro version to account for label imbalance.

Next, we examine results for country-specific, country-agnostic, and multi-country approaches [247]. Finally, for all three approaches, we examine population-level, and hybrid models that correspond to no and partial personalization, respectively, similar to [324, 323, 285] (training and testing splits were always done with 70:30 ratio):

- **Population-Level** model, also known as leave-k-participants-out in country-specific and multi-country approaches, and leave-k-countries-out in country-agnostic approach: the set of participants present in the training set ($\approx 70\%$) and the testing set ($\approx 30\%$) are disjoint. The splitting was done in a stratified manner, meaning each split was made by preserving the percentage of samples for each class. This represents the case where the model was trained on a subset of the population, and a new set of participants joined a system that runs the model and started using it.

- In the country-specific approach, this means that data from disjoint participants are in training and testing splits, and everyone is from the same country. E.g., trained with a set of participants in Italy and tested with another set of participants in Italy who were not in the training set.
- In the country-agnostic approach, this means the training set is from one (Phase I) or four (Phase II) countries, and the testing set is from a country not seen in training. E.g., For Phase I — trained with a set of participants in Italy and tested with a set of participants in Mongolia; Phase II — trained with a set of participants in Italy, Denmark, UK, and Mongolia, and tested with a set of participants in Paraguay.
- In the multi-country approach, this means a disjoint set of participants in training and testing without considering country information. This is the typical way of training models even when data are collected from multiple countries [468]. E.g., trained with a set of participants from all five countries and tested with a set of participants in all five countries who were not in the training set.

- **Hybrid** model, also known as the leave-k-samples-out: the sets of participants in the training and testing splits are not disjoint. Part of the data of some participants present in the testing set ($\approx 70\%$) was used in training the models. Testing is done with the rest of the data from the participants ($\approx 30\%$). This represents the case where the model was trained on the population, and the same participants whose data were used in training continue to use the model. Hence, models are partially personalized.

- In the country-specific setting, this means that some data from participants within a country in the testing set can also be in the training set. This represents a scenario where personalization is examined within the country. E.g., trained with a set of participants in Italy and tested with another set of participants in Italy, whose data (70%) were also used in the training set. The rest of the data (30%) were used in the testing set.
- In the country-agnostic setting, this means the training set is from one/more countries, and the testing set is from another country, where a percentage of their past data (70%) was also included

Table 8.3: Mean (\bar{S}) and Standard Deviation (S_σ) of inference F1-scores, and AUC scores computed from ten iterations using three different models (and two baselines) for each country separately. Results are presented as $\bar{S}(S_\sigma)$, where S is any of the two metrics.

	Baseline I		Baseline II		Random Forest		AdaBoost		MLP	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
	Population-Level									
Italy	0.17 (0.000)	0.50 (0.000)	0.19 (0.001)	0.50 (0.001)	0.41 (0.001)	0.71 (0.001)	0.39 (0.000)	0.71 (0.000)	0.38 (0.002)	0.68 (0.002)
Mongolia	0.26 (0.000)	0.50 (0.000)	0.23 (0.001)	0.50 (0.001)	0.33 (0.002)	0.62 (0.001)	0.33 (0.000)	0.63 (0.000)	0.34 (0.003)	0.61 (0.004)
UK	0.17 (0.000)	0.50 (0.000)	0.18 (0.002)	0.50 (0.001)	0.32 (0.004)	0.63 (0.003)	0.31 (0.000)	0.59 (0.000)	0.22 (0.006)	0.56 (0.003)
Denmark	0.25 (0.000)	0.50 (0.000)	0.24 (0.006)	0.49 (0.003)	0.32 (0.008)	0.61 (0.006)	0.34 (0.000)	0.57 (0.000)	0.25 (0.008)	0.57 (0.006)
Paraguay	0.19 (0.000)	0.50 (0.000)	0.19 (0.006)	0.49 (0.002)	0.30 (0.004)	0.59 (0.003)	0.28 (0.000)	0.56 (0.000)	0.31 (0.009)	0.58 (0.004)
	Hybrid									
	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
Italy	0.17 (0.000)	0.50 (0.000)	0.19 (0.001)	0.50 (0.001)	0.63 (0.001)	0.87 (0.001)	0.40 (0.000)	0.73 (0.000)	0.51 (0.002)	0.81 (0.000)
Mongolia	0.26 (0.000)	0.50 (0.000)	0.23 (0.002)	0.50 (0.001)	0.51 (0.001)	0.79 (0.001)	0.34 (0.000)	0.66 (0.000)	0.45 (0.002)	0.75 (0.002)
UK	0.17 (0.000)	0.50 (0.000)	0.19 (0.003)	0.50 (0.001)	0.66 (0.001)	0.88 (0.006)	0.34 (0.000)	0.68 (0.000)	0.58 (0.003)	0.83 (0.002)
Denmark	0.25 (0.000)	0.50 (0.000)	0.24 (0.003)	0.50 (0.002)	0.69 (0.002)	0.89 (0.001)	0.41 (0.000)	0.66 (0.000)	0.67 (0.002)	0.87 (0.002)
Paraguay	0.18 (0.000)	0.50 (0.000)	0.19 (0.002)	0.49 (0.003)	0.61 (0.003)	0.84 (0.001)	0.30 (0.000)	0.61 (0.000)	0.58 (0.002)	0.79 (0.001)

in the training. This represents a scenario where personalization is examined when deployed to a new country. E.g., Phase I — trained with a set of participants in Italy and tested with a set of participants in Mongolia, whose data (70%) were also used in the training set. Rest of the data (30%) were used in the testing set; Phase II — trained with a set of participants in Italy, Denmark, UK, Mongolia, and tested with a set of participants in Paraguay, whose data (70%) were also used in the training set. The rest of the data (30%) were used in the testing set.

- In the multi-country setting, this means that training and testing participants are not disjoint, and country information is not considered. This is the typical way of partially personalizing models even when data are collected from multiple countries. E.g., trained with a set of participants from all five countries and tested with a set of participants in all five countries, whose data (70%) were also used in the training set. The rest of the data (30%) were used in the testing set.

8.6 Inference Results

In this section, we present the results of the experiments. First, we discuss results from the country-specific and multi-country approaches, shedding light on **RQ2**. Then, the country-agnostic approach is discussed by providing answers to **RQ3** on model generalization.

8.6.1 Country-Specific and Multi-Country Approaches (RQ2)

Country-Specific Approach. We consider this approach to be the base setting that does leverage country-level diversity in building separate models—each country has its own model independently from others. Table 8.3 summarizes the results of experiments following the country-specific approach. In the population-level setting, the three models perform more or less similarly, but the RFs are generally better based on F1 and AUC scores. In the case of the hybrid models, RFs performed the best across the five countries, with AUC scores in the range of 0.79-0.89, where the lowest was for Mongolia, and the highest was for Denmark. Compared to population-level models, we can notice a substantial bump in performance in the hybrid models, showing the effect of personalization within countries. These results suggest that random forest models applied to a partially personalized setting can recognize complex daily activities from passive sensing data with a good performance. Given this conclusion, even though we got results for all model types for subsequent sections, we will present results only using random forest models.

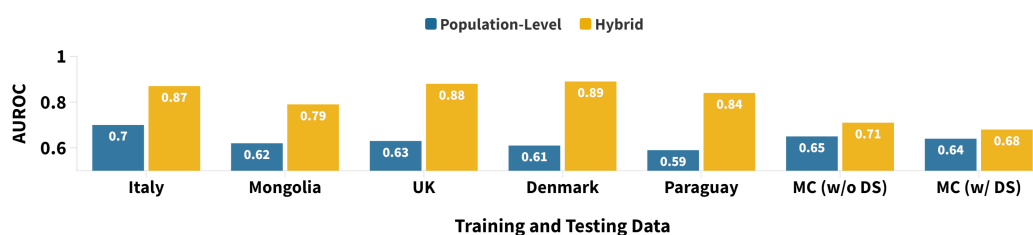


Figure 8.6: Mean AUC score comparison for country-specific and multi-country approaches with population-level and hybrid models. MC: Multi-Country; w/o DS: without downsampling; w/ DS: with downsampling.

Multi-Country Approach. This approach aims at building a generic multi-country or one-size-fits-all model with the expectation that it would capture the diversity of all countries. All five countries are present in both the training and the testing set. We, therefore, consider all participants of the dataset, regardless of their country, similar to an experiment where country-level diversity is ignored. Hence, we can examine population-level and hybrid models for a multi-country approach in this context. Further, models were evaluated with a dataset with an imbalanced representation from five countries (multi-country w/o downsampling — MC w/o DS) and a balanced representation from five countries by randomly downsampling from countries with more data to make it equal to the country with the least number of self-reports (i.e., Paraguay) (multi-country w/ downsampling — MC w/ DS). The results are shown in Figure 8.6 in comparison to country-specific results. MC w/o DS had an AUC of 0.71 while MC w/ DS had an AUC of 0.68, indicating that training on the original data distribution performed better. The reason could in fact be that, more data led to better performance. The expectation of training with downsampled data was to give equal emphasis to each country, expecting that the model would perform well to all countries. However, the result indicates that it is not the case.

These results shed light on our **RQ2**: learning a multi-country model for complex activity recognition solely using passive smartphone sensing data is difficult (AUC: 0.709 with hybrid models). It does not yield better performance than the country-specific approach (AUCs of the range 0.791-0.894). This may stem from the data's imbalance between countries and classes or the context in which the dataset was collected. Another primary reason for this could be behavioral differences in data highlighted in Table 8.2, making it difficult for a model to learn the representation when the diversity of data is unknown. Distributional shifts⁴ across datasets from different countries could be the reason for this. When sensor feature and ground truth distributions (we discussed ground truth distributions in Section 8.4) are different across countries, it could lead to an averaging effect, which would lead to worse-performing models than models for each country. Moreover, it is worth noting that there are not a lot of studies that trained country-specific and multi-country models for performance comparison [402]. In one of the only other studies that we found [247], personality trait inference performance using smartphone sensor data was better when using country-specific models, similar to what we found for complex daily activity inference. Finally, from a human-centered perspective, recruiting participants to collect smartphone sensing data to build machine learning models means that—rather than targeting large samples from a single country, recruiting a reasonable number of participants from diverse countries could help deploy better-performing models to multiple countries.

⁴<https://huyenchip.com/2022/02/07/data-distribution-shifts-and-monitoring.html#data-shifts>

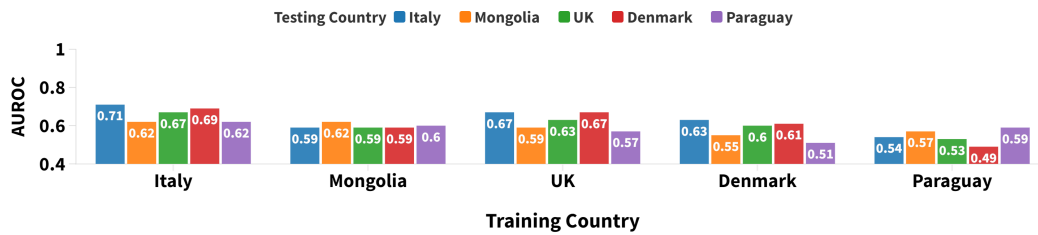


Figure 8.7: Mean AUC scores obtained in the country-agnostic approach with population-level models.

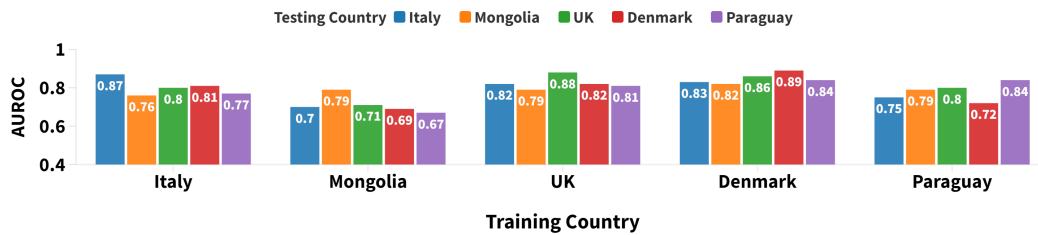


Figure 8.8: Mean AUC scores obtained in the country-agnostic approach with hybrid models.

8.6.2 Generalization Issues with Country-Agnostic and Country-Specific Models (RQ3)

We examined this research question with two phases. During the first phase, to evaluate the extent to which country-specific models generalize to new countries, we tested models trained with a single country’s data in the other four countries separately. In the second phase, to evaluate the extent to which a model trained with four countries generalized to the remaining country, we trained with different combinations of countries and tested on the remaining country.

- Phase I:** Figure 8.7 summarizes results for population-level models and Figure 8.8 summarizes results for hybrid models. To allow easy comparison, in both figures, the result mentioned as the performance of a country, when tested on the same country is the result from Table 8.3. For instance, at the population-level, Italy had an AUC of 0.71 according to Table 8.3, and this is marked in Figure 8.7 where both the Training and Testing country is Italy. Population-level results suggest that the country-agnostic approach tends to perform better in countries geographically close to the country where the model was originally trained. For example, the Italy model had an AUC of 0.71 for the Italian populations in a population-level setting and performed better in Denmark (AUC: 0.69) and the UK (AUC: 0.67) than it did in Mongolia (AUC: 0.62) or Paraguay (AUC: 0.62). Similar results can also be observed for hybrid models, where the Italian model performed better in Denmark and UK. This observation suggests that college students from countries within the same geographic region (Europe) could have behaviors that translate to similar smartphone usage and contexts during periods of doing similar activities. This is consistent with the observations made in the descriptive analysis above, where the countries that deviate from the general trends are usually those outside Europe. In summary, even after using the same experimental protocol when collecting mobile sensing data, we could still observe a distribution shift of data by the performance of models across geographically distant countries.

- Phase II:** The second phase looked into extending the work done in phase I. Instead of testing a country-specific model in a new country, we were interested in testing a model already exposed to

diverse data (e.g., from four countries) in a new country. We present results for random forest models (because they performed the best across experiments) where the training set consisted of data from four countries, and the testing set had data from the fifth. As suggested in prior studies [247], each country contributed equally to the training set in terms of data volume, which means we had to downsample the data from each country to a common count (which was equal to the minimum number of data points available from one country). Table 8.4 presents the results for experiments of the second phase. Similar to previous cases, we observed an increase in performance from population-level to hybrid models. More generally, and by looking at the F1 and AUC scores, the performance of the hybrid models in the country-agnostic approach is lower than that of the same model in the country-specific approach. This is somewhat expected since including data from other distributions (i.e., other countries) in the training set increases the data's variance and makes it more difficult to represent all distributions accurately. This drop in performance could also be due to the downsampling. For instance, in a model where we train with four countries, including Italy and Paraguay, Italy represents the largest portion of the dataset compared to Paraguay, which is the smallest. When reducing the number of samples in each country to that of Paraguay, a lot of information is lost in the other countries: the larger the original dataset is, the larger the loss gets. This could explain the low performance of country-agnostic models in Italy and Mongolia, especially in the hybrid setting.

In addition, when comparing different modeling approaches, the results with Multi-Country w/o Downsampling are similar to those found in Phase II (hybrid) of the country-agnostic approach, which was expected since the training sets are similar. However, the bump in performance when going from population-level to hybrid is less noticeable here compared to previous cases. Furthermore, MC w/ DS performs worse than the previous approach, with an AUC of 0.68 compared to 0.71. This could be because we lose much data from many countries due to downsampling, reducing models' representational ability. To summarize, a hybrid model in a country-agnostic approach can not predict complex activities better than its country-specific counterpart. Furthermore, while more data often means better performances, this does not apply when the data follow different distributions, one per country in this case. This suggests that each country has specific characteristics that make learning one representation difficult.

These results shed light on our **RQ3**: complex activity recognition models trained in specific countries often generalize reasonably to other countries (especially with hybrid models). However, the performance is not comparable to the country-specific approach, suggesting that there is still a distributional shift between countries. In fact, in Section 8.4, we discussed how the labels used in the inference (i.e., shown in Figure 8.4—complex daily activities such as resting, studying, reading, etc.) had different distributions across the five countries. Further, the extent of the generalization often depended on whether countries are geographically closer (i.e., within Europe) or not. This result is in line with findings from previous studies [402, 247] that highlighted the effect of geographic dimensions (i.e., country of data collection) on mobile sensing model performance. For example, [247] found that country-specific models that used mobile sensing data as input, could perform well for the inference of three personality traits—Extraversion, Agreeableness, and Conscientiousness. Furthermore, we would also like to highlight that the issue regarding distributional shifts and generalization is an open problem in multimodal mobile sensing, as highlighted by two recent studies that examined similar datasets collected from the same country in different time periods [569, 7]. This is possibly due to behavioral changes over time leading to different distributions in sensor data and ground truth. Our results go beyond this and show that even if data is collected within the same time period and with the same protocol, distributional shifts could still occur due to country differences.

Table 8.4: Mean (\bar{S}) and Standard Deviation (S_σ) of F1-scores and AUC scores were obtained by testing each Country-Agnostic model (trained in four countries) on data from a new country. Results are presented as $\bar{S}(S_\sigma)$, where S is any of the two metrics.

<i>Test Country</i>	Population-Level		Hybrid	
	<i>F1</i>	<i>AUC</i>	<i>F1</i>	<i>AUC</i>
Italy	0.33 (0.005)	0.65 (0.006)	0.37 (0.004)	0.71 (0.002)
Mongolia	0.30 (0.011)	0.60 (0.004)	0.37 (0.006)	0.67 (0.003)
UK	0.29 (0.004)	0.63 (0.005)	0.47 (0.004)	0.78 (0.002)
Denmark	0.38 (0.006)	0.65 (0.006)	0.63 (0.008)	0.86 (0.004)
Paraguay	0.28 (0.005)	0.59 (0.006)	0.55 (0.006)	0.80 (0.008)

8.7 Discussion

8.7.1 Summary of Results

In this chapter, we examined a multi-country dataset with which we attempted to develop classification models to infer complex daily activities from passive sensing data. Our primary goal was to seek whether reasonably performing complex daily recognition models could be trained using multimodal sensor data from smartphones. Then, our goal was to identify differences among countries visible through smartphone usage and to leverage these differences to decide whether it makes sense to build country-specific or generic multi-country models and whether models generalize well. We believe these findings are of high importance to the community when designing and deploying sensing and machine learning-based apps and systems, in geographically diverse settings. The main findings for the three research questions can be summarized as follows:

- **RQ1:** Different features in each country can characterize an activity. Its distribution throughout the day also varies between countries and seems to be affected by local habits, customs, and technology use. This finding points towards a set of biases that could get transmitted to data if proper care is not taken during the study design and data collection phase of user studies involving smartphones and human participants. In Section 8.7.2, we discuss this in more detail under a set of biases: construct bias [201], sample bias [325], device-type bias [59], and bias from user practices [532].
- **RQ2:** It is feasible to train reasonably performant classification models with the country-specific approach to predict 12 complex daily activities using passive smartphone sensing data. Furthermore, personalization within countries increases performance (AUCs in the range 0.79-0.89). Hence, the country-specific approach outperforms the multi-country approach, which only yields an AUC of 0.71 with hybrid models. However, building multi-country models solely from sensing features is a non-trivial task that might require more effort in the aspects of data balance and feature selection. In the light of prior work that attempted activity recognition, our results show that highly sedentary emerging lifestyles of the post-pandemic world can be captured with country-specific partially personalized machine learning models. In addition, we also show that multimodal smartphone sensors could be used to recognize complex daily activities that go beyond binary inferences to a 12-class inference that captures the everyday life of people. Extending this line of thought, we discuss why in-the-wild studies are important to capture complex emerging lifestyles, in Section 8.7.2. We also discuss how these complex daily activities could be crucial in designing novel context-aware mobile applications, in Section 8.7.2.

- **RQ3:** Under the country-agnostic approach, we found that models generalize reasonably to new countries. However, unsurprisingly, the performance is not as high as when the model was tested in the same country where it was trained. Interestingly, even with language and certain behavior-related differences, models trained in European countries performed better in other European countries than in Paraguay or Mongolia. This problem that we identified in the context of complex daily activity recognition in multiple countries, broadly falls under the bracket of distributional shifts, a topic under explored in mobile sensing literature. We elaborate more on this under Section 8.7.2.

8.7.2 Implications

Stemming from the answers we found to research questions, this work has implications that are aligned to both theoretical and practical aspects.

Accounting for Country Biases in Study Design (RQ1)

Studies using sensing data drawn from geographically diverse samples (i.e., different countries) should account for and understand the *sources of biases* that can occur at different stages of the study. Our study, and also previous studies on human behavior, sociology, and psychology, allow an understanding of these aspects in detail. For example, the following taxonomy can be used to characterize such biases [402]. The *(i) construct bias* occurs when the target is expressed differently across countries, depending on countries' norms or environmental factors [201]. For example, the activity "walking" in one country where physical exercise and fitness are not taken seriously could be labeled as "walking", whereas in a country where it's more of an activity done for fitness by many people, it could be labeled as a "sport" as well. Hence, some behaviors can be specific to a particular environment or group of people. The *(ii) sample bias* concerns the comparability of diverse samples that can be impacted by the recruitment process in each country [325]. For example, if the samples across the country vary in age, profession, or in gender balance, sensing data would likely not have similar distributions across countries. The *(iii) device-type bias* is due to the differences in the devices used by participants in different countries, and environmental factors affecting sensor measurements [59]. Devices worldwide are not equipped with the same software and hardware, and similar sensors can differ in accuracy and precision (e.g., Apple devices are more prominent in developed countries, whereas other android phones are common in others). Finally, the *(iv) bias from user practices* arises when participants from different countries use their mobile phones differently [532]. For example, some people are used to carrying their phones in a pocket and others in their bags, which could distort measurements. Others might occasionally disable some sensors or stop the sensing altogether to save battery or mobile data (i.e., especially in countries where unlimited mobile data plans are not standard). There could be differences in the reason and motivations why people use certain apps in different countries [286]. Phan et al. [402] proposed mitigation strategies to reduce biases and encourage fair approaches to diversity-aware research. During the study planning, they suggest learning about potential cross-country differences, gaining knowledge about relevant countries and environmental factors with the help of local informants, and ensuring that targets exist in every country and are comparable across countries. During the study implementation, they suggest making the recruitment as inclusive as possible such that each sampled country is representative of a given target and inciting participants to keep a consistent behavior.

Activities Captured in In-the-Wild Studies (RQ2)

In terms of theoretical implications, it is worth highlighting that the set of activities that we considered are complex behaviors that can not be typically captured during in-lab studies. Fine-grained sensing-based activity recognition studies help increase performance on simple activities (e.g., walking, running, sitting, climbing stairs, etc. — that have a repetitive nature in sensor data) that can be captured in in-lab settings. However, we believe the only way to build sensing-based machine learning models to capture complex daily behaviors is to conduct in-the-wild studies. For example, an activity like studying, attending classes, working, reading, or shopping is hard to replicate in an in-lab setting, similar to how it occurs in real life. Further, while simple activities might not have led to differences in model performances across countries, complex daily activities tightly bound with cultural, country-level, or geographic norms lead to differences in behaviors, leading to differences in the sensed data. In this context, prior work in the domain has not focused on this aspect enough, in our view. Even more so, we believe that studies must capture data from diverse communities to build models that work for all intended users. While this is a challenging task, it is much needed for the field of research to mature for more real-life use cases.

Novel Applications of Context-Aware Mobile Apps (RQ2)

In terms of practical implications, our findings point towards adding context awareness to mobile applications beyond current norms. Current mobile applications provide context-aware services, interventions, and notifications based on location and simple activity sensing [331, 325]. However, a range of potential applications that go beyond the current offering could become feasible with complex daily activity recognition. For example, previously, a smartphone would only know that a user was sitting in a particular place. With complex activity recognition, it would know that a user is studying, eating, watching something, attending a lecture, or reading, which all entail sitting – however, a user needs contrasting context-awareness-based experiences from the phone in these different settings. For example, if the student is reading, studying, or attending a lecture, automatically putting the phone in the silent or do-not-disturb mode might make sense if users prefer it, even though, in many cases, people forget to. In addition, recognizing that the user is eating could prompt interventions or feedback regarding eating behavior, health, and well-being. Further, after knowing that a user is walking, identifying that the user is shopping would allow intelligent advertising strategies to be employed to provide users with suggestions when they are in the right mood for buying. In addition to recognizing an activity, our model, in a way, recognizes users' intent (i.e., they are in a mentality where they do not want to get disturbed, in a mentality where they are looking to buy items, etc.). Therefore, the model could add benefits on top of location-based services – for instance, a person could be in a shopping area when passing by the area by chance or with the intent of shopping. Hence, knowing that they are shopping instead of being near a shopping could add a layer of intelligence to mobile apps. Hence, similarly, complex daily activity recognition could offer diverse use cases to build mobile applications around in the future.

Domain Adaptation for Multimodal Mobile Sensing (RQ3)

Another theoretical implication can be described in a machine learning sense. In this study, we elaborated on the challenge of generalization and the occurrence of distributional shifts in a smartphone sensor dataset collected through the same protocol in different countries. We described how this shift affects model performance, specifically for complex daily activity recognition with multimodal

sensors. Although biases, distributional shifts, and model generalization have been widely studied in other domains such as natural language processing [144], speech [496], and computer vision [301], smartphone sensing studies have yet to receive sufficient attention [180]. We demonstrated that model personalization (hybrid setting) could reduce distributional shifts to a certain extent. In a way, according to transfer learning-related terms, this approach is similar to fine-tuning an already trained model for a specific user to achieve model personalization [88]. Such strategies for personalization have been used in prior work [330]. However, recent research in domain adaptation has shown limited progress in addressing challenges in the field of mobile sensing, particularly with regard to time series data [562]. The diversity of wearable device positioning poses a persistent issue in human activity recognition, which affects the performance of recognition models [85, 315]. Wilson et al. [562] conducted a study of domain adaptation in datasets captured from individuals of different age groups, yet the findings are limited to simpler time series accelerometer data. Other studies mention that the current lack of solutions for domain adaptation and generalization in multi-modal sensing data originating from smartphones and wearables presents an opportunity for exploration [7, 569]. In light of the results from our study, we add to the literature by confirming the idea that domain adaptation techniques should be explored, more specifically, for multimodal smartphone sensor data collected from diverse countries. In addition, even on a fundamental level, approaches that allow quantifying cross-dataset distributional differences for multimodal sensing features and target labels (e.g., activity, mood, social context, etc.) separately, are lacking in the domain. Research on such aspects could allow us to better understand distributional shifts in sensor data, to better counter it with domain adaptation techniques in multimodal settings.

8.7.3 Limitations

While the dataset covers five different countries, it only spans three continents. Therefore, students' behavior in other continents, such as North America or Oceania, could differ from what we have already encountered. In addition, even though we found geographically closer countries performing well in Europe, such findings need to be confirmed for other regions where geographically closer countries could have contrasting behaviors and norms (e.g., India and China). Furthermore, the weather conditions in different countries during the time period of data collection could be slightly different. All five countries mentioned in this study go through different seasons, as all are somewhat far from the equator. Hence, we could expect changes in features in different seasons. However, in practical terms, collecting data in similar weather conditions is not feasible when mobile sensing apps need to be deployed to collect datasets of the Tera Byte scale from participants. That is why such multicultural in-the-wild studies spanning more extended time periods have rarely been conducted in the past. So, as one of the first studies in the domain to discuss issues of generalization in mobile sensing for activity recognition-related tasks, we believe this chapter provides a reasonable starting point for future studies to expand upon.

When aggregating sensor data around self-reports, the data corresponding to the moment the participant was filling out the self-report is considered a part of the activity he/she was doing at the time. This noise could alter the recognition task if the window's size is small enough. However, even though this could affect results if we intended to increase model performance in a fine-grained sensing task, we do not believe this noise affects the results significantly regarding our findings on diversity awareness. In addition, it is worth noting that the way we model our approach with a tabular dataset is similar to prior ubicomp/mobile sensing studies done in the wild [325] because we do not have continuous ground truth labels. Hence, it restrains us from modeling the task as a time-series problem, which is how a

majority of activity recognition studies [492] with continuous accurate ground truth measurements follow. So, the results should be interpreted with the study's exploratory nature in mind.

Further, it is worth noting that we could expect some of the highly informative features used in models to change over time, with changes to technology use and habits of people, in different countries [569, 7]. For example, a reason for the lack of use of streaming services in certain countries (discussed in Section 8.4) is the lack of laws surrounding the usage of illegally downloaded content (e.g., Germany has strict laws about not using illegal downloads [443]). Changes in the laws of countries could change the behavior of young adults. Further, internet prices could also affect the use of streaming services. While bandwidth-based and cheap internet is common in developed countries, it is not the same in developing nations in Asia, Africa, and South America, where internet usage is expensive, hence demotivating streaming. In addition, income levels could influence captured features a lot. For example, with increasing income levels (usually happens when a country's GDP changes), young adults may use more wearables for fitness tracking, leading to the usage of health and fitness apps on mobile phones.

The amount of data for each country is highly imbalanced. For a fair representation of each country, having the same number of participants and self-reports per country would ensure that a classifier learns to distinguish classes from each country equally. However, Italy and Mongolia are dominant in the current state of the dataset. If not done carefully, down-sampling would result in a loss of expressiveness and variance, making it difficult to discern different classes in a multi-country approach. Another imbalance is found among class labels, where activities such as sleeping or studying are more frequent than others. However, this does make sense since we do not expect all activities to appear at the same frequency in a participant's day or week. Further, we reported F1 and AUC scores that are preferred in such imbalanced settings.

Finally, the dataset was collected in November 2020, during the Covid-19 pandemic, when most students stayed home due to work/study-from-home restrictions. This explains why most of the relevant features found in the statistical analysis are screen events and app events. While some relevant features are relative to proximity and WiFi sensors, there are very few regarding activity and location unless the activity corresponds to physical activities. This is probably an effect of a context where movements were highly discouraged. From another perspective, the behavior of college students from all countries during this time period reflects remote work or study arrangements. We could expect these practices to continue for years as more universities and companies adopt remote work/study culture. Hence, while many prior studies in ubicomp used phone usage features and sensing features for activity/behavior/psychological trait inference tasks, our findings indicate that phone usage features could be even more critical in the future with remote study/work settings due to sedentary behavior, that would limit the informativeness of sensors such as location and inertial sensors.

8.7.4 Future Work

The study's population for the dataset collection consisted of students. Therefore, it might be worth exploring how people from different age groups use their smartphones and how their daily behavior is expressed through that usage. In addition to visible diversity, it is known that deep diversity attributes (innate to humans and not visible) such as personality (captured with Big Five Inventory [138]), values (captured with basic values survey [187] and human values survey [461, 462]), and intelligence (captured with multiple intelligence scale [512]) could also affect smartphone sensor data and activities performed by people [458, 247]. Hence, investigating how such diversity attributes could affect smartphone-based inference models on complex activities, and other target variables, is worth investigating. Further,

future work could investigate how the classification performance is affected when excluding the sensing data corresponding to the time taken to fill the self-report about activities by participants. Finally, domain adaptation for multi-modal smartphone sensor data across time and countries remains an important problem worth investigating in future work.

8.8 Conclusion

In this chapter, we examined the daily behavior of 637 students in Italy, Mongolia, the United Kingdom, Denmark, and Paraguay using over 216K self-reports and passive sensing data collected from their smartphones. The main goal of this chapter was to, first examine whether multimodal smartphone sensor data could be used to infer complex daily activities, which in turn would be useful for context-aware applications. Then, to examine whether models generalize well to different countries. We have a few primary findings: *(i)* While each country has its day distribution of activities, we can observe similarities between the geographically closer countries in Europe. Moreover, features such as the time of the day or the week, screen events, and app usage events are indicative of most daily activities; *(ii)* 12 complex daily activities can be recognized in a country-specific and personalized setting, using passive sensing features with reasonable performance. However, extending this to a multi-country model does not perform well, compared to the country-specific setting; and *(iii)* Models do not generalize well to other countries (at least compared to within-country performance), and especially to geographically distant ones. More studies are needed along these lines regarding complex daily activity recognition and also other target variables (e.g., mood, stress, fatigue, eating behavior, drinking behavior, social context inference, etc.), to confirm the findings. Hence, we believe research around geographic diversity awareness is fundamental for advancing mobile sensing and human behavior understanding for more real-world utility across diverse countries. From a study design sense, we advocate the idea of collecting data from diverse regions and populations to build better-represented machine learning models. From a machine learning sense, we advocate the idea of developing domain adaptation techniques to better handle multimodal mobile sensing data collected from diverse countries.

9 Unsupervised Domain Adaptation for Multimodal Mobile Sensing with Multi-Branch Adversarial Training

Over the years, multimodal mobile sensing has been used extensively for inferences regarding health and well-being, behavior, and context. However, a significant challenge hindering the widespread deployment of such models in real-world scenarios is the issue of distribution shift. This is the phenomenon where the distribution of data in the training set differs from the distribution of data in the real world—the deployment environment. Even though explored extensively in computer vision and natural language processing, and while prior research in mobile sensing briefly addresses this concern, current work primarily focuses on models dealing with a single modality of data, such as audio or accelerometer readings. Consequently, there exists a lack of research on unsupervised domain adaptation when dealing with multimodal sensor data. To address this gap, we propose a novel technique **M3BAT**, unsupervised domain adaptation for **m**ultimodal **m**obile sensing with **m**ulti-**b**ran**b** adversarial training, to account for the multimodality of sensor data during domain adaptation. Through extensive experiments conducted on two multimodal mobile sensing datasets, three inference tasks, and 14 source-target domain pairs, including both regression and classification, we demonstrate that our approach performs effectively on unseen domains. Compared to directly deploying a model trained in the source domain to the target domain, the model shows performance increases up to 12% AUC (area under the receiver operating characteristics curves) on classification tasks, and up to 0.13 MAE (mean absolute error) on regression tasks. The material of this chapter is under review.

9.1 Introduction

In recent years, the prevalence of mobile and wearable devices equipped with multimodal sensors has increased significantly, offering a wide range of applications in health, well-being, context awareness, and user experience [268, 325]. These sensors can capture diverse data, including accelerometers, gyroscopes, photoplethysmography (PPG) readings, and location, as well as device usage data like application usage and typing and touch events. This wealth of data presents exciting opportunities for understanding human behavior [24], physiological responses [168], and contextual information [583] in an unobtrusive manner. Some examples include activity recognition [33, 24, 99], stress detection [318, 202, 340], mood inference [285, 468, 324], eating and drinking behavior understanding [327, 55, 454, 330], and social context recognition [237, 328]. However, despite the growing interest in utilizing multimodal sensor data, several challenges must be addressed to harness their potential fully, in deployment settings. One such under-explored challenges is the generalization of models across different users, populations, and environments [569, 324, 7]. As each individual exhibits unique

behavioral patterns and physiological responses, building models that are robust and adaptable across diverse populations poses a challenge [358]. Additionally, variations in the context of data collection can significantly impact sensor readings and behavior patterns (e.g., training a model in Italy and expecting it to work for people in India) [24, 324]. Achieving generalization is challenging due to distribution shifts in the data.

The data collected from various sensors in different environments may not align perfectly, resulting in a distribution shift between the source dataset (data that the model is trained on) and the target dataset (data that the model would encounter in deployment) [85, 526]. These distribution shifts can have a negative impact on model performance when applied in new and unseen contexts. Addressing distribution shifts require the utilization of transfer learning techniques, including robust domain adaptation approaches [162]. Despite numerous studies exploring the applications of multimodal sensors in mobile and wearable devices, discussions around the challenges of generalization and distributional shifts have been relatively limited [569, 324]. This is in contrast to other domains, such as computer vision, natural language processing, and speech processing, where significant progress has been made in understanding and mitigating domain shifts [595]. However, prior studies have emphasized that blindly adapting techniques from other domains to mobile sensing datasets is not trivial and needs deeper investigation because of the differences in the way data are collected, processed, and made sense of [85, 566, 569]. Therefore, more investigations are needed in multimodal sensing settings to overcome challenges regarding distribution shifts.

In mobile sensing settings, training models often rely on large-scale datasets collected from multiple users. In deployment, models need to personalize for better performance, and having ground truth labels from users is a primary way to do this. However, obtaining labeled ground truth from users poses challenges due to the sparse nature of data collection and difficulties in acquiring accurate and reliable self-reports [568]. Consequently, the lack of labeled data impedes the personalization of models for individual users, making it difficult to cater to their unique characteristics and preferences. Therefore, the crucial step of adapting models to target populations (i.e., genders, age groups, countries, sub-populations, etc.) becomes essential even before personalization [324, 24]. By adapting the models to the target population, we can ensure their effectiveness in diverse contexts, providing a strong foundation for subsequent personalization efforts. Unsupervised domain adaptation (UDA) [161] techniques play a vital role in bridging the gap between different domains, rendering the models more versatile and adaptable to various users and environments. Considering the challenge of obtaining ground truth from users, UDA emerges as a solution to enhance the performance and generalizability of models in the realm of mobile and wearable sensing as well. However, even though UDA has been explored in very few prior studies in mobile sensing [85, 315, 324], how such techniques perform when multimodal data with varying degrees of distribution shift are present, has rarely been explored.

Considering these aspects, in this chapter, we first conducted a statistical analysis of datasets to understand the dynamics of distribution shifts across source-target domain pairs and different sensing modalities. Then, we evaluate unsupervised domain adaptation with domain adversarial training (DANN [162, 85]) on two different multimodal mobile and wearable sensing datasets across both regression and classification tasks. Then, we propose a novel model architecture for Multimodal Mobile Sensing data called Multi-Branch domain Adversarial Training (**M3BAT**), showing improved performance over baselines across a majority of inference tasks. In doing this, we answer the following research questions:

RQ1: What dynamics regarding distribution shift can be observed by conducting a statistical analysis on multimodal sensing datasets?

RQ2: Does DANN on multimodal sensing datasets lead to improved UDA performance? How does it compare to transfer learning-based fine-tuning when labels are available in the target domain?

RQ3: Does having multiple branches for different feature sets (based on modality or distribution shift-based feature groups) lead to improved UDA performance?

By addressing the above research questions, this chapter provides the following contributions:

Contribution 1: We conducted an analysis of two multimodal sensing datasets, namely MUL and WEEE. These datasets provide valuable insights into distribution shifts across various dimensions—MUL explores distribution shifts across different countries with modalities such as wifi, steps, proximity, location, screen events, app usage, activity, etc., while WEEE captures shifts across devices worn on distinct body positions with accelerometer, photoplethysmography (PPG), and gyroscope data. Our approach involved calculating Cohen's-d values for individual features and then aggregating them to discern patterns at the modality and feature set level. This analysis allowed us to pinpoint the modalities that exhibited high distribution shifts across various source and target domain pairs. For instance, activity and screen event data demonstrated minimal difference between Italy and India, while wifi and step count features displayed substantial dissimilarity, attributable to low and high shifts, respectively. These trends contrasted across source-target pairs, underscoring the importance of a multimodality-aware architecture that accounts for individual modality and feature set level shifts during the domain adaptation process.

Contribution 2: The datasets employed in our study provide a platform for exploring diverse inferences, encompassing mood, social context, and energy expenditure estimation via classification and regression tasks. In order to comprehensively assess the impact of multimodality, we transformed the datasets into tabular formats and conducted domain adaptation using domain adversarial training with gradient reversal, employing the DANN approach. Notably, our results showed an improvement in performance, demonstrating an increase of up to 8% in AUC for classification tasks and a reduction of 0.08 in MAE for regression tasks when compared to deploying the model directly on target domains. Remarkably, in the context of the MUL dataset, unsupervised domain adaptation demonstrated competitive performance with transfer learning-based fine-tuning, highlighting its potential to enhance performance even when not explicitly tailored to multimodality. However, for the WEEE dataset, while domain adversarial training led to performance improvement, it fell short of transfer learning. This disparity could be attributed to the presence of high-quality gold standard labels in both source and target domains in WEEE (as opposed to both subjective and objective, but silver-standard labels in MUL), which were effectively harnessed for model fine-tuning when labels were accessible in the target domain. This unique analysis underscores the significance of label quality and its interplay with domain adaptation techniques, offering insights into the diverse impacts of datasets and label types on overall performance.

Contribution 3: We introduce a novel architecture for domain adversarial training, denoted as **M3BAT**, which employs a multi-branch neural network structure featuring multiple encoders tailored to handle different feature sets. Each encoder is designed to accommodate specific features, taking into account factors such as distribution shifts and features stemming from diverse modalities. Through the concatenation of encoder outputs, our architecture incorporates domain adversarial training techniques, including parameter annealing and staged training. Our analysis suggests that employing three branches yields more stable training for the datasets and tasks under consideration for MUL. These branches correspond to high, moderate, and low shift features. For WEEE, we employ 2-3 branches in different setups. Moreover, when these branches are applied to features exhibiting varying degrees

of shifts, our models demonstrate marginally enhanced performance in comparison to conducting unsupervised domain adaptation directly on the entire feature embedding (DANN). On average, we observe an increase of up to 12% in AUC for classification tasks, along with a reduction of up to 0.13 in MAE for regression tasks, as compared to deploying models from the source to the target domain. These findings underscore the potential advantages of our methodology in managing multimodal data and capitalizing on diverse degrees of distribution shift (using a gradient reversal parameter called λ_m that treats branches differently), by having different numbers of modalities within branches (using a parameter called α that determines the number of modalities that fall into high and low shift branches) ultimately enhancing model performance in the context of domain adaptation.

The study is organized as follows. In Section 9.2, we describe the background and related work. Then we describe the proposed architecture in Section 9.4. Section 9.3 provides a description of the data used. In Section 9.5, Section 9.6, and Section 9.7, we define the methods and present results to answer RQ1, RQ2, and RQ3, respectively. We discuss implications, limitations, and future work in Section 9.8, and conclude the chapter in Section 9.9.

9.2 Background and Related Work

9.2.1 Distribution Shift and Unsupervised Domain Adaptation

In the context of machine learning, distribution shift refers to the mismatch between the probability distributions of the data in the source domain (where the model is trained) and the target domain (where the model is deployed) [526]: $p_{X,Y}(\text{source})(x, y) \neq p_{X,Y}(\text{target})(x, y)$. When this mismatch occurs as a result of epistemic uncertainty [222], the model's performance can degrade significantly in the target domain, as it has not seen data from that domain during training. The epistemic uncertainty could be due to many sampling biases such as temporal bias, population bias, and social bias [379]. Hence, in other terms, distribution shift can arise due to various factors, such as differences in data collection settings, user preferences, environmental conditions, and cultural variations.

While there are many nitty-gritty details, three primary types of distribution shift can be identified [526]: covariate shift, label shift, and concept drift. Understanding these types is crucial for effectively addressing the challenges posed by distribution shifts. Covariate shift, also known as input shift or feature shift, occurs when the input data's distribution differs between the source and target domains, but the conditional distribution of the labels given the input remains the same. In other words, the relationship between the input features and the labels is consistent across domains, but the frequency of different feature values may vary. This can be represented as: $p_X(\text{source})(x) \neq p_X(\text{target})(x)$ and $p_{Y|X}(\text{source})(y | x) = p_{Y|X}(\text{target})(y | x)$. Therefore, differences in data collection methods, sensor characteristics, or user behavior across different domains can cause covariate shifts. To illustrate covariate shift, consider a sentiment analysis model trained on movie reviews from the source domain (e.g., American movies) and deployed in the target domain (e.g., Indian movies). The language and writing style of the reviews may differ between the two domains, even though the sentiment expressed by the reviews is the same. In this case, the covariate shift arises from variations in language usage while the sentiment remains consistent. Prior probability shift, also known as label shift, occurs when the label distributions are different between the source and target domains, while the conditional distributions of features given the labels are the same. It can be represented as: $p_Y(\text{source})(y) \neq p_Y(\text{target})(y)$ and $p_{X|Y}(\text{source})(x | y) = p_{X|Y}(\text{target})(x | y)$. Label shift can arise when the labeling process is biased or when the target domain has different class distributions com-

pared to the source domain. Continuing with the sentiment analysis example, label shift may occur if the sentiment expression in movie reviews is perceived differently in different cultures. For instance, positive reviews in the source domain might be labeled as negative in the target domain due to cultural differences in how sentiments are conveyed. Concept drift a much more complex aspect to mitigate [526], and not the focus of this chapter.

Unsupervised domain adaptation (UDA) is a transfer learning technique used to mitigate the effects of distribution shifts between the source and target domains without requiring labeled data from the target domain [526, 161, 85]. The process was primarily developed to handle the covariate shift. While domain adversarial training can indirectly influence the alignment of label distributions or prior probabilities between domains through the shared feature space, it's not the primary mechanism for addressing prior probability shifts. Hence, this process covers both covariate and label shifts to varying extents. Domain adversarial training, introduced by Ganin et al. [162], is a popular approach for UDA. The key idea is to learn a feature representation that is domain-invariant, enabling the model to generalize well across domains. In domain adversarial training, a domain discriminator is introduced along with the primary task model (e.g., classification or regression). The domain discriminator aims to predict the domain of the input data (source or target) based on the feature representation learned by the primary task model. Simultaneously, the primary task model tries to minimize the task-specific loss and maximize the domain discriminator's confusion, effectively aligning the feature distributions between the source and target domains. To increase the confusion, gradient reversal can be used, by multiplying the loss by $-\lambda$ when propagating loss to feature extractor. The domain discriminator, in turn, tries to distinguish between the source and target domains accurately. This adversarial process encourages the primary task model to learn features that are less sensitive to domain variations and more transferable between domains, leading to improved generalization in the target domain. Consequently, unsupervised domain adaptation with domain adversarial training provides a powerful solution to adapt models to new domains and improve their performance in diverse real-world settings, such as multimodal mobile and wearable sensing.

Our work differs from prior work in the sense that we specifically focus on passive sensing datasets captured from mobile and wearable devices, instead of focusing on images or audio. In addition, we focus on multimodal data instead of focusing on a single modality of data, again differentiating our work from previous studies. Studying multimodal mobile sensing is particularly important due to the unique characteristics of mobile sensing data. Unlike computer vision, speech, and natural language processing domains, multimodal mobile sensing deals with data collected from various sensors on smartphones and wearable devices, capturing diverse aspects of users' behavior, environment, and health. The data in multimodal mobile sensing is typically time-series and sequential, providing a continuous stream of information that reflects users' daily activities and interactions. Additionally, multimodal mobile sensing data often includes various modalities such as accelerometer readings, GPS locations, call logs, and app usage patterns, leading to complex and heterogeneous data sources. However, the challenges arising from the heterogeneity, sparsity, and privacy concerns in multimodal mobile sensing data make it essential to explore techniques like unsupervised domain adaptation to ensure robust and effective modeling and generalization across different users and settings.

9.2.2 Mobile Sensing for Inferences Regarding Health and Well-Being, Behavior, and Context

Mobile sensing using smartphones and wearable devices has facilitated the development of context-aware systems that can infer various aspects related to health and well-being, behavior, and context

[268]. These studies leverage diverse sensor data captured by mobile devices, including accelerometer, gyroscope, gps, heart rate monitor, proximity, bluetooth, and app usage, among others, to gain insights into individuals' activities, mood, social context, and energy expenditure [325, 268]. Various studies have explored mood inference, aiming to understand and predict users' emotional states [285, 468]. Servia-Rodríguez et al. [468] collected a large-scale dataset from multiple countries to infer binary mood using the circumplex mood model with population-level models. Mood instability has also been examined using mood reports and phone sensor data [356, 593]. Context-aware systems have been extended to recognize social context, including whether individuals are alone or with others during different activities [328, 327]. For example, Meegahapola et al. [328] used sensor data from Switzerland and Mexico to infer social context during eating activities, while another study [327] examined the social context of young adults during alcohol drinking episodes. Further, energy expenditure estimation (EEE) plays a crucial role in understanding and managing chronic diseases like obesity, diabetes, and metabolic disorders [165]. It also enables personalized health management by providing insights into physical activity, energy consumption, and net calorie intake [265]. Wearable devices such as fitness trackers and smartwatches have been widely used for EEE due to their convenience and capability to measure activity, heart rate, and sleep patterns [109]. These devices overcome the limitations of costly gold standard EEE methods and have been positioned at various body locations to estimate energy expenditure [118, 440]. Overall, the existing literature in the field of multimodal sensing offers valuable insights and tools for inferring various attributes from smartphone and wearable sensor data. However, there is a research gap in understanding generalization and distributional shifts across many different settings, which this chapter aims to address in the context of UDA. The proposed UDA approach seeks to improve the generalization of inference models, making them more adaptable and robust in diverse settings.

9.2.3 Generalization and Distribution Shift in Mobile Sensing

Achieving model generalization across multiple domains is a challenging problem. Transfer learning addresses this issue through domain adaptation, where the model can access some data with labels from the target domains in addition to the source domains [526]. Domain generalization is a more challenging task, where a machine learning model only has access to data from the source domain [569]. In mobile sensing settings, Xu et al. [569] examined this problem and suggested a model based on multi-task neural networks to create a robust model that would work well in target domains without access to data or labels when training. Even though the performance increase that they reported was not high, and in the range of 2%-5%, it was justifiable given that the inference they performed regarding depression detection is already a challenging one. In addition, they highlighted that deep learning-based domain generalization approaches designed for computer vision tasks do not fare reasonably on longitudinal passive sensing data. While they also used multimodal data, they focused on domain generalization, and not unsupervised domain adaptation, which is a different problem setting.

Our analysis focuses on domain adaptation instead of domain generalization, allowing the model to access some data from a target domain. The generic topic of domain adaptation falls under transfer learning. Leveraging transfer learning techniques in domain adaptation could potentially address the distribution shift at the user level, and move towards personalization. Meegahapola et al. [324] and Assi et al. [24] explored how such approaches that lead to the personalization of models affect generalization performance to new countries, in the case of mood inference and complex activity recognition. They also used multimodal sensor data from smartphones in their studies. However, they

did not consider the unsupervised setup where labels might not be available from the target domain. This setting represents a real-life deployment scenario where a trained model is directly applied to new users with access to their unlabeled data.

Finally, it is also worth mentioning the UDA has been previously tried for mobile sensor data by Chang and Mathur et al. [85]. They also used adversarial domain adaptation, similar to ours. However, they only considered a single modality of data in their experiments, hence making the task simpler compared to our experiments, which consider features from multimodal data with varying degrees of shifts across source and target domains. For example, in single modality settings, if the data are accelerometer or audio data, depending on the shift for the specific modality, UDA techniques would facilitate adaptation. However, in mobile sensing, multiple modalities of data are present. The multimodal setting has been studied in a recent study [566], by using all features as input using a single encoder. They also performed domain adversarial training and obtained decent results. However, they do not look into how different modalities could have different levels of shift, and try to accommodate that as their focus was mainly on using different types of labels with a multi-target setup. Therefore, in multimodal settings, some modalities/feature sets might have a high amount of shift, whereas some other modalities/feature sets might have a low or no shift across source and target domains. The effectiveness of UDA for such settings has rarely been studied. In addition, techniques to handle this multimodality in UDA, have rarely been studied in mobile sensing. Therefore, this chapter aims to cover this research gap.

9.3 Datasets

To examine our architecture, we used multiple datasets. Both these datasets have been used in previous publications, and inferences that can be made with them too, are defined. Hence, the objective is to perform the same inferences while examining the proposed architectures.

9.3.1 MUL: Multimodal Smartphone Sensing Dataset from 8 Countries

The MUL dataset was used [324], and this dataset was presented in Chapter 2. The used features are summarized in Table 2.4.

9.3.2 WEEE: Multimodal Wearable Sensing Dataset for Energy Expenditure Estimation

The WEEE dataset was collected from 17 participants (12 men and 5 women) by the authors of [166]. The processed version of the dataset we used in this chapter is from our prior work [18]. The data collection process involved capturing information during the execution of three specific activities: resting, cycling, and running. Participants were equipped with eight different wearable devices, including an Indirect Calorimeter device, which served as the ground-truth measurement for energy expenditure estimation. Alongside sensor data, the dataset also encompasses demographic and body composition details, activity specifics, and questionnaire-based data obtained from each participant. Despite the use of eight devices during data collection, only three devices were selected for this study due to inconsistencies, missing data, and also because they contain comparable multimodal data, allowing us to conduct domain adaptation. These devices and their respective sensors are abbreviated as follows: *i*) Nokia Bell Labs Earbuds: accelerometer, gyroscope, PPG; *ii*) Empatica E4 Wristband: accelerometer, PPG; *iii*)

Chapter 9. Unsupervised Domain Adaptation for Multimodal Mobile Sensing with Multi-Branch Adversarial Training

Table 9.1: Summary of datasets, source and target domains, modalities used, and the performed inferences. C stands for classification and R stands for regression.

Dataset	Source	Targets	Modalities	Inferences
MUL	Italy	India, China, Mexico, Paraguay, UK, Denmark	Location, Bluetooth, Wifi, Cellular, Notifications, Proximity, Activity Type, Steps, Screen Events, User presence, Touch events, App events	Mood (C) [324] Social Context (C) [328, 237]
MUL	Mongolia	India, China, Mexico, Paraguay, UK, Denmark	Location, Bluetooth, Wifi, Cellular, Notifications, Proximity, Activity Type, Steps, Screen Events, User presence, Touch events, App events	Mood (C) [324] Social Context (C) [328, 237]
WEEE	EarBuds	Empatica	Accelerometer, Photoplethysmography	EEE (R) [18]
WEEE	EarBuds	Muse	Accelerometer, Gyroscope	EEE (R) [18]

Muse S Headband: accelerometer, gyroscope. Hence, the objective is to infer and estimate a person’s energy expenditure in a particular moment (the ground truth values come from VO2 Master Analyzer Face Mask: VO2 (ml/kg/min) as the gold standard ground truth¹), based on data from wearable data. Source and target domain combinations were determined based on the availability of common sensors in wearable devices. Further, more details regarding the used features can be found in Appendix A.2.

9.3.3 Domains and Inferences

Our intention here is to delineate the various experimental settings encompassing multiple datasets and inferences. As depicted in Table 9.1, our focus is on two specific datasets: MUL and WEEE.

MUL dataset was employed to facilitate domain transfer across distinct countries, a problem setting motivated by previous studies [324, 24]. Accordingly, we considered Italy ($N_{italy}=151,342$ from 240 users) and Mongolia ($N_{mongolia}=94,006$ from 214 users) as source domains as they have large datasets. Multiple target domains, namely India ($N_{india}=4,233$ from 39 users), China ($N_{china}=22,289$ from 41 users), Mexico ($N_{mexico}=11,662$ from 20 users), Paraguay ($N_{paraguay}=9,744$ from 28 users), UK ($N_{uk}=26,688$ from 72 users), and Denmark ($N_{denmark}=10,010$ from 24 users) were considered for each source domain. Consequently, our analysis spanned a total of 12 source-target pairs within the MUL dataset. For each pair, we undertook two classification tasks previously defined: a two-class mood inference (positive vs. negative) [324] and a two-class social context inference (alone vs. with others) [328, 237]. Both these tasks hold significance in the context of mobile health and mobile food diary applications. It is worth noting that the ground truth for the inferences are: i) mood, which is subjective, and silver-standard² because it is captured with self-reports; and ii) social context, which is more objective, but still silver-standard because of self-reports.

Our approach with the WEEE dataset revolved around domain transfer across diverse devices sharing common sensor modalities. This task is also motivated by prior work that highlights the importance of domain adaptation for devices across body positions [85]. It is noteworthy that not all devices in the original dataset possess identical sensors. For instance, the Nokia Bell Labs earbuds comprise

¹The term "gold-standard" refers to the highest level of accuracy or reliability in ground truth labels or annotations. Gold-standard annotations are often obtained using methods that are considered highly reliable or accurate, such as expert manual annotations, precise measurements, or comprehensive and well-established criteria [566, 142].

²The term "silver-standard" is used to describe annotations or ground truth labels that are of lower accuracy or reliability compared to the gold-standard because of uncertainty, bias, and/or noise. They are still considered useful and informative, and are commonly used in inferences. These annotations might be obtained through less stringent methods, such as automated algorithms, self-reports, surrogate measures, or indirect observations [566, 142].

an accelerometer, gyroscope, and PPG sensor, while the Empatica E4 wristband lacks a gyroscope. Consequently, our experiments encompass two setups: firstly, treating EarBuds ($N_{earbuds}=9,226$ from 17 users) as the source domain and Empatica ($N_{empatica}=9,226$ from 17 users) as the target domain, utilizing accelerometer and PPG data from both devices; secondly, considering EarBuds as the source domain and the Muse S headband ($N_{muse}=9,226$ from 17 users) as the target domain, leveraging accelerometer and gyroscope as the sensing modalities. It is worth noting that data from all devices were collected simultaneously from different body positions, hence the same number of data points. In both setups, we adopted energy expenditure estimation [166] as the target inference, constituting a regression task. This inference is further characterized and validated in prior studies [18, 12, 120, 382]. It is worth noting that the ground truth here is the gold standard for energy expenditure estimation.

9.4 M3BAT Architecture

In this section, we aim to define the proposed architecture, including the intuition behind it. We will first define an unsupervised domain adaptation setting for classification and regression (Section 9.4.1). This can be represented in a generic form similar to DANN [162], as shown in Figure 9.1. Then in Section 9.4.2, we define how multiple branches could be used in both classification and regression instead of a single encoder that outputs a feature embedding. This is also shown in Figure 9.2. Finally, in Section 9.4.3, we describe how different λ could be used for different branches, depending on the shift of input features to that branch in source and target domains, to improve the performance of the model. This is summarized in Figure 9.3.

9.4.1 Unsupervised Domain Adaptation with Domain Adversarial Training

Given two domains, a source domain $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ and a target domain $\mathcal{D}_t = \{x_j^t\}_{j=1}^{n_t}$, where x_i^s and x_j^t represent the input feature vectors from the source and target domains, respectively, n_s and n_t represent the number of samples in the source and target domains, respectively, and y_i^s represents the corresponding target variables in the source domain, the goal is to learn a classifier or regressor $f(x)$ that can accurately infer targets y in the target domain using the labeled source domain data and the unlabeled target domain data. The domain adversarial training process consists of three main components:

- **Encoder:** A multi-layer perceptron neural network represented by $G(x)$, which maps the input feature vectors in dimensionally reduced shared feature space, where $G_s = G(x_i^s)$ and $G_t = G(x_j^t)$ represent the features of the source and target domain samples, respectively.
- **Target Classifier or Regressor:** A head represented by $C(G_s)$, which takes the shared features G_s as input and infers y^s in the source domain.
- **Domain Classifier:** A domain discriminator represented by $D(G_s)$ and $D(G_t)$, which takes the shared features G_s and G_t as input, respectively, and infers whether the features are from the source or target domain.

The overall objective function for unsupervised domain adaptation with domain adversarial training for classification or regression can be written as:

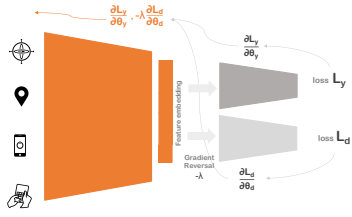


Figure 9.1: Base architecture for UDA with features from multimodal sensors, encoder, domain and target classifier/regressor, and gradient reversal layer.

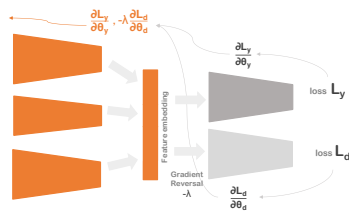


Figure 9.2: Modification to the base architecture to have multiple branches that concatenate to create a feature embedding.

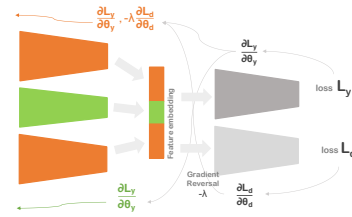


Figure 9.3: Using different λ for branches depending on the average distribution shift of features in the branch. When there is little to no shift, $\lambda \approx 0$ (green).

$$\min_{G,C} \max_D \frac{1}{n_s} \sum_{i=1}^{n_s} \mathcal{L}_y(C(G(x_i^s)), y_i^s) - \frac{\lambda}{n_s + n_t} \sum_{i=1}^{n_s+n_t} \mathcal{L}_d(D(G(x_i^s)), D(G(x_i^t)))$$

where \mathcal{L}_y is the classification loss (e.g., cross-entropy loss) or the regression loss (e.g., mean squared error) function for the source domain samples; \mathcal{L}_d is the adversarial loss function, such as the binary cross-entropy loss, for the domain discriminator to distinguish between the source and target domain features; λ is a parameter that controls the trade-off between the classification/regression and adversarial loss—also known as gradient reversal layer (usually $0 \leq \lambda \leq 1$); The first term aims to minimize the classification or regression loss for the source domain samples, encouraging the model to infer the targets in the source domain accurately; and the second term aims to maximize the domain discriminator’s confusion between the source and target domain features, effectively aligning the feature distributions of the two domains in the shared feature space.

In both classification and regression settings, unsupervised domain adaptation with domain adversarial training is a powerful technique to adapt models trained on a labeled source domain to perform well on a different, unlabeled target domain. The adversarial training process encourages the model to learn domain-invariant features, thereby improving the model’s generalization to new, unseen data from the target domain. In addition, when defining, whether to use $-\lambda$ or $+\lambda$ depends on the loss function used for the domain discriminator. When the domain discriminator is binary cross-entropy which provides a negative value, using $-\lambda$ as above works [162].

9.4.2 Multiple Branches to Process Multimodal Data

To represent the setup with multiple branches for processing modalities or feature sets from multiple modalities, we can introduce separate branches. Let’s denote the encoders as $E^{(m)}(x^{(m)})$, where m represents the number of branches and $x^{(m)}$ is the input data from branch m , which could be multiple features in the tabular datasets that we consider. Each encoder processes the input data from its corresponding modality and maps it to a shared feature space. In this case, the overall feature extractor $G(x)$, which used to be a single encoder in the previous setups mentioned in Section 9.4.1, can now be defined as a combination of these multiple encoders for each modality. We can represent this as: $G(x) = \text{concat}(E^{(1)}(x^{(1)}), E^{(2)}(x^{(2)}), \dots, E^{(M)}(x^{(M)}))$. Here, $G(x)$ contains the outputs from all the encoders corresponding to the different branches, and these outputs are concatenated to form a shared feature representation that captures information from all features. With the multiple branches, the objective function for unsupervised domain adaptation with domain adversarial training can

be extended to include all the modalities. With this setup, each specific encoder learns a feature representation specific to its input data, and the shared feature space created by combining the outputs of these encoders captures information from all the modalities. This approach allows the model to adapt to multiple data modalities simultaneously and improves the domain adaptation performance by considering the shared information among different modalities.

9.4.3 Training Process with Multimodal Domain Adversarial Training

The training process for multimodal domain adversarial training involves a staged approach to adapt the model to the target domain while considering the distribution shift across different modalities. The process includes the following steps:

Step 1: Train Common Encoder with Target Discriminator

In the first step, we train a common encoder, $G(x)$, with only the target discriminator (classifier or regressor), $D(G(x_j^s))$. The target discriminator is responsible for performing either classification or regression. During this step, only the source domain data is used for training. The objective function for this step is to minimize the target discrimination loss, depending on whether it is regression or classification (Section 9.4.1).

Step 2: Introduce Unlabeled Target Domain Data

After training the common encoder with the target discriminator, we introduce the unlabeled target domain data, $\mathcal{D}_t = \{x_j^t\}_{j=1}^{n_t}$, into the training process together with domain discriminator. This is done to further align the feature distributions of the source and target domains in the shared feature space. During this step, the objective function changes a bit as we aim to perform both domain and target inferences, similar to what would happen if we used a multi-task neural network. In terms of gradient reversal, the $\lambda=0$ at this stage.

Step 3a: Increase λ for Adversarial Objective with Annealing

To increase the impact of the adversarial objective gradually, we anneal the value of λ from zero to one during training [573, 161]. The parameter λ controls the trade-off between the target loss and the domain loss. By increasing λ over time, we encourage the common encoder to learn more domain-invariant representations. In Figure 9.1, we show the architecture at this stage, which is similar to the DANN architecture [162]. Annealing λ from zero to one works because it facilitates a controlled and adaptive process of aligning feature representations between the source and target domains, ultimately leading to improved domain adaptation performance. The rationale behind this approach is to start with minimal domain alignment ($\lambda=0$), allowing the model to initially focus on learning source domain knowledge without being influenced by the target domain. As training progresses and the λ parameter gradually increases, the model is encouraged to align the feature representations of both domains [161].

Step 3b: Replace Encoder with Multiple Branches

After training with the common encoder and gradually increasing the adversarial objective, we replace the common encoder, $G(x)$, with a multi-branch setup (Section 9.4). Each encoder, $E^{(m)}(x^{(m)})$, processes the input data from its corresponding modality or a set of features and maps it to the shared feature space. The combined feature representation is then formed by concatenating the outputs of all the branches. With this setup, steps 1 and step 2 can be followed to train the model with staging and annealing. Here $\lambda = 1$ across all branches. In Figure 9.2, we show the architecture at this stage.

Step 3c: Increase λ for Different Branches with Annealing

We train the model as in Step 3b. Then, we adaptively decrease the value of λ from 1 if needed, for each branch with annealing, until it reaches λ_m ($0 \leq \lambda_m \leq \lambda = 1$). The λ_m values, which control the impact of the adversarial objective for each specific encoder, are determined based on the average Cohen's-d value [101] for each feature group. Then, the Cohen's-d values across the branches were normalized to a value between 0 and 1. If the Cohen's-d of the lowest branch is above 0.2 (above small effect size), a zero was introduced artificially before normalizing to ensure a considerable shift does not go unnoticed when performing adversarial training with different λ_m . As an example, if the Cohen's-d values were 0.8, 0.6, and 0.05, the λ_m values would be 1, 0.58, and 0. If the Cohen's-d values from branches were 0.9 and 0.4, we would introduce a 0 to make it 0.9, 0.4, and 0 because 0.4 is above small effect size, and obtain λ_m s 1 and 0.44 for the two branches. Hence, this accounts for the distribution shift between the source and target domains specific to input features to each branch. Figure 9.3, we show the architecture at this stage.

With this staged training process, the multimodal domain adversarial training algorithm can effectively adapt the model to the target domain while considering the distribution shift across different modalities or feature groups (e.g., regardless of the modality, features with a high, moderate, and low distribution shift in separate branches). The hypothesis is that this approach would allow the model to learn domain-invariant representations capable of capturing relevant information from all modalities, improving the generalization and adaptation performance to new, unseen data in the target domain.

9.5 Using Statistical Tests to Quantify Distribution Shift of Sensors (RQ1)

9.5.1 Methodology

The aim of this analysis is to provide empirical evidence for the rationale behind the development of a multi-branch architecture. In accordance with previous studies [526], two primary methods for quantifying distribution shift are statistical tests [363] and inference performance metrics [324]. Statistical test-based techniques are known for their cost-effectiveness and ability to offer a general estimation of the shift for each sensing modality [363]. Thus, we could employ common statistical tests such as t-test [249], PERMANOVA and PERMDISP [363], and Cohen's-d [101] to assess the distribution shift of sensor modalities for each target inference. In this context, after an initial analysis of these tests, we selected Cohen's-d for our analysis due to its relatively linear distribution of values [324, 24, 363], within a range approximating 0 and 1. Most importantly, it allowed the best downstream performance for domain adaptation. In addition, the rule of thumb of 0.8 or above: large effect size, 0.5: moderate effect size, and 0.2: small effect size allows easy interpretation [101]. This characteristic facilitated the

9.5 Using Statistical Tests to Quantify Distribution Shift of Sensors (RQ1)

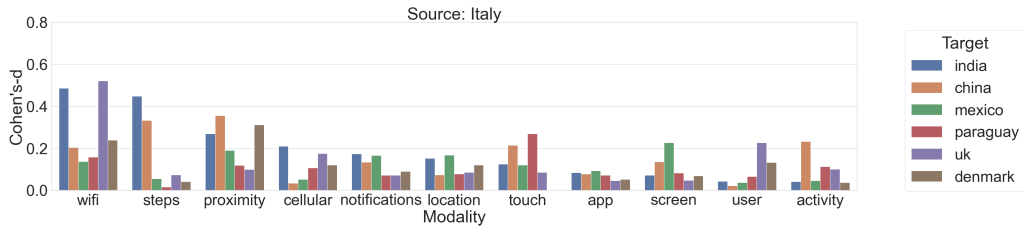


Figure 9.4: Average Cohen's-d Values for Modalities. Italy is the Source Domain.

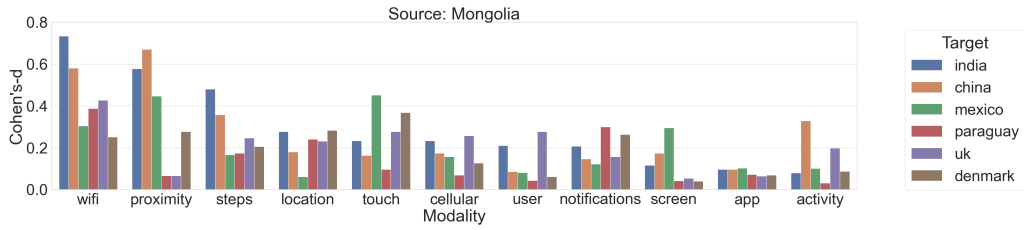


Figure 9.5: Average Cohen's-d Values for Modalities. Mongolia is the Source Domain.

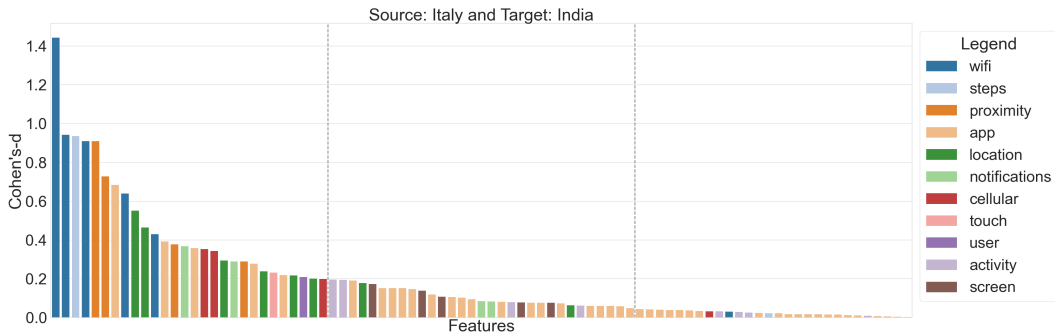


Figure 9.6: Cohen's-d Values Italy and India, Sorted in the Descending Order. Modalities Marked in Different Colors.

utilization of normalized values in our architecture for λ_m (Section 9.4.3)³.

For the MUL dataset, our initial step involved calculating Cohen's-d values for all captured features. Subsequently, we aggregated these values by computing the mean for each modality (e.g., wifi, steps, proximity, location, etc.). By designating Italy and Mongolia as source domains, we plotted the results for other target domains (Figure 9.4 and Figure 9.5). This approach enabled us to comprehend how modalities could exhibit varying degrees of distribution shifts for the same source and target domains. The underlying concept is to demonstrate that aggregating features by modalities facilitate the differentiation of high and low levels of distribution shifts. Continuing, we proceeded to visualize Cohen's-d values for all features, assigning distinct modality-specific colors to each bar for enhanced clarity (Figure 9.6). This visualization aimed to provide insights into whether distribution shifts often emanate from the same set of modalities or if there are instances of outliers with high Cohen's-d values from specific modalities exhibiting relatively low levels of distribution shift overall. Due to space limitations,

³Please note that we use terms distribution shift or shift between two modalities interchangeably to refer to Cohen's-d, from this point onward.

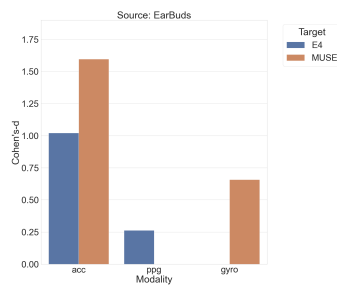


Figure 9.7: Average Cohen's-d Values for Modalities. EarBuds is the Source Domain

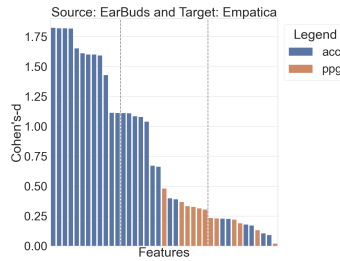


Figure 9.8: Cohen's-d Values for EarBuds and Empatica, Sorted in the Descending Order. Modalities Marked in Different Colors.

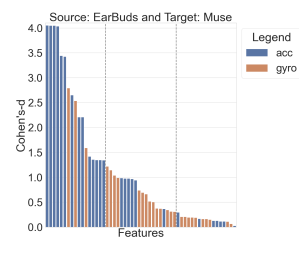


Figure 9.9: Cohen's-d Values for EarBuds and Muse, Sorted in the Descending Order. Modalities Marked in Different Colors.

we only show the distribution for the case when Italy is the domain, and India is the target domain.

For the WEEE dataset, we did a similar analysis. We aggregated Cohen's-d values by computing the mean for each modality (e.g., acc, ppg, gyro — Figure 9.7). It is worth noting that, in the setup of transferring from EarBuds to Empatica, only accelerometer and PPG data are available. In the other case of EarBuds to Muse, only accelerometer and gyroscope data are available as common features. Continuing, we proceeded to visualize Cohen's-d values for all features, assigning distinct modality-specific colors to each bar for enhanced clarity. This was done separately for pairs EarBuds and Empatica (Figure 9.8) and EarBuds and Muse (Figure 9.9).

9.5.2 Results

For MUL, Figures 9.4 and 9.5 present the quantification of distribution shift using Cohen's-d values, which indicate the effect size. The x-axes represent modalities, while the y-axes show the shift. Each modality is color-coded to represent the target domain. Notably, the figures reveal that various countries exhibit diverse modalities that exhibit the highest and lowest distribution shifts compared to Italy. For instance, WiFi is prominent for India and the UK, proximity for China and Denmark, screen for Mexico, and touch for Paraguay. This pattern remains consistent when the source is Mongolia, except for touch in Mexico and Denmark, and WiFi in Paraguay. Moreover, when analyzing shifts within the target countries, the values for different modalities contrast. For instance, with Italy as the source, the target domain India exhibits WiFi and steps with a Cohen's-d of around 0.5 (medium effect size). Conversely, all modalities such as app, screen, user, and activity have values below 0.1 (very small effect size), indicating minimal distribution shift. Similar trends are observed for other countries, even when Mongolia is the source. However, it is important to note that these diagrams do not account for individual features within modalities, which could have high distribution shifts, but their impact might be mitigated by numerous other features within the same modality with low shifts. This aspect is demonstrated in Figure 9.6, where the distribution shift of each feature for source Italy and target India is plotted, with colors denoting modalities. While we can't visualize all such features across various source-target domain pairs, we have identified a considerable number of such cases that highlight significant shifts despite the overall low shift in the modality.

Moving to the WEEE dataset, in Figure 9.7, we illustrate the modality-specific shifts for the source-target pairs of Earbuds and Empatica, as well as Earbuds and Muse. For the EarBuds and Empatica pair, it's

evident that accelerometer features exhibit a substantial shift (Cohen's-d of ≈ 1 —indicating a large effect size), whereas PPG features demonstrate only a smaller shift (Cohen's-d of ≈ 0.2 —reflecting a small effect size). Similarly, a contrasting pattern is observed for the EarBuds and Muse pair, with accelerometer and gyroscope modalities showcasing noticeable shifts. Notably, even the gyroscope exhibits a considerable shift in this context (Cohen's-d around 0.6—indicating an above-medium effect size). To delve deeper into this analysis, we present feature-specific shifts in Figure 9.8 and Figure 9.9. Regarding the EarBuds and Empatica pair, fewer outliers are observed compared to the modality-specific pattern. Most features with above-medium effect sizes primarily originate from the accelerometer, while PPG features tend to exhibit small to medium effect sizes, indicating smaller shifts. Conversely, in the case of the EarBuds and Muse pair, the features exhibit a more mixed distribution, deviating from the modality-specific shifts highlighted in Figure 9.7.

In summary, in answering RQ1, the statistical analysis suggests that it might be possible to categorize features into groups with high, medium, or low distribution shifts based on modalities or by considering mixed feature groups derived from multiple modalities with prominent effect sizes. These insights directly address the design of the architecture that we use to answer RQ3, where we propose the use of multiple branches for distinct feature groups, as detailed in Section 9.4.2.

9.6 Domain Adversarial Training with Multimodal Sensing Features (RQ2)

9.6.1 Methodology

As the next step, we performed domain adversarial training using the base architecture described in Sections 9.4.1. Our experiments were implemented in Python, with TensorFlow [2] and Keras [91]. The architecture consists of an encoder without considering the multimodality of the data. Our dataset splitting involved separating training (70%) and testing (30%) sets to ensure non-overlapping users, facilitating leave-one-out cross-validation [200]. We conducted five such random training and testing splits to ensure robustness and reported the average results. This experimental setup is similar to the approach proposed in [85], with the distinction that we employed processed tabular features from multiple modalities, similar to [566, 330], instead of the raw sensor values with a feature extractor.

For the MUL dataset, we initiated our analysis by training models on the training sets for Italy and Mongolia as source domains. Our model architecture was designed to infer Mood and Social Context, with intermediate layer sizes of 128, 128, and 64, all with the ReLU activation [9]. Dropout [487] was used with rates of 50%, 50%, and 20% to mitigate overfitting. We used sigmoid activation [140] and binary cross-entropy loss function [542], suitable for the two-class nature of our inferences. Adam optimizer [250] and a batch size of 32 was chosen. We also implemented early stopping to prevent overfitting after five epochs of non-improving validation loss within {0, 300} epochs. Performance evaluation employed the area under the receiver operating characteristic curve (AUC) with macro averaging, which considers class imbalance [24, 324]. Evaluation of the models began by assessing the performance of the Source model on the Source testing set (S->S). These results were averaged across the five iterations. Subsequently, we evaluated the Source model on the Target datasets (S->T). Due to multiple targets for each source, we averaged the results. We also fine-tuned the source model on target training sets with transfer learning and evaluated on the target domain testing set (S->T (w/ TL)). Note that this setup is not unsupervised and needs labels in the target domain. We then proceeded with domain adversarial training (DANN [162]), as outlined in Section 9.4.3—Step 3a, where we first trained

the encoder (with layer sizes 128 and 64) and later the target classifier (with intermediate layer sizes 64 and 32) and domain classifier (with intermediate layer sizes 64 and 32). These classifiers employed sigmoid activation and binary cross-entropy loss at their respective final layers. A fixed $\lambda = 1$ was used for gradient reversal during encoder training. Data with labels from the source domain contributed to the loss for both domain and target classifiers during training, while unlabelled target domain data only contributed to the domain classifier loss.

Even for WEEE dataset, the methodology was similar to that of the MUL dataset, albeit with smaller models due to the dataset's size. We trained models for EarBuds, utilizing accelerometer and PPG data, and for EarBuds with accelerometer and gyroscope data. These modality combinations were chosen to allow domain adaptation for two devices, as described in Table 9.1. The models were designed to infer energy expenditure estimation, with intermediate layer sizes of 64 and 32, and using the ReLU activation function. Dropout with rates of 30% and 20% was used for regularization. The mean squared error loss function was utilized given that it is a regression—Adam optimizer and a batch size of 16 facilitated model training. Early stopping was also implemented. The evaluation process mirrored that of the MUL dataset. **DANN**, following the process outlined in Section 9.4.3, involved training the encoder (with layer sizes 64 and 32) and later the target regressor (with intermediate layer size 32) and domain classifier (with intermediate layer sizes 32 and 16). The target regressor employed mean squared error, and the domain classifier employed sigmoid activation and binary cross-entropy loss at their final layers. A $\lambda = 1$ was used for gradient reversal during encoder training.

Hence, in summary, given below are the inferences that we conducted across both datasets.

































- **S->S**: performance of the model trained in the source domain, for the source testing set. This provides an upper bound for the possible results in the target domain.
- **S->T (w/ TL)**: performance of the model trained in the source domain, for the target testing set, after fine-tuning the target training set with transfer learning. This setup assumes that labels are available in the target domain, hence could lead to higher performance and act as another ceiling for performance. It is also worth noting that ground truth labels used in training models can be gold-standard or silver-standard as mentioned in Section 9.3. In MUL, ground truth is the silver standard as they are self-reports that can be noisy. However, mood reports can be subjective and social context reports are more objective, potentially leading to differences in the accuracy and noisyness [325]. In WEEE, the ground truth is the gold standard.
- **Random**: performance of a random classifier/regressor on the target testing set.
- **S->T**: performance of the model trained in the source domain for the target testing set.
- **DANN**: performance of the model trained in the source domain and unlabelled target data (training set), on the target testing set.

9.6.2 Results

Table 9.2 presents the classification outcomes for the MUL dataset, for mood and social context inferences. The results show the model's performance under different scenarios. Specifically, **S->S** showcases how the model performs when evaluated on the source testing set. The inferred values fall within the range of 0.59 to 0.61 AUC. Although the performance in the source domain is not notably high, this aligns with trends observed in recent research focused on mental well-being and contextual

9.6 Domain Adversarial Training with Multimodal Sensing Features (RQ2)

Table 9.2: Results for classification tasks. Results are presented as average AUC scores (higher the better). TL refers to transfer learning where labelled target domain data are available.

Dataset	MUL	MUL	MUL	MUL
Inference	Mood	Mood	Social Context	Social Context
Source	Italy	Mongolia	Italy	Mongolia
Targets	Avg. of Countries	Avg. of Countries	Avg. of Countries	Avg. of Countries
RQ2				
S->S	0.61±0.04 	0.59±0.06 	0.63±0.03 	0.65±0.04 
Random	0.45±0.12 	0.46±0.08 	0.42±0.14 	0.44±0.11 
S->T	0.46±0.08 	0.48±0.07 	0.44±0.03 	0.49±0.08 
S->T (w/ TL)	0.51±0.05 	0.52±0.06 	0.56±0.05 	0.55±0.07 
DANN [162]	0.52±0.07 	0.53±0.02 	0.52±0.03 	0.54±0.05 
RQ3				
Ours ($\lambda = 1$, Setup1)	0.55±0.05 	0.52±0.06 	0.55±0.06 	0.57±0.05 
Ours (w/ λ_m, Setup1)	0.56±0.04 	0.53±0.07 	0.56±0.04 	0.57±0.05 
Ours (w/ λ_m, Setup2)	0.58±0.04 	0.54±0.03 	0.55±0.05 	0.55±0.03 

inference using multimodal mobile sensing datasets [569, 324]. Despite not achieving high levels of performance, these results still provide a foundation for investigating domain adaptation techniques, where even small enhancements in performance on target domains are crucial. **S->T** is where the model trained on the source domain is evaluated on the target domain’s testing set. As expected, performance experiences a decline across all four inferences compared to **S->S**. With transfer learning fine-tuning (**S->T (w/ TL)**), the performance improves across all four inferences as expected because it uses labels in the target domain. Interestingly, the application of **DANN** leads to further performance enhancements across social context inferences while showing a slight performance decline compared to **S->T (w/ TL)** for mood inference. This could be because mood labels are more subjective; hence even having labels in the target domain is less useful, whereas **DANN** leads to marginally better results. However, for social context, which is more objective ground truth, having labels led to increased performance, even more than the adapted model with **DANN**. Hence, while more exploration is needed, this could be evidence that the label source quality and objectivity might have an effect on fine-tuning or domain adaptation performance.

Table 9.3 presents the regression outcomes for the WEEE dataset, focusing on energy expenditure estimation (EEE) inference. In this context, EarBuds serves as the source domain, while Empatica or Muse serves as the target domain. The source domain performance (**S->S**) yields MAE values of 0.62 and 0.59 for EarBuds, representing the desired ceiling performance. The random baseline, on the other hand, delivers poor results. In the **S->T** scenario, the performance exhibits a reduction of approximately 0.17 MAE and 0.21 MAE for Empatica and Muse, respectively. Notably, the application of transfer learning (**S->T (w/ TL)**) results in improved performance compared to **DANN**. This stands in contrast to mood inference in the MUL dataset results, where **DANN** marginally outperformed transfer learning for mood inference. The divergence in results can be attributed to the nature of labels; the MUL dataset employs silver standard labels derived from user self-reports, however, with high and low subjectivity for mood and social context, respectively, while both labels could also be susceptible to noise. On the other hand, the WEEE dataset leverages gold-standard labels derived from lab-based measurements. Our findings underscore that while transfer learning with silver standard labels, especially when having subjective ground truth, does not universally guarantee performance gains, it presents an alternative ceiling. It is noteworthy however, that transfer learning necessitates the availability of labels in the target domain, which is not the primary scenario we are addressing.

In conclusion, when tackling RQ2, our exploration revealed that domain adversarial training applied to

multimodal mobile sensing datasets translates to enhanced performance compared to **S->T** setting. Notably, for mood inference in the MUL dataset, the observed increase even surpassed/equaled that achieved through transfer learning. This phenomenon can likely be attributed to the presence of silver standard, and potentially subjective labels in this dataset. On the other hand, when examining the WEEE dataset, **DANN** contributed to improved performance, although not reaching the same level as transfer learning, which was to be expected. This discrepancy could be attributed to the presence of high-quality gold standard labels available in both the source and target domains, which were effectively utilized to train models. This calls for deeper investigations in this direction in the future.

9.7 Multi-Branch Domain Adversarial Training (RQ3)

9.7.1 Methodology

Experiments with Multiple Branches

The next step is to replace the encoder with multiple branches as described in Section 9.4.2. We considered experimental approaches with two distinct setups. These setups were based on the results we obtained for experiments in Section 9.5. There, we discussed how distribution shift could be quantified by aggregating features based on modalities (e.g., activity, steps, wifi, location, etc.) or by aggregating based on feature level shift as quantified by statistical tests (e.g., top 33% of features, bottom 33% of features, and the rest, etc. regardless of the modality).

Setup1—Branches Based on Modalities: In the MUL dataset, there are over ten modalities. Having separate branches for all modalities leads to a complex optimization problem. Hence, in this chapter, we focus on having three branches for which we were able to obtain decent results. Beyond the three branches, we did not obtain good results, as it became a difficult optimization given the dataset size and the challenging nature of the task, as described in Section 9.6. Hence, when having three branches, for each target country, we used the modality with the highest shift as one branch, the modality with the lowest shift as another branch, and the rest of the modalities in one branch, as visualized in Figure 9.2. When considering modalities, we normalized the highest Cohen's-d modalities λ_m to 1, and if the Cohen's-d of the lowest modality was below 0.2 (below small effect size), normalized it such that it is $\lambda_m = 0$ (all source-target pairs had below 0.1 modalities, as shown in Figure 9.4 and Figure 9.5). λ_m for the set of features in the middle was normalized to a suitable value, as described in Section 9.4.3—Step 3c. We only have two modalities in the WEEE dataset for both inferences. Hence, we used two branches and again normalized between 0 and 1 to obtain the λ_m values for the two modalities, following Section 9.4.3. Finally, for this setup, we first conducted experiments with $\lambda = 1$ for all branches ($\lambda = 1$, **Setup1**). Then, we used different λ_m values for branches based on the shift and conducted experiments (**w/ λ_m , Setup1**).

Setup2—Branches Based on Feature Group: In both datasets, we could sort all features in descending

Table 9.3: Results for regression tasks. Results are presented as mean absolute errors (MAE) (the lower the better).

Dataset	WEEE	WEEE
Inference	EEE	EEE
Source	EarBuds	EarBuds
Target	Empatica	Muse
RQ2		
S->S	0.62±0.11	0.52±0.06
Random	1.35±0.31	1.41±0.43
S->T	0.79±0.15	0.73±0.10
S->T (w/ TL)	0.67±0.07	0.56±0.04
DANN [162]	0.73±0.05	0.65±0.06
RQ3		
Ours ($\lambda = 1$, Setup1)	0.74±0.04	0.62±0.06
Ours (w/ λ_m , Setup1)	0.69±0.06	0.60±0.03
Ours (w/ λ_m , Setup2)	0.69±0.05	0.64±0.03

order based on shift for each source-target pair. Then, we could consider three groups similar to Setup 1, by considering 33% of data with the highest shift, 33% of data with the lowest shift, and finally, the rest of the 33% of data in the middle. While the percentage could be changed, we did not delve deeper into that in this analysis and focused on obtaining equal splits for the three feature sets. In Figures 9.6, 9.8, and 9.9, these splits are marked with vertical dotted lines. Finally, for this setup, we used different λ_m values for branches, based on the shift, as suggested in Section 9.4.3 (**w/ λ_m , Setup2**).

Experiments with Multiple Modalities in High and Low Branches for Setup1

An identified limitation of Setup1 is evident in instances where a single modality, representing branches with the highest or lowest shift, encompasses only a limited number of features. For instance, when the source domain is Italy and the target is India, the modality with the highest shift, ‘wifi’ ($\lambda_0 = 1$), includes merely seven features. In contrast, the ‘activity’ modality with the lowest shift contains eight features ($\lambda_2 = 0$), while the remaining features (over 80) fall within the moderate shift range ($\lambda_1 = 0.62$). This scenario results in an imbalance among the branches. However, such an imbalance does not manifest in Setup2, where the equitability of sizes across branches is maintained. We performed another experiment to assess the potential implications of this limitation on performance. Specifically, we introduced additional modalities to the high and low-shift branches. To accomplish this, we define α to indicate the number of modalities present within the high and low shift branches. Thus, $\alpha = 1$ corresponds to one modality each in the high and low branches, $\alpha = 2$ signifies two modalities each, $\alpha = 3$ represents three modalities each within these branches, and so on. Subsequently, we conduct a series of experiments akin to those described in Section 9.7.1, systematically varying the α values. This experimentation enables us to gauge the influence of modifying the number of modalities on performance. Given the inadequacy of modalities within the WEEE dataset, it is important to underscore that this specific experiment pertains only to the MUL dataset.

9.7.2 Results

Experiments with Multiple Branches

The results for the MUL dataset are presented in Table 9.2. When λ is set to 1, indicating domain adversarial training with uniform λ values across all branches, the performance exceeds that of **DANN** in all instances except for mood inference with Mongolia as the source. Within Setup1, the introduction of distinct λ_m values for branches, determined by modality distribution, yields minor performance enhancements across all scenarios compared to the uniform $\lambda = 1$ configuration, except for social context inference with Mongolia as the source where both setups yielded an AUC of 0.57. Moreover, the superiority of Setup1 over Setup2 is not clear, as each configuration displayed better performance in different inferences. However, a discernible trend emerged with mood inference, which is a more subjective task reliant on nuanced labeling, indicating that fine-tuning with labels (**S->T (w/ TL)**) failed to elevate performance, even relative to **DANN** and our proposed approach. Conversely, transfer learning achieved performance on par with our method for the objective ground truth of social context inference, again underscoring the possible impact of ground truth nature on inference accuracy as discussed in Section 9.6.2. From another sense, this highlights that our technique achieves decent performance, even compared to transfer learning, for social context inferences in the MUL dataset.

The findings for the WEEE dataset are shown in Table 9.3. Here, our approach once again outperformed **DANN** and other baselines across both inferences. However, the performance does not surpass that of

fine-tuning with transfer learning. This discrepancy can be attributed to the presence of gold-standard ground truth labels in this dataset, allowing fine-tuning to exhibit superior performance. Furthermore, distinguishing between the efficacy of Setup1 and Setup2 remains inconclusive, despite Setup1’s superior performance for the Earbuds and Muse source-target pair. As mirrored in the results for the MUL dataset, even here, employing diverse λ values for the branches proved to be more effective than adopting a uniform $\lambda = 1$ strategy. Thus, the adjustment of λ based on branch-specific shift statistics yielded more better results.

Experiments with Multiple Modalities in High and Low Branches for Setup1

The outcomes of the experiment conducted to explore the effects of varying α values are presented in Figure 9.10. The results exhibit distinct patterns across different inferences and source-target domain pairings. In specific scenarios, such as the ‘Italy-All Mood’ and ‘Mongolia-All Social Context’, employing $\alpha = 2$ yielded slightly superior results compared to when α was set to 1. Similarly, for the ‘Mongolia-All Mood’ inference, $\alpha = 2$ yielded results akin to those of $\alpha = 1$, while $\alpha = 3$ led to notably improved performance. Intriguingly, in all instances, setting $\alpha = 4$ resulted in consistently subpar performance. These findings underscore the nuanced nature of determining optimal configurations for the number of branches and the number of modalities within each branch during model training. There appears to be no universal formula for these selections, as their efficacy depends on the specific inference and source-target domains. Nevertheless, a recurring trend across the analyses is that employing multiple branches with an appropriately chosen α and λ consistently outperforms utilizing a single encoder with fixed λ , except in very few experiments (e.g., WEEE Earbuds-Empatica DANN=0.73 performed better than $\lambda = 1$, Setup1; WEEE Earbuds-Muse $\lambda = 1$, Setup1 performed better than w/ λ_m Setup2; MUL Mood w/ source Mongolia DANN = 0.53 performed similar to $\lambda = 1$, Setup1 and w/ λ_m , Setup1).

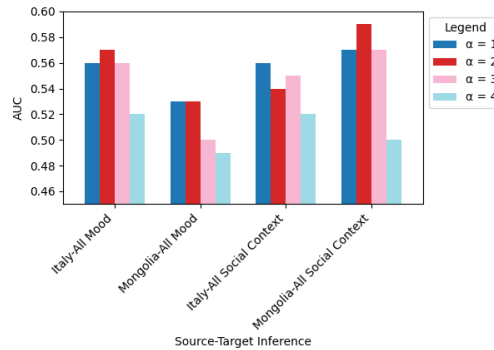


Figure 9.10: Inference results for various α values.

As a summary, in answering RQ3, the analysis shed light on the nuanced interplay between diverse factors such as the nature of ground truth (gold/silver standard, subjective/objective), modality distributions and related distribution shifts, and branch-specific adaptation dynamics. Hence, the conducted experiments provide evidence that incorporating multiple branches for different feature sets, based on either modality or distribution shift-based feature groups, yields improved performance in unsupervised domain adaptation. The results for both the MUL and WEEE datasets consistently indicate that this approach outperforms baseline methods and enhances domain adaptation performance. Further, the findings emphasize the importance of adaptability in adjusting parameters such as λ and α based on specific contexts, inference tasks, and source-target domain pairings. This adaptability proves crucial in effectively harnessing the advantages of domain adversarial training, highlighting its potential to significantly enhance model generalization and performance across diverse multimodal mobile sensing datasets.

9.8 Discussion

9.8.1 Implications

The use of multi-branch encoders with varying lambda values in domain adaptation carries both theoretical and practical implications. First, given below are some theoretical implications.

1. **Adaptation Strategy Customization:** The insights drawn from the experiments underscore the importance of adopting good adaptation strategies. The observed performance improvements achieved through tuning parameters such as λ and α values emphasize the significance of tailoring the training process to the specific characteristics of the data and the nature of the inference tasks. This departure from the prevalent one-size-fits-all approach in unsupervised domain adaptation demonstrates a more effective and nuanced approach to enhancing model performance. The results highlight the need for flexibility in adaptation techniques, acknowledging that different data modalities and inference tasks may demand distinct strategies for optimal performance. This implication encourages researchers and practitioners to consider the intricacies of the data and the problem domain when crafting adaptation methodologies, contributing to a more effective approach to domain adaptation.

2. **Impact of Ground Truth Nature:** The divergent performance trends observed across various inference tasks, particularly distinguishing between subjective tasks like mood inference, more objective tasks like social context inference, and gold standard ground truth-based tasks like energy expenditure estimation, unveil evidence of the impact of ground truth nature on the domain adaptation model's performance. These findings highlight the intricate interplay between ground truth labels' quality and reliability and domain adaptation methods' success. The outcomes showcase the need to take into account the inherent subjectivity and potential noise in ground truth labels, particularly in contexts where human judgment and perception play a crucial role. The implications underscore the pivotal role of domain adaptation not only in mitigating distribution shifts but also in accommodating the peculiarities of the ground truth data. This revelation aligns with the ongoing discourse about the importance of robust label acquisition strategies in ensuring the efficacy of domain adaptation approaches.

In addition, in the deployment of this kind of model, the following practical implications could be considered.

1. **Enhancing Real-world Applicability:** The tangible performance improvements showcased by the multi-branch adaptation strategies carry direct implications for real-world applications reliant on multimodal mobile sensing data. These findings suggest that practitioners navigating the challenges of integrating data from diverse sensors can effectively enhance the utility of their models by customizing their adaptation strategies based on the nuanced characteristics of the data and the demands of specific inference tasks. This adaptability offers a concrete means to elevate predictive accuracy and obtain deeper insights from data-driven applications such as mental well-being monitoring, personalized healthcare, and behavior analysis. By fine-tuning adaptation techniques to the intricacies of data distributions and domain shifts, practitioners can achieve more robust and meaningful outcomes in these critical domains.

2. **Guidance for Model Design:** The insights obtained from the experiments offer valuable guidance for practitioners and researchers engaged in designing and deploying adaptation models for multimodal data. By comprehending the impact of factors such as ground truth quality and modality distribution

on model performance, informed decisions could be made about the architecture and configuration of their models. This understanding streamlines experimentation, minimizes trial-and-error efforts, and accelerates the development of effective adaptation techniques tailored to specific use cases. The findings enable researchers to navigate the complex landscape of parameter tuning, modality selection, and ground truth consideration with greater clarity and purpose, thereby facilitating the more efficient development of robust adaptation models.

In conclusion, the theoretical and practical implications derived from the experimental results collectively underscore the inherent flexibility, adaptability, and performance-enhancing capabilities of the M3BAT architecture for unsupervised domain adaptation. These implications advance the theoretical understanding of adaptation techniques and provide actionable insights for practitioners and researchers seeking to harness these strategies to elevate model performance and extend generalization capabilities. As the field of multimodal mobile sensing continues to mature, the insights derived from these implications inform the development of sophisticated techniques that effectively address the complexities of domain shifts in real-world settings.

9.8.2 Limitations and Future Work

The presented domain adaptation technique for multimodal mobile sensing exhibits promising potential in addressing domain shift and enhancing generalization across datasets. However, several limitations warrant consideration and provide avenues for future research to enrich the methodology's applicability and effectiveness. While domain adversarial training serves as a foundational technique, it might not comprehensively capture the intricacies of real-world domain variations. Moreover, domain adaptation techniques are widely separated into discrepancy-based and adversarial-based [178, 566]. The model we proposed in this chapter is an adversarial-based technique. Further exploration into other adaptation techniques across both types is needed to handle more intricate and challenging domain shifts. To this end, future investigations could delve into approaches like maximum mean discrepancy (MMD) [571, 85], adversarial discriminative domain adaptation [517], self-ensembling [399], and moment matching [396], expanding the set of techniques available for mitigating domain shift in mobile sensing settings.

Moreover, the assumption of prior probability shift or label shift, while valuable, might not consistently mirror the complexities of real-world data distributions. While this is not necessarily a problem only for the ubiquitous and mobile sensing domain, given the nature of data and involved intricacies, exploring alternate assumptions or techniques for scenarios involving unknown label distributions is crucial for capturing a more comprehensive range of domain variations and challenges. The need for more work in these directions in ubiquitous and mobile sensing is highlighted in recent studies [363]. Hence, future work could look into these areas. Further, the current methodology leverages separate branches for individual sensor modalities (Setup1), a pragmatic approach that effectively considers the unique characteristics of each modality. However, exploiting potential synergies between modalities remains a compelling avenue for future research. Even though this was partially done with Setup2, the development of more sophisticated multimodal fusion techniques could enable a more comprehensive integration of information across different sensor sources, potentially yielding further performance gains. Scalability to large-scale datasets and applicability in real-world deployment scenarios are vital considerations for the practical utility of the technique. To this end, future research should prioritize refining the method's efficiency and robustness in diverse real-world settings, facilitating its seamless integration into practical applications across domains.

Expanding the technique's versatility and usefulness could involve extending it to encompass transfer learning and personalized model scenarios. This broader scope could cater to diverse application needs and user-specific requirements, enhancing the technique's adaptability and applicability. For example, given our findings, sometimes it might make sense to do domain adaptation first and personalize a model to target users in a target domain rather than directly personalizing a model. These directions need further investigation. Additionally, the technique's application has been primarily centered around time series data processed through time windows, as is customary in this context across many studies [566, 324, 485, 24]. Future research could explore its adaptability when confronted with raw time series data alongside convolutional neural network-based feature extraction, similar to how multimodal data are handled in [569]. This expansion would offer insights into the method's effectiveness under different data representations and processing techniques. Furthermore, while the technique has been demonstrated within the domain of mobile sensing, its fundamental principles render it applicable to various tabular datasets with multimodal attributes, extending beyond mobile sensing applications. Future investigations could explore its utility across diverse domains, enriching the understanding of its generalizability and impact. Finally, it is also worth noting that the set of results we included in the chapter is the core set of results that would allow us to convey the idea and feasibility while adhering to page limits and conciseness. Hence, there are aspects such as different percentages of data splits for Setup2 and optimizing λ_m as a hyper-parameter (which could be expensive computationally) that could be explored in future work, in more detail. Moreover, while datasets to explore our research questions are limited, the technique's generalizability for other inferences and datasets with different source-target domain pairs could be explored in future work. In conclusion, while the proposed domain adaptation technique exhibits promising results, these limitations and unexplored avenues provide the impetus for further research and innovation, fostering the continuous evolution and advancement of multimodal domain adaptation methodologies in the context of mobile sensing and beyond.

9.9 Conclusion

In this chapter, we explored the effectiveness of a multi-branch domain adaptation technique for multimodal mobile sensing data. Our experiments on the MUL and WEEE datasets highlight the adaptability of the approach through parameter customization, leading to enhanced performance and generalization. The results underscore the need for tailored adaptation strategies, while the distinction between subjective and objective tasks emphasizes the role of ground truth quality. The technique's potential for scenarios with limited labeled data and its applicability to practical settings further demonstrate its significance. However, challenges remain, and future research should focus on refining the technique's scalability, real-world deployment, and fusion of multimodal data. As mobile sensing gains momentum in various domains, this study contributes to the advancement of unsupervised domain adaptation with M3BAT architecture, with potential implications for a wide range of real-world machine learning applications.

10 Conclusion

10.1 Summary of Contributions

This thesis represents an in-depth exploration into the realm of multimodal mobile sensing. It uncovered the potential of smartphone sensors in understanding various facets of everyday life and human behavior, emphasizing the critical challenges of generalization and personalization of models. From discerning drinking social contexts and food consumption levels to identifying eating events and inferring mood, this thesis contributes both empirically and methodologically to the fields of ubiquitous computing and digital health. Moreover, by addressing the issue of model generalization across different source and target domains and proposing innovative domain adaptation techniques, this thesis provided advances towards the practical implementation of mobile sensing models in real-world, diverse environments, ultimately enhancing their utility and impact.

Chapter 2 introduced two mobile sensing in the wild data collections for which we contributed specific design elements as part of a large team of partners. These data were collected during pilots of the European WeNet project, with the aim of understanding the everyday behavior and well-being of young adults. The first dataset (MEX) was collected from over 80 college students in Mexico. The second dataset (MUL) was a multi-country dataset from 8 countries (Italy, Denmark, UK, India, China, Mongolia, Mexico, and Paraguay). This dataset was collected from over 680 college students across these countries. Most of the work in this thesis builds upon these two datasets, in addition to using two other datasets from previous work.

Chapter 3 studied mobile sensing capabilities to infer social contexts, particularly focusing on the drinking behavior of young adults. By analyzing data collected from 241 young adults during weekend nightlife in Switzerland, we unveiled the potential of smartphone sensor data to discern diverse drinking social contexts with reasonable accuracy. These findings have the potential for future interventions related to alcohol-drinking behavior, emphasizing the importance of incorporating social context information.

Chapter 4 investigated into eating behavior, bridging the gap between traditional nutritional research and smartphone sensing. Our study, involving 84 college students in Mexico, introduced the novel concept of inferring self-perceived food consumption levels using passive smartphone sensing data and personalized machine learning models. By uncovering associations between sociability, activity types, and food consumption, we presented a potential direction to improve context-aware mobile food diaries. This research has the potential to be used as part of user-friendly food diary applications

that rely on smartphone data to infer food consumption levels, reducing user burden and increasing data quality for self-tracking tools, personalized interventions, and public health studies.

Chapter 5 again focused on eating behavior. Hence, we demonstrated how smartphone sensors can distinguish between eating and non-eating events as standalone devices. Through meticulous analysis of data from 58 college students, we established that time of day and various sensor modalities can serve as indicators of eating events. We also demonstrate the feasibility of leveraging these relationships to infer eating events with personalized machine learning models. Our work not only contributes to scalable sensing-based eating studies but also opens doors for interventions related to healthy eating practices, effectively supporting users in monitoring their food consumption in mobile health applications, including mobile food diaries.

Chapter 6 delved into eating behavior and its intricate interplay with mood, an under-explored area within mobile sensing. By leveraging data from college students, we found that generic mood inference models trained with data collected for varying contexts struggle to generalize to specific contexts like eating occasions. Further, this chapter also underscored the need for personalized approaches for mood inference, especially when dealing with limited label settings. Through a novel community-based personalization technique, we demonstrated the feasibility of inferring mood-while-eating with reasonable accuracy, contributing to context-aware and personalized mobile health applications.

After exploring a situation where model generalization comes into question, the thesis studied model generalization further in Chapter 7, where we looked into the problem of cross-country generalization. Utilizing data from eight countries, we examined the performance of mood inference models in diverse geographical contexts. We also introduced a coherent approach to understanding generalization for different source-target domain pairs with country-specific, country-agnostic, and multi-country approaches. The findings revealed the challenges of cross-country generalization and highlighted the importance of country-specific models, emphasizing the significance of considering geographical diversity when developing mobile sensing models. We also emphasized how models underperform in an unseen country, even after personalizing the model to an extent with re-training or fine-tuning. We also highlighted the need for domain adaptation techniques, focusing on multimodal mobile sensing data, due to distributional shifts across domains.

In Chapter 8, we used the same analytical framework as in the previous chapter and extended our understanding of human behavior recognition to encompass complex daily activities. Through the analysis of data from five countries, we defined a 12-class recognition task, emphasizing the significance of detecting complex daily activities in today's evolving lifestyles. The chapter not only showcased the potential of multimodal smartphone sensors, but also underscored the importance of considering geographical diversity for context-aware applications. Furthermore, we emphasized the need for domain adaptation and generalization in multimodal sensing settings in tackling distribution shifts, similar to the previous chapter.

Finally, Chapter 9 introduced a novel architectural solution to address the issue of generalization and distribution shifts in mobile sensing models. Building upon the lessons learned from previous chapters, we proposed the M3BAT architecture for unsupervised domain adaptation in multimodal mobile sensing. Our extensive experiments across various domains, including both classification and regression tasks, underscored the effectiveness of M3BAT in adapting models to unseen domains, and represent a first step towards deploying these models in diverse real-world scenarios for better model generalization.

10.2 Limitations and Future Work

In all the chapters, we discussed many limitations and future directions specific to those chapters. Here, we summarize some commonly emerging areas.

10.2.1 Data Imbalance, Diversity, and Sampling Biases

Addressing multifaceted challenges such as data imbalance, diversity, and sampling biases is pivotal to advancing the robustness and fairness of multimodal mobile sensing models. While this thesis briefly looked at model personalization with limited labeled data availability (Chapter 6), this thesis did not specifically look into data imbalance. Hence, we used available techniques such as SMOTE, undersampling, and oversampling to handle the situation across many sections. However, researchers should delve into advanced data augmentation techniques for mobile sensing. Generative Adversarial Networks (GANs) and self-training algorithms [157], for instance, can be explored to deal with imbalanced datasets. These techniques generate synthetic data instances to supplement underrepresented classes, improving the model's ability to generalize across diverse behavioral patterns.

Furthermore, fostering collaborations with research groups worldwide is crucial. A global approach to data collection will facilitate the acquisition of datasets that transcend geographical and cultural boundaries. While such efforts are being orchestrated in clinical domains such as Sepsis prediction [351, 311, 499], in mobile sensing, much more coordination and discussion is needed to initiate such an effort. The large-scale data collection campaigns in multiple countries, with the WeNet project (Chapter 2) is a testament to studies in this direction. It is essential to acknowledge that behaviors and emotional cues can significantly differ across countries and other situated contexts. By collaborating with researchers from different regions, we can ensure that mobile sensing models are not only sensitive to these differences but also respectful of cultural nuances. Hence, diversity in data collection is essential, as highlighted in Chapter 7 and Chapter 8.

Beyond country diversity, the scope of data collection should extend to cover diverse contextual settings. As highlighted in Chapter 6, the contextual underpinning of data collection could have a large effect on model generalization. Even though this aspect has rarely been discussed in prior work, it is imperative that future work look into this crucial aspect. Moreover, by considering different age groups, varying lifestyles, and groups of individuals with different conditions, researchers can capture a more holistic view of human behavior. This multifaceted approach to data diversity will not only enhance the generalizability of behavioral analysis models but also contribute to their ethical and equitable applications in real-world scenarios.

While several steps were taken in this thesis to enhance data quality and reduce biases in collecting self-reports, more work is needed to mitigate sampling biases. Researchers should explore visual or audio-based verification methods, such as image, video, or audio analysis, to validate self-reported data points effectively. Such techniques might also require privacy protection considerations. This approach enhances the credibility of collected data, especially in scenarios where self-reports are prone to exaggeration or underreporting. Furthermore, the development of techniques for passive verification of self-reports through sensor data analysis is crucial. This will minimize the reliance on user-generated reports, reducing the potential for biased or inaccurate data for both model training and validation.

10.2.2 Privacy and Ethical Considerations

Privacy and ethical considerations present significant challenges and opportunities for future work in this field. As mobile sensing-based behavioral understanding delves deeper into understanding human actions and emotions, the potential invasion of privacy escalates. Striking a delicate balance between extracting valuable insights and respecting individuals' privacy is essential. Future research should emphasize the development of robust anonymization and aggregation techniques to ensure data is utilized in a way that maintains privacy. Moreover, establishing clear ethical guidelines and standards for data collection, storage, and usage is paramount. This includes obtaining informed consent, transparently communicating data usage, and implementing mechanisms for individuals to have control over their data. Addressing these concerns will not only fortify the credibility of behavioral analysis research but also ensure that the benefits of this technology are ethically grounded and widely accepted.

10.2.3 Generalization, Distribution Shift, and Cross-Cultural Analysis

Overcoming the multifaceted challenges of generalization, distribution shift, and cross-cultural analysis is crucial for the practical deployment of multimodal mobile sensing systems. While we presented M3BAT architecture in Chapter 9, researchers should dedicate more efforts to developing multimodal domain adaptation techniques tailored explicitly for sensing data. These techniques should emphasize joint alignment strategies for multiple data modalities, ensuring that models adapt seamlessly to changing domains.

Understanding the temporal dynamics of distribution shift is also essential. This has been initially examined by Xu et al. [569], in a recent study. Hence, longitudinal studies that examine individuals over extended periods will help uncover how behavior evolves over time and its implications for domain adaptation. These longitudinal datasets are particularly valuable for healthcare applications, where tracking behavioral changes over time is vital for early diagnosis and intervention.

As we have shown in Chapter 7 and Chapter 8, cultural nuances can also impact distribution shifts to a considerable extent. Therefore, conducting cross-cultural studies and integrating cultural factors into domain adaptation models will be vital to addressing this limitation comprehensively. Hence, future models should not only adapt to changes in behavior but also be sensitive to cultural variations, ensuring their relevance and effectiveness across diverse populations.

10.2.4 Feature Extraction and Interpretability

Enhancing feature interpretability is pivotal for the practical utility of multimodal mobile sensing models, especially when moving toward clinical relevance in certain applications. In this thesis, we limited ourselves to typical feature engineering techniques aligning with prior studies. However, in future work, researchers should invest in sophisticated feature engineering methods that capture both high-level behavioral aspects and low-level sensor data characteristics. Feature engineering can encompass the extraction of meaningful features from raw sensor data, such as identifying patterns in accelerometer data indicative of specific activities or behavioral states. In another direction, with the emergence of deep learning, many studies are also exploring the use of raw time series data from multimodal sensors for automatic feature extraction [569, 315, 85]. While they work effectively for a limited number of sensing modalities, handling a large number of modalities, as we commonly encounter in mobile sensing, remains challenging. Hence, more work could also be done in this

direction while also ensuring interpretability.

10.2.5 Clinical Validity of Prediction Outcomes

Most of the inferences used in this thesis lack clinical validity, and are self-reports. However, this is common practice in the domain of ubiquitous computing where clinical validity is not always the ultimate goal [325]. While such clinical validity is not essential for social context inference, complex daily activity recognition, and energy expenditure estimation because they have value in other dimensions such as user experience and general well-being, ensuring the clinical validity of mood inferences is fundamental for many healthcare applications. Hence, researchers should focus on integrating established clinical instruments for mood assessment into data collection protocols. The inclusion of instruments like the Positive and Negative Affect Schedule (PANAS) [553] or the Patient Health Questionnaire (PHQ-9) [299] (used for quantifying depression) will facilitate a direct comparison between model-based inferences and clinically validated assessments. This ensures that the models provide meaningful and actionable insights to healthcare professionals.

Additionally, expanding mood inference models to encompass both valence and arousal dimensions, as per the circumplex mood model, is an improvement to collecting only the first dimension, as we did here [445]. Valence represents the positivity or negativity of an emotional state, while arousal indicates its intensity. While prior work in mobile sensing focused on both these aspects [285, 468], the cross-country analysis we provided only focused on valence due to ground truth availability. Hence, future work should also look into arousal, which would allow a holistic understanding of mood. Understanding these dual dimensions of mood during various behavioral contexts is crucial for providing more clinically relevant insights, especially in settings where precise mood assessment is critical, such as mental health diagnosis and treatment planning.

10.2.6 Transfer Learning and Personalization

The future of mobile sensing-based behavioral analysis models lies in their adaptability to various contexts and user groups. Transfer learning techniques should be explored to enable models to leverage knowledge from one domain or user group to enhance their performance in another, as discussed in Chapter 4. This can significantly reduce the resource-intensive process of training models from scratch for each unique scenario. Additionally, personalization methods should be at the forefront of research efforts. However, there is a lack of mobile sensing research on model personalization. These methods will allow behavioral analysis models to cater to individual needs and preferences, making them more valuable in real-world applications. By understanding and adapting to the unique behavioral traits and emotional responses of individuals, these models can provide tailored insights and interventions, fostering better mental health, well-being, and personal development. This focus on adaptability and personalization will be central to unlocking the full potential of behavioral analysis in a multitude of contexts, ensuring that it genuinely serves the needs of individuals and diverse user groups.

10.2.7 Scalability and User-Friendliness

Ensuring the scalability and real-world applicability of multimodal mobile sensing-based behavioral analysis models is paramount for their widespread adoption. The models used in this thesis solely focused on server-based training with large datasets. Future efforts should focus on optimizing model

Chapter 10. Conclusion

training and deployment processes to accommodate large-scale datasets efficiently. This includes exploring distributed computing solutions and cloud-based infrastructure to handle the computational demands of training complex models. To this end, Federated Learning has been explored in recent work to allow distributed training while also preserving user privacy [287].

Moreover, user-friendliness should be a key consideration in the design of mobile sensing systems. These systems should be accessible and intuitive for a wide range of users, including healthcare professionals, educators, and individuals seeking self-improvement. User-friendly interfaces and integration into existing platforms or applications will facilitate the seamless adoption of behavioral analysis technology across diverse industries and domains, enhancing its practical utility. This is especially needed since the progress in this direction is limited, except for mobile health applications integrated with smartwatches.

A Appendix

A.1 Feature Groups Used in Different Inference Models (Chapter 5)

Table A.1: Feature Groups Used in Different Inference Models

Table	Feature Group	Features
Table 5.2	F1	screen_on_count, screen_off_count, facebook, whatsapp, googlequicksearchbox, microsoft_launcher, instagram, youtube, chrome, spotify, android_dialer, youtube_music, battery_level, charging_true_count, charging_false_count, charging_ac, charging_usb, charging_unknown, minutes_elapsed, hours_elapsed, weekend, acc_x_bef, acc_y_bef, acc_z_bef, acc_x_aft, acc_y_aft, acc_z_aft, acc_yabs, acc_zabs, acc_xabs_bef, acc_yabs_bef, acc_xabs_aft, acc_yabs_aft, acc_zabs_aft, radius_of_gyration
	F3	screen_on_count, screen_off_count, facebook, whatsapp, googlequicksearchbox, microsoft_launcher, instagram, youtube, chrome, spotify, android_dialer, youtube_music, charging_true_count, charging_false_count, charging_ac, charging_usb, charging_unknown, minutes_elapsed, hours_elapsed, weekend, acc_z_bef, acc_x_aft, acc_z_aft, acc_yabs, radius_of_gyration
Table 5.3	F7	googlequicksearchbox, microsoft_launcher, instagram, youtube, charging_false_count
	F1	screen_on_count, screen_off_count, facebook, whatsapp, googlequicksearchbox, microsoft_launcher, instagram, youtube, chrome, spotify, android_dialer, youtube_music, battery_level, charging_true_count, charging_false_count, charging_ac, charging_usb, charging_unknown, minutes_elapsed, hours_elapsed, weekend, acc_x_bef, acc_y_bef, acc_z_bef, acc_x_aft, acc_y_aft, acc_z_aft, acc_yabs, acc_zabs, acc_xabs_bef, acc_yabs_bef, acc_xabs_aft, acc_yabs_aft, acc_zabs_aft, radius_of_gyration

A.2 WEEE Dataset Features from [18] (Chapter 9)

Table A.2: Summary of the features used in the analysis.

Modality	Description
Accelerometer	Statistical features calculated using tsfresh [95]: sum_values, median, mean, length, standard_deviation, variance, root_mean_square, maximum, absolute_maximum, minimum
Gyroscope	Statistical features calculated using tsfresh [95]: sum_values, median, mean, length, standard_deviation, variance, root_mean_square, maximum, absolute_maximum, minimum
Photoplethysmography	Features derived using HeartPy [169]: bpm, ibi, sdn, sdsd, rmsd, pnn20, pnn50, hr_mad, sd1, sd2, s, sd1/sd2, breathingrate

A.3 Selecting an Initial Threshold for Target Users with MEX dataset (Chapter 6)

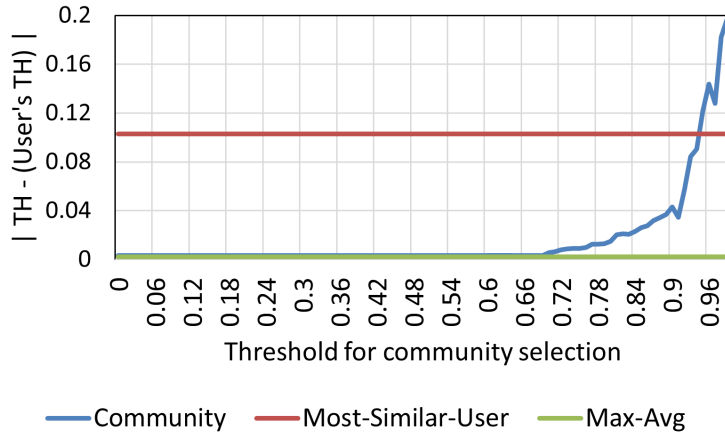


Figure A.1: Averaged threshold distribution difference between target users and three threshold value distributions of MEX dataset

The threshold selection process when discovering the community is a bit time-consuming. For example, we used over twenty th values and trained models and selected the th that gave the highest accuracy. However, in real-world deployments, having a long list of thresholds in the very beginning could make the process of creating a model impractical. Hence, a way to estimate a good threshold is required. Given that each target user has different communities, corresponding to different thresholds (which can vary between 0 and 1), we need a threshold-searching algorithm to find the ideal threshold value. Figure A.1 shows the absolute value of threshold distribution difference between the averaged optimum threshold value of target users (0.77) and three threshold value distributions: (i) Community: averaged threshold values of the community where each user in the community yields the highest accuracy. (ii) Most-Similar-User: averaged threshold values of the most similar user (obtained using the similarity metric) of each target user, where the most similar user yields the highest accuracy (this is a fixed value of 0.88). (iii) Max-Avg: Averaged threshold values corresponding to the maximum accuracy obtained for all the users (this is a fixed value of 0.78 – see Figure 6.4a MAX column). In each of the given three distributions, the one that shows the minimum difference (Max-Avg) can be taken as the first best option when selecting a threshold value for each new user, which can be used to optimize the threshold selection process. The conclusion from this analysis is it is always better to use the Max-Avg of other's thresholds ($th = 0.78$ for this set of users) as a th for any new user as it yields the lowest difference compared to the best threshold that can be obtained for any user. By using this average value as the threshold for any new user, the resource-consuming threshold selection process can be eliminated in the very beginning. In addition, this value provides a reasonably high performance compared to both PLMs and HMs. However, note that this dynamic might change for other datasets and users.

A.4 Algorithm for Data Aggregation (Chapter 6)

Algorithm 1: User-based Data Aggregation

Data:

u : a single user where $u \in U$; U : all the users in the dataset.

f : a single feature where $f \in F$; F : all the features in the dataset.

D_u : the dataset of a user u where $D_u \in D$; D : complete dataset.

Result:

U_{aggr} : user-based data aggregated matrix of size $(|U| \times |F|)$

```

1  $U_{aggr} = [ ]_{|U| \times |F|};$  // Initialize user-based data aggregated matrix
2 for  $u$  in  $U$  do
3    $D_u = GetUserDataMatrix(D, u);$  // Get all the row vectors(1x|F|) of the
   user  $u$ 
4    $u_{aggr} = [ ]_{1 \times |F|};$  // Initialize aggregated user vector
5   for  $f$  in  $F$  do
6      $u_{aggr}[f] = MeanValueOfColumn(D_u, f);$  // Get mean of column  $f$  for each
     user
7   end
8    $U_{aggr}[u] = u_{aggr}$ 
9 end
10 Return( $U_{aggr}$ )

```

A.5 Algorithm for Community Detection (Chapter 6)

Algorithm 2: User-Level-Community Detection

Data: U, u same as Algorithm 1 T : pre-defined threshold array. th : selected threshold value where $th \in T$ U_{aggr} : return value of Algorithm 1 Sim_U : similarity matrix of all the users of size $(|U| \times |U|)$ Sim_{u_t} : similarity vector of the target user u_t of size $(1 \times (|U|-1))$ **Result:** U_{u_t} : community of the target user u_t

```

1  $Sim_U = [ ]_{|U| \times |U|};$  // Initialize user-similarity matrix
2  $Sim_U = CosineSimilarity(U_{aggr});$  // Calculate the similarity metric among all
   user pairs
3  $Sim_{u_t} = GetUserSimilarityVector(Sim_U, u_t);$  // Get similarity vector of the
   target user
4  $U_{u_t} = EmptyArray();$  // Initialize the community matrix of the target user
5 for  $u$  in  $(U - u_t)$  do
6    $Corr_u = Sim_{u_t}[u];$  // Get similarity value between  $u_t$  and  $u$ 
7   if  $Corr_u \geq th$  then
8      $U_{u_t} = Append(U_{u_t}, u);$  // Append user  $u$  to the community of the target
       user
9   end
10 end
11  $Return(U_{u_t})$ 

```

A.6 MEX Dataset Features (Chapter 2)

Table A.3: Summary of Mobile Sensing Features Extracted from Smartphone Sensors

Sensor	Sensor Description
– Acronym (# of features)	Example Features
Location	Using location data, the radius of gyration [581, 37] associated with the one-hour episode was calculated. It is a commonly used metric in UbiComp research. Moreover, in the calculation, location coordinate values were lowered in precision for location privacy reasons (using only four decimal points).
– LOC (1)	radius_of_gyration
Accelerometer	For each ten-minute time window of the day, features that represent the mean of all values and the mean of absolute values (abs) were generated using accelerometer data for axes x, y, and z separately. Using them, values corresponding to the one-hour eating/non-eating event windows were calculated by taking the mean of six ten-minute time bins. Further, by using the half an hour before and after the T_{anc} , more features were generated that correspond to the time before (bef) and after (aft) the eating time [330].
– ACC (18)	considering the one-hour window: $acc_x, acc_y, acc_z, acc_xabs, acc_yabs, acc_zabs$ before and after T_{anc} : $acc_xbef, acc_ybef, acc_zbef, acc_xabs_bef, acc_yabs_bef, acc_zabs_bef, acc_xaft, acc_yaft, acc_zaft, acc_xabs_aft, acc_yabs_aft, acc_zabs_aft$
Application	Prior UbiComp studies have used app usage as a proxy for the behavior of participants [454]. Similarly, the ten most frequently used apps in the dataset were selected. Then, for each hour of consideration, whether each app was used during the episode was derived.
– APP (10)	facebook, whatsapp, instagram, youtube, chrome, spotify, android dialer, youtube music, google quick search box, microsoft launcher
Battery	Battery level and charging state have been used as proxies for smartphone usage behavior, which also represents the behavior of study participants [30, 5]. Hence, the average battery level during the one hour was calculated. In addition, whether the phone is charging or not was detected together with a possible source (ac - alternative current, USB, or unknown).
– BAT (6)	battery_level, charging_true, charging_false, charging_ac, charging_usb, charging_unknown
Screen	Similar to prior work [5, 30], screen-on and screen-off events during the one-hour time window were calculated.
– SCR (2)	screen_on, screen_off
Date and Time	The hour of the day and the minute of the day were derived. In addition, another feature captures whether the day is a weekend or not. Prior work has shown that the behavior (e.g., mobility, food consumption, etc.) of people could differ significantly during weekdays and weekends [326].
– TIME (3)	hours_elapsed, minutes_elapsed, weekend

Bibliography

- [1] *A more personal Health app. For a more informed you.* 2020. URL: <https://www.apple.com/ios/health/>.
- [2] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from [tensorflow.org](https://www.tensorflow.org). 2015.
- [3] S. Abdullah, N. Lane, and T. Choudhury. “Towards population scale activity recognition: A framework for handling data diversity”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 26. 1. 2012.
- [4] S. Abdullah, M. Matthews, E. L. Murnane, G. Gay, and T. Choudhury. “Towards Circadian Computing: “Early to Bed and Early to Rise” Makes Some of Us Unhealthy and Sleep Deprived”. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp '14. Seattle, Washington: Association for Computing Machinery, 2014, 673–684.
- [5] S. Abdullah, E. L. Murnane, M. Matthews, M. Kay, J. A. Kientz, G. Gay, and T. Choudhury. “Cognitive Rhythms: Unobtrusive and Continuous Sensing of Alertness Using a Mobile Phone”. In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp '16. Heidelberg, Germany: ACM, 2016, pp. 178–189.
- [6] S. F. Abraham and P. Beumont. “How patients describe bulimia or binge eating”. In: *Psychological medicine* 12.3 (1982), pp. 625–635.
- [7] D. A. Adler, F. Wang, D. C. Mohr, and T. Choudhury. “Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies”. In: *Plos one* 17.4 (2022), e0266516.
- [8] T. Aflague, C. Boushey, R. Guerrero, Z. Ahmad, D. Kerr, and E. Delp. “Feasibility and Use of the Mobile Food Record for Capturing Eating Occasions among Children Ages 3–10 Years in Guam”. In: *Nutrients* 7.6 (2015), 4403–4415.
- [9] A. F. Agarap. “Deep learning using rectified linear units (relu)”. In: *arXiv preprint arXiv:1803.08375* (2018).
- [10] A. C. Aguiar-Bloemer and R. W. Diez-Garcia. “Influence of emotions evoked by life events on food choice”. en. In: *Eating and Weight Disorders - Studies on Anorexia, Bulimia and Obesity* 23.1 (Feb. 2018), pp. 45–53.

Bibliography

- [11] A. Aguilera, S. M. Schueller, and Y. Leykin. “Daily mood ratings via text message as a proxy for clinic based depression assessment”. In: *Journal of Affective Disorders* 175 (2015), pp. 471–474.
- [12] F. Albinali, S. Intille, W. Haskell, and M. Rosenberger. “Using wearable activity type detection to improve physical activity energy expenditure estimation”. In: *Proceedings of the 12th ACM international conference on Ubiquitous computing*. 2010, pp. 311–320.
- [13] J. Alexander. “Can journaling help you heal from your eating disorder?” In: (2017).
- [14] D. A. Alsaleh, M. T. Elliott, F. Q. Fu, and R. Thakur. “Cross-cultural differences in the adoption of social media”. In: *Journal of Research in Interactive Marketing* (2019).
- [15] N. Alshurafa, J. Jain, R. Alharbi, G. Iakovlev, B. Spring, and A. Pfammatter. “Is More Always Better? Discovering Incentivized MHealth Intervention Engagement Related to Health Behavior Trends”. In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2.4 (Dec. 2018).
- [16] G. Altheimer and H. L. Urry. “Do Emotions Cause Eating? The Role of Previous Experiences and Social Context in Emotional Eating”. In: *Current Directions in Psychological Science* 28.3 (2019), pp. 234–240.
- [17] T. Althoff, R. Sosič, J. L. Hicks, A. C. King, S. L. Delp, and J. Leskovec. “Large-scale physical activity data reveal worldwide activity inequality”. In: *Nature* 547.7663 (2017), pp. 336–339.
- [18] Y. Amarasinghe, D. Sandaruwan, T. Madusanka, I. Perera, and L. Meegahapola. “Multimodal Earable Sensing for Human Energy Expenditure Estimation”. In: *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2023.
- [19] B. Ander, A. Abrahamsson, and D. Bergnehr. “‘It is ok to be drunk, but not too drunk’: party socialising, drinking ideals, and learning trajectories in Swedish adolescent discourse on alcohol use”. In: *Journal of youth studies* 20.7 (2017), pp. 841–854.
- [20] A. H. Andrew, G. Borriello, and J. Fogarty. “Simplifying mobile phone food diaries”. In: *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*. 2013, pp. 260–263.
- [21] *Apple AppStore*, <https://www.apple.com/ios/app-store/>. 2019. URL: <https://www.apple.com/ios/app-store/>.
- [22] Z. Arnold, D. Larose, and E. Agu. “Smartphone inference of alcohol consumption levels from gait”. In: *2015 International Conference on Healthcare Informatics*. IEEE, 2015, pp. 417–426.
- [23] J. Ashurst, I. van Woerden, G. Dunton, M. Todd, P. Ohri-Vachaspati, P. Swan, and M. Bruening. “The Association among Emotions and Food Choices in First-Year College Students Using mobile-Ecological Momentary Assessments”. In: *BMC PUBLIC HEALTH* 18 (2018).
- [24] K. Assi, L. Meegahapola, W. Droz, P. Kun, A. De Götzen, M. Bidoglia, S. Stares, G. Gaskell, A. Chagnaa, A. Ganbold, et al. “Complex daily activities, country-level diversity, and smartphone sensing: A study in denmark, italy, mongolia, paraguay, and uk”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–23.
- [25] S. Attwood, H. Parke, J. Larsen, and K. L. Morton. “Using a mobile health application to reduce alcohol consumption: a mixed-methods evaluation of the drinkaware track & calculate units application”. In: *BMC public health* 17.1 (2017), pp. 1–21.
- [26] M. Atzmueller and K. Hilgenberg. “Towards Capturing Social Interactions with SDCF: An Extensible Framework for Mobile Sensing and Ubiquitous Data Collection”. In: *Proceedings of the 4th International Workshop on Modeling Social Media*. MSM '13. Paris, France: Association for Computing Machinery, 2013.

- [27] M. Azevedo, C. Araújo, F. Reichert, F. Siqueira, M. Silva, and P. Hallal. "Gender differences in leisure-time physical activity". In: *International journal of public health* 52 (Feb. 2007), pp. 8–15.
- [28] N. H. Azrin, M. J. Kellen, J. Brooks, C. Ehle, and V. Vinas. "Relationship Between Rate of Eating and Degree of Satiation". In: *Child & Family Behavior Therapy* 30.4 (2008), pp. 355–364.
- [29] S. Bae, T. Chung, D. Ferreira, A. K. Dey, and B. Suffoletto. "Mobile phone sensors and supervised machine learning to identify alcohol use events in young adults: Implications for just-in-time adaptive interventions". In: *Addictive behaviors* 83 (2018), pp. 42–47.
- [30] S. Bae, D. Ferreira, B. Suffoletto, J. C. Puyana, R. Kurtz, T. Chung, and A. K. Dey. "Detecting Drinking Episodes in Young Adults Using Smartphone-based Sensors". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1.2 (June 2017), 5:1–5:36.
- [31] J. D. Baker, D. A. Williamson, and C. Sylve. "Body image disturbance, memory bias, and body dysphoria: Effects of negative mood induction". In: *Behavior Therapy* 26.4 (1995), pp. 747–759.
- [32] D. Bakker and N. Rickard. "Engagement in mobile phone app for self-monitoring of emotional wellbeing predicts changes in mental health: MoodPrism". In: *JOURNAL OF AFFECTIVE DISORDERS* 227 (2018), 432–442.
- [33] A. Baldominos, A. Cervantes, Y. Saez, and P. Isasi. "A comparison of machine learning and deep learning techniques for activity recognition using mobile devices". In: *Sensors* 19.3 (2019), p. 521.
- [34] Y. P. S. Balhara and S. Mathur. "Alcohol: a major public health problem—South Asian perspective". In: *Addictive Disorders & Their Treatment* 11.2 (2012), pp. 101–120.
- [35] L. Bao and S. S. Intille. "Activity Recognition from User-Annotated Acceleration Data". In: *Pervasive Computing*. Ed. by A. Ferscha and F. Mattern. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 1–17.
- [36] J. E. Bardram and A. Matic. "A decade of ubiquitous computing research in mental health". In: *IEEE Pervasive Computing* 19.1 (2020), pp. 62–72.
- [37] G. Barlacchi, C. Perentis, A. Mehrotra, M. Musolesi, and B. Lepri. "Are you getting sick? Predicting influenza-like symptoms using human mobility behaviors". In: *EPJ Data Science* 6 (Dec. 2017), p. 27.
- [38] N. S. Baron and Y. H. af Segerstad. "Cross-cultural patterns in mobile-phone use: public space and reachability in Sweden, the USA and Japan". In: *New Media & Society* 12.1 (2010), pp. 13–34.
- [39] A. Bauman, G. Ma, F. Cuevas, Z. Omar, T. Waqanivalu, P. Phongsavan, K. Keke, and A. Bhushan. "Cross-national comparisons of socioeconomic differences in the prevalence of leisure-time and occupational physical activity, and active commuting in six Asia-Pacific countries". In: 65.1 (2011), pp. 35–43.
- [40] A. Baumel, F. Muench, S. Edan, J. M. Kane, et al. "Objective user engagement with mental health apps: systematic search and panel-based usage analysis". In: *Journal of medical Internet research* 21.9 (2019), e14567.
- [41] A. Bayat, M. Pomplun, and D. A. Tran. "A Study on Human Activity Recognition Using Accelerometer Data from Smartphones". In: *Procedia Computer Science* 34 (2014). The 9th International Conference on Future Networks and Communications (FNC'14)/The 11th International Conference on Mobile Systems and Pervasive Computing (MobiSPC'14)/Affiliated Workshops, pp. 450–457.
- [42] M. C. Becht and A. J. Vingerhoets. "Crying and mood change: A cross-cultural study". In: *Cognition & Emotion* 16.1 (2002), pp. 87–101.

Bibliography

- [43] K. H. Beck, A. M. Arria, K. M. Caldeira, K. B. Vincent, K. E. O'Grady, and E. D. Wish. "Social context of drinking and alcohol problems among college students". In: *American journal of health behavior* 32.4 (2008), pp. 420–430.
- [44] A. Bedri, R. Li, M. Haynes, R. P. Kosaraju, I. Grover, T. Prioleau, M. Y. Beh, M. Goel, T. Starner, and G. Abowd. "EarBit: using wearable sensors to detect eating episodes in unconstrained environments". In: *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1.3 (2017), pp. 1–20.
- [45] E. J. Bedrick. "Biserial Correlation". In: *Encyclopedia of Biostatistics*. American Cancer Society, 2005.
- [46] M. H. Bekker, C. Van De Meerendonk, and J. Mollerus. "Effects of negative mood induction and impulsivity on self-perceived emotional eating". In: *International Journal of Eating Disorders* 36.4 (2004), pp. 461–469.
- [47] B. M. Bell, R. Alam, N. Alshurafa, E. Thomaz, A. S. Mondol, K. de la Haye, J. A. Stankovic, J. Lach, and D. Spruijt-Metz. "Automatic, wearable-based, in-field eating detection approaches for public health research: a scoping review". In: *NPJ digital medicine* 3.1 (2020), pp. 1–14.
- [48] J. Benesty, J. Chen, Y. Huang, and I. Cohen. "Pearson Correlation Coefficient". In: *Noise Reduction in Speech Processing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 1–4.
- [49] J. Bennett, G. Greene, and D. Schwartz-Barcott. "Perceptions of emotional eating behavior. A qualitative study of college students". In: *Appetite* 60 (2013), pp. 187–192.
- [50] L. Beretta and A. Santaniello. "Nearest neighbor imputation algorithms: a critical evaluation". In: *BMC medical informatics and decision making* 16.3 (2016), pp. 197–208.
- [51] E. M. Berke, T. Choudhury, S. Ali, and M. Rabbi. "Objective Measurement of Sociability and Activity: Mobile Sensing in the Community". In: *The Annals of Family Medicine* 9.4 (2011), pp. 344–350.
- [52] M. Bersamin, S. Lipperman-Kreda, C. Mair, J. Grube, and P. Gruenewald. "Identifying strategies to limit youth drinking in the home". In: *Journal of studies on alcohol and drugs* 77.6 (2016), pp. 943–949.
- [53] S. Bi, T. Wang, N. Tobias, J. Nordrum, S. Wang, G. Halvorsen, S. Sen, R. Peterson, K. Odame, K. Caine, et al. "Auracle: Detecting eating episodes with an ear-mounted sensor". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.3 (2018), pp. 1–27.
- [54] S. J. Biddle et al. "Emotion, mood and physical activity". In: *Physical activity and psychological well-being* 63 (2000).
- [55] J.-I. Biel, N. Martin, D. Labbe, and D. Gatica-Perez. "Bites'N'Bits: Inferring Eating Behavior from Contextual Mobile Data". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1.4 (Jan. 2018), 125:1–125:33.
- [56] A. J. Birney, R. Gunn, J. K. Russell, and D. V. Ary. "MoodHacker Mobile Web App With Email for Adults to Self-Manage Mild-to-Moderate Depression: Randomized Controlled Trial". In: *JMIR mHealth uHealth* 4.1 (2016), e8.
- [57] C. A. Bisogni, L. W. Falk, E. Madore, C. E. Blake, M. Jastran, J. Sobal, and C. M. Devine. "Dimensions of everyday eating and drinking episodes". In: *Appetite* 48.2 (2007), pp. 218–231.
- [58] U. Blanke and B. Schiele. "Daily Routine Recognition through Activity Spotting". In: *Location and Context Awareness*. Ed. by T. Choudhury, A. Quigley, T. Strang, and K. Sugiuma. Vol. 5561. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 192–206.

- [59] H. Blunck, N. O. Bouvin, T. Franke, K. Grønbaek, M. B. Kjaergaard, P. Lukowicz, and M. Wüstenberg. "On heterogeneity in mobile sensing applications aiming at representative data collection". In: *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. 2013, pp. 1087–1098.
- [60] A. Bogomolov, B. Lepri, and F. Pianesi. "Happiness Recognition from Mobile Phone Data". In: *2013 International Conference on Social Computing*. IEEE. 2013, pp. 790–795.
- [61] A. Bogomolov, B. Lepri, M. Ferron, F. Pianesi, and A. S. Pentland. "Daily Stress Recognition from Mobile Phone Data, Weather Conditions and Individual Traits". In: *Proceedings of the 22Nd ACM International Conference on Multimedia*. MM '14. Orlando, Florida, USA: ACM, 2014, pp. 477–486.
- [62] M. Böhmer, B. Hecht, J. Schöning, A. Krüger, and G. Bauer. "Falling asleep with Angry Birds, Facebook and Kindle: a large scale study on mobile application usage". In: *Proceedings of the 13th international conference on Human computer interaction with mobile devices and services*. 2011, pp. 47–56.
- [63] P. Bongers, A. de Graaff, and A. Jansen. "'Emotional' does not even start to cover it: Generalization of overeating in emotional eaters". In: *Appetite* 96 (2016), pp. 611–616.
- [64] P. Bongers, A. Jansen, R. Havermans, A. Roefs, and C. Nederkoorn. "Happy eating. The underestimated role of overeating in a positive mood". In: *Appetite* 67 (2013), pp. 74–80.
- [65] L. Bossard, M. Guillaumin, and L. Van Gool. "Food-101 – Mining Discriminative Components with Random Forests". In: *Computer Vision – ECCV 2014*. Ed. by D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars. Cham: Springer International Publishing, 2014.
- [66] N. Bostrom and E. Yudkowsky. "The ethics of artificial intelligence". In: *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 57–69.
- [67] M. Boukhechba, A. R. Daros, K. Fua, P. I. Chow, B. A. Teachman, and L. E. Barnes. "DemoniSalmon: Monitoring mental health and social interactions of college students using smartphones". In: *Smart Health* 9-10 (2018). CHASE 2018 Special Issue, pp. 192–203.
- [68] E. Bouton-Bessac, L. Meegahapola, and D. Gatica-Perez. "Your Day in Your Pocket: Complex Activity Recognition from Smartphone Accelerometers". In: *Proceedings of the 16th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)* (2022).
- [69] A. P. Bradley. "The use of the area under the ROC curve in the evaluation of machine learning algorithms". In: *Pattern recognition* 30.7 (1997), pp. 1145–1159.
- [70] L. Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32.
- [71] L. Brown, B. Grundlehner, and J. Penders. "Towards wireless emotional valence detection from EEG". In: *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2011, pp. 2188–2191.
- [72] H. Bruch. "Psychological Aspects of Overeating And Obesity". In: *Psychosomatics* 5.5 (1964), pp. 269–274.
- [73] M. N. Burns, M. Begale, J. Duffecy, D. Gergle, C. J. Karr, E. Giangrande, and D. C. Mohr. "Harnessing Context Sensing to Develop a Mobile Intervention for Depression". In: *J Med Internet Res* 13.3 (2011), e55.
- [74] Y. S. Can, B. Arnrich, and C. Ersoy. "Stress detection in daily life scenarios using smart phones and wearable sensors: A survey". In: *Journal of biomedical informatics* 92 (2019), p. 103139.

Bibliography

- [75] L. Canetti, E. Bachar, and E. M. Berry. “Food and emotion”. In: *Behavioural processes* 60.2 (2002), pp. 157–164.
- [76] N. Canton. *Cell phone culture: How cultural differences affect mobile use*. Sept. 2012.
- [77] L. Canzian and M. Musolesi. “Trajectories of Depression: Unobtrusive Monitoring of Depressive States by Means of Smartphone Mobility Traces Analysis”. In: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp '15. Osaka, Japan: ACM, 2015, pp. 1293–1304.
- [78] B. Cao, L. Zheng, C. Zhang, P. S. Yu, A. Piscitello, J. Zulueta, O. Ajilore, K. Ryan, and A. D. Leow. “DeepMood: Modeling Mobile Phone Typing Dynamics for Mood Detection”. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '17. Halifax, NS, Canada: ACM, 2017, pp. 747–755.
- [79] G. Carrà, C. Crocamo, F. Bartoli, D. Carretta, A. Schivalocchi, P. E. Bebbington, and M. Clerici. “Impact of a mobile E-Health intervention on binge drinking in young people: The digital-alcohol risk alertness notifying network for adolescents and young adults project”. In: *Journal of Adolescent Health* 58.5 (2016), pp. 520–526.
- [80] E. A. Carroll, M. Czerwinski, A. Roseway, A. Kapoor, P. Johns, K. Rowan, and M. C. Schraefel. “Food and Mood: Just-in-Time Support for Emotional Eating”. In: *Humaine Association Conference on Affective Computing and Intelligent Interaction*. 2013, pp. 252–257.
- [81] G. S. Castell, L. Serra-Majem, and L. Ribas-Barba. “What and how much do we eat? 24-hour dietary recall method”. In: *Nutricion hospitalaria* 31.3 (2015), pp. 46–48.
- [82] D. Castelvechi. “Is facial recognition too biased to be let loose?” In: *Nature* 587.7834 (2020), pp. 347–350.
- [83] S. Chan, L. Li, J. Torous, D. Gratzler, and P. M. Yellowlees. “Review of Use of Asynchronous Technologies Incorporated in Mental Health Care”. en. In: *Current Psychiatry Reports* 20.10 (Oct. 2018), p. 85.
- [84] T.-C. Chan, T.-J. Yen, Y.-C. Fu, and J.-S. Hwang. “ClickDiary: Online Tracking of Health Behaviors and Mood”. In: *JOURNAL OF MEDICAL INTERNET RESEARCH* 17.6 (2015).
- [85] Y. Chang, A. Mathur, A. Isopoussu, J. Song, and F. Kawsar. “A systematic study of unsupervised domain adaptation for robust human-activity recognition”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.1 (2020), pp. 1–30.
- [86] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. “SMOTE: Synthetic Minority Over-sampling Technique”. In: *Journal of Artificial Intelligence Research* 16 (2002), 321–357.
- [87] T. Chen and C. Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *CoRR* abs/1603.02754 (2016), pp. 785–794.
- [88] Y. Chen, X. Qin, J. Wang, C. Yu, and W. Gao. “Fedhealth: A federated transfer learning framework for wearable healthcare”. In: *IEEE Intelligent Systems* 35.4 (2020), pp. 83–93.
- [89] M. Chiva. “Cultural aspects of meals and meal frequency”. In: *British Journal of Nutrition* 77.S1 (1997), S21–S28.
- [90] J.-H. Choi, M. Constantinides, S. Joglekar, and D. Quercia. “KAIROS: Talking heads and moving bodies for successful meetings”. In: *Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications*. 2021, pp. 30–36.
- [91] F. Chollet. *keras*. <https://github.com/fchollet/keras>. 2015.

- [92] V. Chotpitayasunondh and K. M. Douglas. “How “phubbing” becomes the norm: The antecedents and consequences of snubbing via smartphone”. In: *Computers in human behavior* 63 (2016), pp. 9–18.
- [93] T. Choudhury, G. Borriello, S. Consolvo, D. Haehnel, B. Harrison, B. Hemingway, J. Hightower, P. P. Klasnja, K. Koscher, A. LaMarca, J. A. Landay, L. LeGrand, J. Lester, A. Rahimi, A. Rea, and D. Wyatt. “The Mobile Sensing Platform: An Embedded Activity Recognition System”. In: *IEEE Pervasive Computing* 7.2 (Apr. 2008), pp. 32–41.
- [94] P. I. Chow, K. Fua, Y. Huang, W. Bonelli, H. Xiong, L. E. Barnes, and B. A. Teachman. “Using Mobile Sensing to Test Clinical Models of Depression, Social Anxiety, State Affect, and Social Isolation Among College Students”. In: *JOURNAL OF MEDICAL INTERNET RESEARCH* 19.3 (2017).
- [95] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr. “Time series feature extraction on basis of scalable hypothesis tests (tsfresh—a python package)”. In: *Neurocomputing* 307 (2018), pp. 72–77.
- [96] L. Christensen. “Effects of eating behavior on mood: a review of the literature”. In: *International Journal of Eating Disorders* 14.2 (1993), pp. 171–183.
- [97] J. Chua, S. Touyz, and A. Hill. “Negative mood-induced overeating in obese binge eaters: An experimental study”. In: *International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity* 28 (May 2004), pp. 606–10.
- [98] K. S. Chun, S. Bhattacharya, and E. Thomaz. “Detecting eating episodes by tracking jawbone movements with a non-contact wearable sensor”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.1 (2018), pp. 1–21.
- [99] S. Chung, J. Lim, K. J. Noh, G. Kim, and H. Jeong. “Sensor data acquisition and multimodal sensor fusion for human activity recognition using deep learning”. In: *Sensors* 19.7 (2019), p. 1716.
- [100] D. G. Cobian, N. S. Daehn, P. A. Anderson, and B. C. Heiderscheidt. “Active cervical and lumbar range of motion during performance of activities of daily living in healthy young adults”. In: *Spine* 38.20 (2013), pp. 1754–1763.
- [101] J. Cohen. *Statistical Power Analysis for the behavioral sciences*. L. Erlbaum Associates, 1988.
- [102] L. Cohen, G. C. Curhan, and J. P. Forman. “Influence of Age on the Association between Lifestyle Factors and Risk of Hypertension”. In: *J Am Soc Hypertens* (2012).
- [103] L. Coles-Kemp, R. B. Jensen, and R. Talhouk. “In a New Land: Mobile Phones, Amplified Pressures and Reduced Capabilities”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI ’18. Montreal QC, Canada: ACM, 2018, 584:1–584:13.
- [104] *Commuter health*. 2020. URL: <https://www.apple.com/ios/health/>.
- [105] S. Consolvo, D. W. McDonald, T. Toscos, M. Y. Chen, J. Froehlich, B. Harrison, P. Klasnja, A. LaMarca, L. LeGrand, R. Libby, I. Smith, and J. A. Landay. “Activity Sensing in the Wild: A Field Trial of Ubifit Garden”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’08. Florence, Italy: ACM, 2008, pp. 1797–1806.
- [106] M. Constantinides, J. Busk, A. Matic, M. Faurholt-Jepsen, L. V. Kessing, and J. E. Bardram. “Personalized versus generic mood prediction models in bipolar disorder”. In: *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. 2018, pp. 1700–1707.

Bibliography

- [107] L. Coorevits and T. Coenen. “The rise and fall of wearable fitness trackers”. en. In: *Academy of Management Proceedings* 2016.1 (Jan. 2016), p. 17305.
- [108] V. P. Cornet and R. J. Holden. “Systematic review of smartphone-based passive sensing for health and wellbeing”. In: *Journal of Biomedical Informatics* 77 (2018), pp. 120–132.
- [109] S. S. Coughlin and J. Stewart. “Use of consumer wearable devices to promote physical activity: a review of health intervention studies”. In: *Journal of environment and health sciences* 2.6 (2016).
- [110] R. Cowie. “Ethical issues in affective computing”. In: *The Oxford handbook of affective computing* (2015), pp. 334–348.
- [111] M. Cox, K Sewell, K. Egan, S Baird, C Eby, K Ellis, and J Kuteh. “A systematic review of high-risk environmental circumstances for adolescent drinking”. In: *Journal of Substance Use* 24.5 (2019), pp. 465–474.
- [112] D. Crane, C. Garnett, S. Michie, R. West, and J. Brown. “A smartphone app to reduce excessive alcohol consumption: Identifying the effectiveness of intervention components in a factorial randomised control trial”. In: *Scientific reports* 8.1 (2018), p. 4384.
- [113] A. Crosby, J. Gfroerer, B. Han, L. Ortega, and S. E. Parks. “Suicidal thoughts and behaviors among adults aged 18 Years–United States, 2008–2009”. In: (2011).
- [114] T. Cruwys, K. E. Bevelander, and R. C. Hermans. “Social modeling of eating: A review of when and why social influence affects food intake and choice”. In: *Appetite* 86 (2015). Social Influences on Eating, pp. 3–18.
- [115] A. Cultures. *Neurips 2022 workshop on AI Cultures*. 2022. URL: <https://ai-cultures.github.io/>.
- [116] A. Cuthbertson. *Self-driving cars are be more likely to drive into black people, study claims*. 2019.
- [117] A. Cutler, D. Cutler, and J. Stevens. “Random Forests”. In: vol. 45. Jan. 2011, pp. 157–176.
- [118] B. Cvetković, B. Kaluža, R. Milić, and M. Luštrek. “Towards human energy expenditure estimation using smart phone inertial sensors”. In: *Ambient Intelligence: 4th International Joint Conference, Aml 2013, Dublin, Ireland, December 3-5, 2013. Proceedings 4*. Springer. 2013, pp. 94–108.
- [119] J. Dai, J. Teng, X. Bai, Z. Shen, and D. Xuan. “Mobile phone based drunk driving detection”. In: *2010 4th International Conference on Pervasive Computing Technologies for Healthcare*. IEEE. 2010, pp. 1–8.
- [120] K. L. Dannecker, N. A. Sazonova, E. L. Melanson, E. S. Sazonov, and R. C. Browning. “A comparison of energy expenditure estimation of several physical activity monitors”. In: *Medicine and science in sports and exercise* 45.11 (2013), p. 2105.
- [121] V.-A. Darvariu, L. Convertino, A. Mehrotra, and M. Musolesi. “Quantifying the relationships between everyday objects and emotional states through deep learning based image analysis using smartphones”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.1 (2020), pp. 1–21.
- [122] B. V. Dasarathy. “Sensor fusion potential exploitation-innovative architectures and illustrative applications”. In: *Proceedings of the IEEE* 85.1 (1997), pp. 24–38.
- [123] E. L. Davies, A. J. Lonsdale, S. E. Hennesly, A. R. Winstock, and D. R. Foxcroft. “Personalized digital interventions showed no impact on risky drinking in young adults: a pilot randomized controlled trial”. In: *Alcohol and alcoholism* 52.6 (2017), pp. 671–676.
- [124] M. De Nadai, A. Cardoso, A. Lima, B. Lepri, and N. Oliver. “Strategies and limitations in app usage and human mobility”. In: *Scientific Reports* 9 (July 2019), p. 10935.

- [125] J. B. De Wit, F. M. Stok, D. J. Smolenski, D. D. de Ridder, E. de Vet, T. Gaspar, F. Johnson, L. Nureeva, and A. Luszczynska. "Food culture in the home environment: Family meal practices and values can support healthy eating and self-regulation in young people in four European countries". In: *Applied Psychology: Health and Well-Being* 7.1 (2015), pp. 22–40.
- [126] M. B. Del Rosario, K. Wang, J. Wang, Y. Liu, M. Brodie, K. Delbaere, N. H. Lovell, S. R. Lord, and S. J. Redmond. "A comparison of activity classification in younger and older cohorts using a smartphone". In: *Physiological Measurement* 35.11 (Nov. 2014), pp. 2269–2286.
- [127] C. Deng, W. Lin, X. Ye, Z. Li, Z. Zhang, and G. Xu. "Social media data as a proxy for hourly fine-scale electric power consumption estimation". In: *Environment and Planning A: Economy and Space* 50.8 (2018), pp. 1553–1557.
- [128] Y. Deng. "Deep Learning on Mobile Devices - A Review". In: *CoRR* abs/1904.09274 (2019).
- [129] T. Denning, A. Andrew, R. Chaudhri, C. Hartung, J. Lester, G. Borriello, and G. Duncan. "BAL-ANCE: Towards a Usable Pervasive Wellness Application with Accurate Activity Inference". In: *Proceedings of the 10th Workshop on Mobile Computing Systems and Applications*. HotMobile '09. Santa Cruz, California: Association for Computing Machinery, 2009.
- [130] S. Dernbach, B. Das, N. C. Krishnan, B. L. Thomas, and D. J. Cook. "Simple and Complex Activity Recognition through Smart Phones". In: *2012 Eighth International Conference on Intelligent Environments*. Guanajuato, Mexico: IEEE, June 2012, pp. 214–221.
- [131] G. Deubler and M. Swaney-Stueve. "The K-State emoji scale, a cross-cultural validation with adults". In: *Journal of Sensory Studies* 35.4 (2020), e12573.
- [132] D. M. Dick, J. L. Pagan, C. Holliday, R. Viken, L. Pulkkinen, J. Kaprio, and R. J. Rose. "Gender differences in friends' influences on adolescent drinking: A genetic epidemiological study". In: *Alcoholism: clinical and experimental research* 31.12 (2007), pp. 2012–2019.
- [133] E. Diener, S. Kanazawa, E. M. Suh, and S. Oishi. "Why people are in a generally good mood". In: *Personality and Social Psychology Review* 19.3 (2015), pp. 235–256.
- [134] V. Dissanayake, S. Seneviratne, R. Rana, E. Wen, T. Kaluarachchi, and S. Nanayakkara. "SigRep: Toward Robust Wearable Emotion Recognition With Contrastive Representation Learning". In: *IEEE Access* 10 (2022), pp. 18105–18120.
- [135] T. M. T. Do, J. Blom, and D. Gatica-Perez. "Smartphone Usage in the Wild: A Large-scale Analysis of Applications and Context". In: *Proceedings of the 13th International Conference on Multimodal Interfaces*. ICMI '11. Alicante, Spain: ACM, 2011, pp. 353–360.
- [136] T.-M.-T. Do and D. Gatica-Perez. "By Their Apps You Shall Understand Them: Mining Large-scale Patterns of Mobile Phone Usage". In: *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*. MUM '10. Limassol, Cyprus: ACM, 2010, 27:1–27:10.
- [137] Y. Dong, J. Scisco, M. Wilson, E. Muth, and A. Hoover. "Detecting periods of eating during free-living by tracking wrist motion". In: *IEEE journal of biomedical and health informatics* 18.4 (2013), pp. 1253–1260.
- [138] M. B. Donnellan, F. L. Oswald, B. M. Baird, and R. E. Lucas. "The mini-IPIP scales: tiny-yet-effective measures of the Big Five factors of personality." In: *Psychological assessment* 18.2 (2006), p. 192.
- [139] M Dubad, C Winsper, C Meyer, M Livanou, and S. Marwaha. "A systematic review of the psychometric properties, usability and clinical impacts of mobile mood-monitoring applications in young people". In: *Psychological medicine* 48.2 (2018), pp. 208–228.

Bibliography

- [140] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri. “Activation functions in deep learning: A comprehensive survey and benchmark”. In: *Neurocomputing* (2022).
- [141] P. Dulin, V. Gonzalez, D. King, D. Giroux, and S. Bacon. “Smartphone-Based, Self-Administered Intervention System for Alcohol Use Disorders: Theory and Empirical Evidence Basis”. In: *Alcoholism treatment quarterly* 31 (July 2013).
- [142] A. Dy, N.-N. J. Nguyen, S. H. Mirjahanmardir, D. Androutsos, M. Dawe, A. Fyles, W. Shi, F.-F. Liu, S. Done, and A. Khademi. “Domain Adaptation using Silver Standard Labels for Ki-67 Scoring in Digital Pathology A Step Closer to Widescale Deployment”. In: *Medical Imaging with Deep Learning*. 2023.
- [143] N. Eagle and A. (Sandy) Pentland. “Reality mining: sensing complex social systems”. en. In: *Personal and Ubiquitous Computing* 10.4 (2006), pp. 255–268.
- [144] H. Elsahar and M. Gallé. “To annotate or not? predicting performance drop under domain shift”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 2163–2173.
- [145] N. C. Engard. “LimeSurvey <http://limesurvey.org>: Visited: Summer 2009”. In: (2009).
- [146] W. F. Epling and W. D. Pierce. “Activity-based anorexia: A biobehavioral perspective”. In: *International Journal of Eating Disorders* 7.4 (1988), pp. 475–485.
- [147] C. G. Esteban Ortiz-Ospina and M. Roser. “Time Use”. In: *Our World in Data* (2020).
- [148] C. Evers, M. Adriaanse, D. T. de Ridder, and J. C. de Witt Huberts. “Good mood food. Positive emotion as a neglected trigger for food intake”. In: *Appetite* 68 (2013), pp. 1–7.
- [149] L. J. Faherty, L. Hantsoo, D. Appleby, M. D. Sammel, I. M. Bennett, and D. J. Wiebe. “Movement patterns in women at risk for perinatal depression: use of a mood-monitoring mobile application in pregnancy”. In: *Journal of the American Medical Informatics Association* 24.4 (2017), pp. 746–753.
- [150] A. A. Farhan, C. Yue, R. Morillo, S. Ware, J. Lu, J. Bi, J. Kamath, A. Russell, A. Bamis, and B. Wang. “Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data”. In: *2016 IEEE Wireless Health (WH)*. IEEE. 2016, pp. 1–8.
- [151] J. Farrington, A. J. Moore, N. Tilbury, J. Church, and P. D. Biemond. “Wearable sensor badge and sensor jacket for context awareness”. In: *Digest of Papers. Third International Symposium on Wearable Computers*. 1999, pp. 107–113.
- [152] A. Ferrari, D. Micucci, M. Mobilio, and P. Napolitano. “On the personalization of classification models for human activity recognition”. In: *IEEE Access* 8 (2020), pp. 32066–32079.
- [153] A. E. Field, C. B. Taylor, A. Celio, and G. A. Colditz. “Comparison of self-report to interview assessment of bulimic behaviors among preadolescent and adolescent girls and boys”. In: *International Journal of Eating Disorders* 35.1 (2004), pp. 86–92.
- [154] K. Fisher and J. Robinson. *Daily routines in 22 countries diary evidence of average daily time spent in thirty activities*. Tech. rep. University of Oxford, Centre for Time Use Research., 2010.
- [155] J. C. Franklin, J. D. Ribeiro, K. R. Fox, K. H. Bentley, E. M. Kleiman, X. Huang, K. M. Musacchio, A. C. Jaroszewski, B. P. Chang, and M. K. Nock. “Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research.” In: *Psychological bulletin* 143.2 (2017), p. 187.

- [156] B. Freisthler, S. Lipperman-Kreda, M. Bersamin, and P. J. Gruenewald. "Tracking the when, where, and with whom of alcohol use: Integrating ecological momentary assessment and geospatial data to examine risk for alcohol-related problems". In: *Alcohol research: current reviews* 36.1 (2014), p. 29.
- [157] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan. "Synthetic data augmentation using GAN for improved liver lesion classification". In: *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE. 2018, pp. 289–293.
- [158] B. Friese and J. W. Grube. "Teen parties: who has parties, what predicts whether there is alcohol and who supplies the alcohol?" In: *The journal of primary prevention* 35 (2014), pp. 391–396.
- [159] J. Froehlich, M. Y. Chen, S. Consolvo, B. Harrison, and J. A. Landay. "MyExperience: A System for in Situ Tracing and Capturing of User Feedback on Mobile Phones". In: *Proceedings of the 5th International Conference on Mobile Systems, Applications and Services*. MobiSys '07. San Juan, Puerto Rico: ACM, 2007, pp. 57–70.
- [160] O. Gallupe and M. Bouchard. "Adolescent parties and substance use: A situational approach to peer influence". In: *Journal of Criminal Justice* 41.3 (2013), pp. 162–171.
- [161] Y. Ganin and V. Lempitsky. "Unsupervised domain adaptation by backpropagation". In: *International conference on machine learning*. PMLR. 2015, pp. 1180–1189.
- [162] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. "Domain-adversarial training of neural networks". In: *The journal of machine learning research* 17.1 (2016), pp. 2096–2030.
- [163] Y. Gao, N. Zhang, H. Wang, X. Ding, X. Ye, G. Chen, and Y. Cao. "iHear food: eating detection using commodity bluetooth headsets". In: *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE. 2016, pp. 163–172.
- [164] P. Garcia-Segovia, R. J. Harrington, and H.-S. Seo. "Influences of table setting and eating location on food acceptance and intake". In: *Food Quality and Preference* 39 (2015), pp. 1–7.
- [165] J. A. Álvarez García, B. Cvetković, and M. Luštrek. "A Survey on Energy Expenditure Estimation Using Wearable Devices". In: *ACM Computing Surveys* (2020).
- [166] S. Gashi, C. Min, A. Montanari, S. Santini, and F. Kawsar. "A multidevice and multimodal dataset for human energy expenditure estimation using wearable devices". In: *Scientific Data* 9.1 (2022), p. 537.
- [167] D. Gatica-Perez, J.-I. Biel, D. Labbe, and N. Martin. "Discovering eating routines in context with a smartphone app". In: *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 2019, pp. 422–429.
- [168] S. Gedam and S. Paul. "A review on mental stress detection using wearable sensors and machine learning techniques". In: *IEEE Access* 9 (2021), pp. 84045–84066.
- [169] P. van Gent, H. Farah, N. Nes, and B. Arem. "Heart Rate Analysis for Human Factors: Development and Validation of an Open Source Toolkit for Noisy Naturalistic Heart Rate Data". In: June 2018.
- [170] I. Giannopoulou, M. Kotopoulea-Nikolaidi, S. Daskou, K. Martyn, and A. Patel. "Mindfulness in eating is inversely related to binge eating and mood disturbances in university students in health-related disciplines". In: *Nutrients* 12.2 (2020), p. 396.
- [171] E. L. Gibson. "Emotional influences on food choice: sensory, physiological and psychological pathways". In: *Physiology & behavior* 89.1 (2006), pp. 53–61.

Bibliography

- [172] C. Gilligan, K. Kypri, N. Johnson, M. Lynagh, and S. Love. “Parental supply of alcohol and adolescent risky drinking”. In: *Drug and alcohol review* 31.6 (2012), pp. 754–762.
- [173] F. Giunchiglia. *A Diversity-aware Internet, When Technology Works for People*. (2020). 2020.
- [174] F. Giunchiglia, E. Bignotti, and M. Zeni. “Personal context modelling and annotation”. In: *2017 IEEE international conference on pervasive computing and communications workshops (PerCom workshops)*. IEEE. 2017, pp. 117–122.
- [175] F. Giunchiglia, I. Bison, M. Busso, R. Chenu-Abente, M. Rodas, M. Zeni, C. Gunel, G. Veltri, A. D. Götzen, P. Kun, A. Ganbold, A. Chagnaa, G. Gaskell, S. Stares, M. Bidoglia, L. Cernuzzi, A. Hume, J. L. Zarza, H. Xu, D. Song, S. Diwakar, C. Nutakki, S. R. Correa, A.-R. Mendoza, L. Meegahapola, and D. Gatica-Perez. *A worldwide diversity pilot on daily routines and social practices (2020-2021)*. University of Trento Technical Report - DataScientia dataset descriptors. <https://iris.unitn.it/handle/11572/338382>. 2022.
- [176] T. E. Glasgow, H. T. K. Le, E. S. Geller, Y. Fan, and S. Hankey. “How transport modes, the built and natural environments, and activities influence mood: A GPS smartphone app study”. In: *JOURNAL OF ENVIRONMENTAL PSYCHOLOGY* 66 (2019).
- [177] G. Gmel, J. Gaume, M. Faouzi, J.-P. Kulling, and J.-B. Daeppen. “Who drinks most of the total alcohol in young men—risky single occasion drinking as normative behaviour”. In: *Alcohol & Alcoholism* 43.6 (2008), pp. 692–697.
- [178] P. Goel and A. Ganatra. “Unsupervised Domain Adaptation for Image Classification and Object Detection Using Guided Transfer Learning Approach and JS Divergence”. In: *Sensors* 23.9 (2023), p. 4436.
- [179] J. Gong, Y. Huang, P. I. Chow, K. Fua, M. S. Gerber, B. A. Teachman, and L. E. Barnes. “Understanding behavioral dynamics of social anxiety among college students through smartphone sensors”. In: *Information Fusion* 49 (2019), pp. 57–68.
- [180] T. Gong, Y. Kim, A. Orzikulova, Y. Liu, S. J. Hwang, J. Shin, and S.-J. Lee. “DAPPER: Performance Estimation of Domain Adaptation in Mobile Sensing”. In: *arXiv preprint arXiv:2111.11053* (2021).
- [181] Z. Gong, P. Zhong, and W. Hu. “Diversity in Machine Learning”. In: *IEEE Access* 7 (2019), 64323–64350.
- [182] H. C. Gooding, C. Milliren, C. M. Shay, T. K. Richmond, A. E. Field, and M. W. Gillman. “Achieving Cardiovascular Health in Young Adulthood—Which Adolescent Factors Matter?” In: *Journal of Adolescent Health* 58.1 (2016), pp. 119–121.
- [183] Google. *Adapt your app by understanding what users are doing*. 2022. URL: <https://developers.google.com/location-context/activity-recognition>.
- [184] Google. *Choose a category and tags for your app or game*. 2021. URL: <https://bit.ly/39r8zPp>.
- [185] Google Fit - *Coaching you to a healthier and more active life*. 2020. URL: <https://www.google.com/fit/>.
- [186] Google Playstore. 2019. URL: <https://play.google.com/store?hl=en>.
- [187] V. V. Gouveia, T. L. Milfont, and V. M. Guerra. “Functional theory of human values: Testing its content and structure hypotheses”. In: *Personality and Individual Differences* 60 (2014), pp. 41–47.
- [188] B. R. Greenberg. “Predictors of binge eating in bulimic and nonbulimic women”. In: *International Journal of Eating Disorders* 5.2 (1986), pp. 269–284.

- [189] B. R. Greenberg and P. D. Harvey. "Affective lability versus depression as determinants of binge eating". In: *Addictive Behaviors* 12.4 (1987), pp. 357–361.
- [190] Greenland Sander, Senn Stephen J., Rothman Kenneth J., Carlin John B., Poole Charles, Goodman Steven N., and Altman Douglas G. "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations". In: *European Journal of Epidemiology* 31.4 (2016), pp. 337–350.
- [191] Greenland Sander, Senn Stephen J., Rothman Kenneth J., Carlin John B., Poole Charles, Goodman Steven N., and Altman Douglas G. "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations". In: *European Journal of Epidemiology* 31.4 (2016), pp. 337–350.
- [192] U. Grittner, S. Wilsnack, S. Kuntsche, T. K. Greenfield, R. Wilsnack, A. Kristjanson, and K. Bloomfield. "A multilevel analysis of regional and gender differences in the drinking behavior of 23 countries". In: *Substance use & misuse* 55.5 (2020), pp. 772–786.
- [193] P. J. Grother, M. L. Ngan, K. K. Hanaoka, et al. "Face recognition vendor test part 3: demographic effects". In: (2019).
- [194] D. H. Gustafson, F. M. McTavish, M.-Y. Chih, A. K. Atwood, R. A. Johnson, M. G. Boyle, M. S. Levy, H. Driscoll, S. M. Chisholm, L. Dillenburg, et al. "A smartphone application to support recovery from alcoholism: a randomized clinical trial". In: *JAMA psychiatry* 71.5 (2014), pp. 566–572.
- [195] R. Guthold, M. J. Cowan, C. S. Autenrieth, L. Kann, and L. M. Riley. "Physical Activity and Sedentary Behavior Among Schoolchildren: A 34-Country Comparison". In: *The Journal of Pediatrics* 157.1 (2010), 43–49.e1.
- [196] G. Harari, S. Mueller, C. Stachl, R. Wang, W. Wang, M. Buehner, P. Rentfrow, A. Campbell, and S. Gosling. "Sensing Sociability: Individual Differences in Young Adults' Conversation, Calling, Texting, and App Use Behaviors in Daily Life". In: *Journal of Personality and Social Psychology* 119 (May 2019).
- [197] G. M. Harari, S. D. Gosling, R. Wang, F. Chen, Z. Chen, and A. T. Campbell. "Patterns of behavior change in students over an academic term: A preliminary study of activity and sociability behaviors using smartphone sensing methods". In: *Computers in Human Behavior* 67 (2017), pp. 129–138.
- [198] D. A. Harrison, K. H. Price, and M. P. Bell. "Beyond relational demography: Time and the effects of surface-and deep-level diversity on work group cohesion". In: *Academy of management journal* 41.1 (1998), pp. 96–107.
- [199] T. Hastie, S. Rosset, J. Zhu, and H. Zou. "Multi-class AdaBoost". en. In: *Statistics and Its Interface* 2.3 (2009), pp. 349–360.
- [200] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.
- [201] J. He and F. van de Vijver. "Bias and equivalence in cross-cultural research". In: *Online readings in psychology and culture* 2.2 (2012), pp. 2307–0919.
- [202] J. A. Healey and R. W. Picard. "Detecting stress during real-world driving tasks using physiological sensors". In: *IEEE Transactions on intelligent transportation systems* 6.2 (2005), pp. 156–166.
- [203] HealthLinkBC. *Alcohol and Drug Use in Young People*. 2019.
- [204] HealthLinkBC. *Statistics of Alcohol and Drug Use in Young People*. 2019.

Bibliography

- [205] Heatherton and Baumeister. “Binge eating as escape from self-awareness”. In: *Psychol Bull.* (Sept. 1991), pp. 86–108.
- [206] V. T. van Hees, R. Golubic, U. Ekelund, and S. Brage. “Impact of study design on development and evaluation of an activity-type classifier”. en. In: *Journal of Applied Physiology* 114.8 (Apr. 2013), pp. 1042–1051.
- [207] S. Hemminki, P. Nurmi, and S. Tarkoma. “Gravity and Linear Acceleration Estimation on Mobile Devices”. In: *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. MOBIQUITOUS '14. London, United Kingdom: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2014, 50–59.
- [208] I. M. Henriksen, M. Skaar, and A. Tjora. “The Constitutive Practices of Public Smartphone Use”. In: *Societies* 10.4 (2020).
- [209] C. P. Herman and J. Polivy. “Dieting as an exercise in behavioral economics.” In: *Time and decision: Economic and psychological perspectives on intertemporal choice* (2003).
- [210] C. Herman, J. Polivy, and T. Leone. “6 - The psychology of overeating”. In: *Food, Diet and Obesity*. Ed. by D. J. Mela. Woodhead Publishing Series in Food Science, Technology and Nutrition. Woodhead Publishing, 2005, pp. 115–136.
- [211] M. M. Hetherington. “Cues to overeat: psychological factors influencing overconsumption”. In: *Proceedings of the Nutrition Society* 66.1 (2007), 113–123.
- [212] L. Hides, C. Quinn, W. Cockshaw, S. Stoyanov, O. Zelenko, D. Johnson, D. Tjondronegoro, L.-H. Quek, and D. J. Kavanagh. “Efficacy and outcomes of a mobile app targeting alcohol use in young people”. In: *Addictive behaviors* 77 (2018), pp. 89–95.
- [213] S. Higgs and J. Thomas. “Social influences on eating”. In: *Current Opinion in Behavioral Sciences* 9 (2016). Diet, behavior and brain function, pp. 1–6.
- [214] L. M. Hilty. “Ethical issues in ubiquitous computing—three technology assessment studies revisited”. In: *Ubiquitous Computing in the Workplace*. Springer, 2015, pp. 45–60.
- [215] G. R. J. Hockey, A. John Maule, P. J. Clough, and L. Bdzola. “Effects of negative mood states on risk in everyday decision making”. In: *Cognition & Emotion* 14.6 (2000), pp. 823–855.
- [216] R. Horlings, D. Datcu, and L. J. Rothkrantz. “Emotion recognition using brain activity”. In: *Proceedings of the 9th international conference on computer systems and technologies and workshop for PhD students in computing*. 2008, pp. II–1.
- [217] S. A. Hoseini-Tabatabaei, A. Gluhak, and R. Tafazolli. “A survey on smartphone-based systems for opportunistic user context recognition”. In: *ACM Computing Surveys (CSUR)* 45.3 (2013), pp. 1–51.
- [218] K. Houston, K. Hawton, and R. Shepperd. “Suicide in young people aged 15–24: a psychological autopsy study”. In: *Journal of Affective Disorders* 63.1 (2001), pp. 159–170.
- [219] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications”. In: *ArXiv abs/1704.04861* (2017).
- [220] <https://www.alz.org>. *Younger/Early Onset of Alzheimer's Disease*. 2019.
- [221] Y.-L. Huang, W. O. Song, R. A. Schemmel, and S. M. Hoerr. “What do college students eat? Food selection and meal pattern”. In: *Nutrition Research* 14.8 (1994), pp. 1143–1153.

- [222] E. Hüllermeier and W. Waegeman. “Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods”. In: *Machine Learning* 110 (2021), pp. 457–506.
- [223] G. C.-L. Hung, P.-C. Yang, C.-C. Chang, J.-H. Chiang, and Y.-Y. Chen. “Predicting negative emotions based on mobile phone usage patterns: an exploratory study”. In: *JMIR research protocols* 5.3 (2016), e5551.
- [224] T. Huynh, M. Fritz, and B. Schiele. “Discovery of activity patterns using topic models”. en. In: *Proceedings of the 10th international conference on Ubiquitous computing - UbiComp '08*. Seoul, Korea: ACM Press, 2008, p. 10.
- [225] L. H. Hwang JH. “A Study on Lifestyles, Dietary Habits, Nutrition Knowledge and Dietary behaviors of Male University Students According to Residence Type.” In: *Korean Journal of Community Nutrition* (2007), pp. 381–395.
- [226] J. Ifland, H. Preuss, M. Marcus, K. Rourke, W. Taylor, K. Burau, W. Jacobs, W. Kadish, and G. Manso. “Refined food addiction: A classic substance use disorder”. In: *Medical Hypotheses* 72.5 (2009), pp. 518–526.
- [227] K. M. Jackson, J. E. Merrill, N. P. Barnett, S. M. Colby, C. C. Abar, M. L. Rogers, and K. L. Hayes. “Contextual influences on early drinking: Characteristics of drinking and nondrinking days.” In: *Psychology of Addictive Behaviors* 30.5 (2016), p. 566.
- [228] N. Jaques, S. Taylor, A. Azaria, A. Ghandeharioun, A. Sano, and R. Picard. “Predicting students’ happiness from physiology, phone, mobility, and behavioral data”. In: *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2015, pp. 222–228.
- [229] M. M. Jastran, C. A. Bisogni, J. Sobal, C. Blake, and C. M. Devine. “Eating routines. Embedded, value based, modifiable, and reflective”. In: *Appetite* 52.1 (2009), pp. 127–136.
- [230] S. Jiang, W. Min, L. Liu, and Z. Luo. “Multi-Scale Multi-View Deep Feature Aggregation for Food Recognition”. In: *IEEE Transactions on Image Processing* 29 (2020), pp. 265–276.
- [231] W. Jiang, Q. Li, L. Su, C. Miao, Q. Gu, and W. Xu. “Towards personalized learning in mobile sensing systems”. In: *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*. IEEE. 2018, pp. 321–333.
- [232] E. C. Johnson-Sabine, K. H. Wood, and A. Wakeling. “Mood changes in bulimia nervosa”. In: *The British Journal of Psychiatry* 145.5 (1984), pp. 512–516.
- [233] C. Jung. *Psychological types*. Routledge, 2016.
- [234] J. Jung, L. Wellard-Cole, C. Cai, I. Koprinska, K. Yacef, M. Allman-Farinelli, and J. Kay. “Foundations for systematic evaluation and benchmarking of a mobile food logger in a large-scale nutrition study”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.2 (2020), pp. 1–25.
- [235] A. Kabir, S. Miah, and A. Islam. “Factors influencing eating behavior and dietary intake among resident students in a public university in Bangladesh: A qualitative study”. In: *PloS one* 13.6 (2018), e0198801.
- [236] A. Kadamura, C.-Y. Li, K. Tsukada, H.-H. Chu, and I. Siio. “Persuasive technology to improve eating behavior using a sensor-embedded fork”. In: *Proceedings of the 2014 acm international joint conference on pervasive and ubiquitous computing*. 2014, pp. 319–329.
- [237] N. Kammoun, L. Meegahapola, and D. Gatica-Perez. “Understanding the Social Context of Eating with Multimodal Smartphone Sensing: The Role of Country Diversity”. In: (2023).
- [238] E. Kanjo, D. J. Kuss, and C. S. Ang. “NotiMind: utilizing responses to smart phone notifications as affective sensors”. In: *IEEE Access* 5 (2017), pp. 22023–22035.

Bibliography

- [239] M. Kanning and W. Schlicht. “Be active and become happy: an ecological momentary assessment of physical activity and mood”. In: *Journal of Sport and Exercise Psychology* 32.2 (2010), pp. 253–261.
- [240] H.-T. Kao, S. Yan, H. Hosseinmardi, S. Narayanan, K. Lerman, and E. Ferrara. “User-Based Collaborative Filtering Mobile Health System”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.4 (2020), pp. 1–17.
- [241] H.-L. Kao, B.-J. Ho, A. C. Lin, and H.-H. Chu. “Phone-based gait analysis to detect alcohol usage”. In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. 2012, pp. 661–662.
- [242] D. Karantonis, M. Narayanan, M. Mathie, N. Lovell, and B. Celler. “Implementation of a Real-Time Human Movement Classifier Using a Triaxial Accelerometer for Ambulatory Monitoring”. In: *Information Technology in Biomedicine, IEEE Transactions on* 10 (Feb. 2006), pp. 156–167.
- [243] K. Karkkainen and J. Joo. “Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 1548–1558.
- [244] G. S. Keenan, L. Childs, P. J. Rogers, M. M. Hetherington, and J. M. Brunstrom. “The portion size effect: Women demonstrate an awareness of eating more than intended when served larger than normal portions”. In: *Appetite* 126 (2018), pp. 54–60.
- [245] R. Kenny, B. Dooley, and A. Fitzgerald. “Feasibility of “CopeSmart”: A Telemental Health App for Adolescents”. In: *JMIR MENTAL HEALTH* 2.3 (2015).
- [246] J. Kerr, L. Frank, J. F. Sallis, B. Saelens, K. Glanz, and J. Chapman. “Predictors of trips to food destinations”. In: *International Journal of Behavioral Nutrition and Physical Activity* 9.1 (2012), pp. 1–10.
- [247] M. Khwaja, S. S. Vaid, S. Zannone, G. M. Harari, A. A. Faisal, and A. Matic. “Modeling personality vs. modeling personalidad: In-the-wild mobile data analysis in five countries suggests cultural impact on personality models”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3.3 (2019), pp. 1–24.
- [248] H.-Y. Kim. “Analysis of variance (ANOVA) comparing means of more than two groups”. In: *Restorative dentistry & endodontics* 39.1 (2014), pp. 74–77.
- [249] T. Kim. “T test as a parametric statistic”. In: *Korean Journal of Anesthesiology* 68.6 (Nov. 2015), p. 540.
- [250] D. P. Kingma and J. Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).
- [251] S. Klenk, D. Reifegerste, and R. Renatus. “Gender differences in gratifications from fitness app use and implications for health interventions”. In: *Mobile Media & Communication* 5.2 (2017), pp. 178–193.
- [252] T. Kondo, H. Kamachi, S. Ishii, A. Yokokubo, and G. Lopez. “Robust classification of eating sound collected in natural meal environment”. In: *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 2019, pp. 105–108.
- [253] E. P. Köster and J. Mojet. “From mood to food and from food to mood: A psychological perspective on the measurement of food-related emotions in consumer research”. In: *Food research international* 76 (2015), pp. 180–191.

- [254] D Kromhout, A Keys, C Aravanis, R Buzina, F Fidanza, S Giampaoli, A Jansen, A Menotti, S Nedeljkovic, and M Pekkarinen. "Food consumption patterns in the 1960s in seven countries". In: *The American Journal of Clinical Nutrition* 49.5 (May 1989), pp. 889–894.
- [255] S. Kumar, W. J. Nilsen, A. Abernethy, A. Atienza, K. Patrick, M. Pavel, W. T. Riley, A. Shar, B. Spring, D. Spruijt-Metz, et al. "Mobile health technology evaluation: the mHealth evidence workshop". In: *American journal of preventive medicine* 45.2 (2013), pp. 228–236.
- [256] E. Kuntsche and G. Gmel. "Alcohol consumption in late adolescence and early adulthood—where is the problem?" In: *Swiss medical weekly* 143.2930 (2013), w13826–w13826.
- [257] E. Kuntsche, R. Knibbe, G. Gmel, and R. Engels. "Why do young people drink? A review of drinking motives". In: *Clinical psychology review* 25.7 (2005), pp. 841–861.
- [258] E. Kuntsche and S. Kuntsche. "Development and validation of the drinking motive questionnaire revised short form (DMQ–R SF)". In: *Journal of Clinical Child & Adolescent Psychology* 38.6 (2009), pp. 899–908.
- [259] E. Kuntsche, S. Kuntsche, J. Thrul, and G. Gmel. "Binge drinking: Health impact, prevalence, correlates and interventions". In: *Psychology & health* 32.8 (2017), pp. 976–1017.
- [260] E. Kuntsche and F. Labhart. "The future is now—Using personal cellphones to gather data on substance use and related factors." In: (2014).
- [261] F. Labhart, M. Livingston, R. Engels, and E. Kuntsche. "After how many drinks does someone experience acute consequences—determining thresholds for binge drinking based on two event-level studies". In: *Addiction* 113.12 (2018), pp. 2235–2244.
- [262] F. Labhart, S. Muralidhar, B. Massé, L. Meegahapola, E. Kuntsche, and D. Gatica-Perez. "Ten seconds of my nights: Exploring methods to measure brightness, loudness and attendance and their associations with alcohol use from video clips". In: *PLoS one* 16.4 (2021), e0250443.
- [263] F. Labhart, F. Tarsetti, O. Bornet, D. Santani, J. Truong, S. Landolt, D. Gatica-Perez, and E. Kuntsche. "Capturing drinking and nightlife behaviours and their social and physical context with a smartphone application—investigation of users' experience and reactivity". In: *Addiction Research & Theory* 28.1 (2020), pp. 62–75.
- [264] F. Labhart, S. Wells, K. Graham, and E. Kuntsche. "Do individual and situational factors explain the link between predrinking and heavier alcohol consumption? An event-level study of types of beverage consumed and social context". In: *Alcohol and Alcoholism* 49.3 (2014), pp. 327–335.
- [265] Y. T. Lagerros and P. Lagiou. "Assessment of physical activity and energy expenditure in epidemiological research of chronic diseases". In: *European journal of epidemiology* 22.6 (2007), pp. 353–362.
- [266] D. Lakens. "Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front Psychol* 4: 863". In: *Frontiers in psychology* 4 (Nov. 2013), p. 863.
- [267] D. Lakens. "Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front Psychol* 4: 863". In: *Frontiers in psychology* 4 (Nov. 2013), p. 863.
- [268] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell. "A survey of mobile phone sensing". In: *IEEE Communications Magazine* 48.9 (2010), pp. 140–150.
- [269] N. D. Lane, M. Lin, M. Mohammod, X. Yang, H. Lu, G. Cardone, S. Ali, A. Doryab, E. Berke, A. T. Campbell, et al. "Bewell: Sensing sleep, physical activities and social interactions to promote wellbeing". In: *Mobile Networks and Applications* 19 (2014), pp. 345–359.

Bibliography

- [270] N. D. Lane, Y. Xu, H. Lu, S. Hu, T. Choudhury, A. T. Campbell, and F. Zhao. "Enabling Large-scale Human Activity Inference on Smartphones Using Community Similarity Networks (Csn)". In: *Proceedings of the 13th International Conference on Ubiquitous Computing*. UbiComp '11. Beijing, China: ACM, 2011, pp. 355–364.
- [271] G. Laput and C. Harrison. "Sensing fine-grained hand activity with smartwatches". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–13.
- [272] A.-M. Laslett, R. Room, J. Ferris, C. Wilkinson, M. Livingston, and J. Mugavin. "Surveying the range and magnitude of alcohol's harm to others in Australia". In: *Addiction* 106.9 (2011), pp. 1603–1611.
- [273] N. Lathia, G. M. Sandstrom, C. Mascolo, and P. J. Rentfrow. "Happier People Live More Active Lives: Using Smartphones to Link Happiness and Physical Activity". In: *PLOS ONE* 12.1 (Jan. 2017), pp. 1–13.
- [274] J. K. Laurila, D. Gatica-Perez, I. Aad, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, M. Miettinen, et al. *The mobile data challenge: Big data for mobile computing research*. Tech. rep. 2012.
- [275] K. Leadley, C. L. Clark, and R. Caetano. "Couples' drinking patterns, intimate partner violence, and alcohol-related partnership problems". In: *Journal of substance abuse* 11.3 (2000), pp. 253–263.
- [276] D. K. Lee. "Alternatives to P value: confidence interval and effect size". In: *Korean journal of anesthesiology* 69.6 (2016), pp. 555–562.
- [277] I.-M. Lee and E. J. Shiroma. "Using accelerometers to measure physical activity in large-scale epidemiological studies: issues and challenges". In: *British Journal of Sports Medicine* 48.3 (2014), pp. 197–201.
- [278] R. A. Lee and M. E. Jung. "Evaluation of an mHealth App (DeStressify) on University Students' Mental Health: Pilot Trial". In: *JMIR Ment Health* 5.1 (2018), e2.
- [279] K. E. Leonard and P. Mudar. "Peer and partner drinking and the transition to marriage: a longitudinal examination of selection and influence processes." In: *Psychology of addictive behaviors* 17.2 (2003), p. 115.
- [280] B. Li and A. Sano. "Extraction and Interpretation of Deep Autoencoder-based Temporal Features from Wearables for Forecasting Personalized Mood, Health, and Stress". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.2 (2020), pp. 1–26.
- [281] M. Li. "An improved fcm clustering algorithm based on cosine similarity". In: *Proceedings of the 2019 International Conference on Data Mining and Machine Learning*. 2019, pp. 103–109.
- [282] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma. "Mining user similarity based on location history". In: *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*. 2008, pp. 1–10.
- [283] T. Li, M. Zhang, Y. Li, E. Lagerspetz, S. Tarkoma, and P. Hui. "The Impact of Covid-19 on Smartphone Usage". In: *IEEE Internet of Things Journal* 8.23 (Dec. 2021), pp. 16723–16733.
- [284] Y. Liang and J. Li. "Computer vision-based food calorie estimation: dataset, method, and experiment". In: *ArXiv abs/1705.07632* (2017).
- [285] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong. "MoodScope: Building a Mood Sensor from Smartphone Usage Patterns". In: *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*. MobiSys '13. Taipei, Taiwan: ACM, 2013, pp. 389–402.

- [286] S. L. Lim, P. Bentley, N. Kanakam, F. Ishikawa, and S. Honiden. “Investigating Country Differences in Mobile App User Behavior and Challenges for Software Engineering”. In: *IEEE Transactions on Software Engineering* 41.1 (Sept. 2014).
- [287] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao. “Federated learning in mobile edge networks: A comprehensive survey”. In: *IEEE Communications Surveys & Tutorials* 22.3 (2020), pp. 2031–2063.
- [288] M. Lin, N. D. Lane, M. Mohammad, X. Yang, H. Lu, G. Cardone, S. Ali, A. Doryab, E. Berke, A. T. Campbell, and T. Choudhury. “BeWell+: Multi-dimensional Wellbeing Monitoring with Community-guided User Feedback and Energy Optimization”. In: *Proceedings of the Conference on Wireless Health*. WH '12. San Diego, California: ACM, 2012, 10:1–10:8.
- [289] T. Lintonen, S. Ahlström, and L. Metso. “The reliability of self-reported drinking in adolescence”. In: *Alcohol and alcoholism* 39.4 (2004), pp. 362–368.
- [290] S. Lipperman-Kreda, L. J. Finan, and J. W. Grube. “Social and situational characteristics associated with adolescents’ drinking at party and non-party events”. In: *Addictive behaviors* 83 (2018), pp. 148–153.
- [291] C. G. Lisco, D. J. Parrott, and A. T. Tharp. “The role of heavy episodic drinking and hostile sexism in men’s sexual aggression toward female intimate partners”. In: *Addictive behaviors* 37.11 (2012), pp. 1264–1270.
- [292] B. Liu, H. Young, F. L. Crowe, V. S. Benson, E. A. Spencer, T. J. Key, P. N. Appleby, and V. Beral. “Development and evaluation of the Oxford WebQ, a low-cost, web-based method for assessment of previous 24 h dietary intakes in large-scale prospective studies”. In: *Public health nutrition* 14.11 (2011), pp. 1998–2005.
- [293] R. Liu, A. A. Ramli, H. Zhang, E. Henricson, and X. Liu. “An Overview of Human Activity Recognition Using Wearable Sensors: Healthcare and Artificial Intelligence”. In: *arXiv:2103.15990 [cs, eess]* (Aug. 2021). arXiv: 2103.15990.
- [294] T. Liu, J. Hernandez, M. Gonzalez-Franco, A. Maselli, M. Kneisel, A. Glass, J. Chudge, and A. Miller. “Characterizing and Predicting Engagement of Blind and Low-Vision People with an Audio-Based Navigation App”. In: *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI EA '22. New Orleans, LA, USA: Association for Computing Machinery, 2022.
- [295] E. V. Long, L. R. Vartanian, C. P. Herman, and J. Polivy. “What does it mean to overeat?” In: *Eating Behaviors* 37 (2020).
- [296] O. Lopez-Fernandez, D. J. Kuss, L. Romo, Y. Morvan, L. Kern, P. Graziani, A. Rousseau, H.-J. Rumpf, A. Bischof, A.-K. Gässler, A. Schimmenti, A. Passanisi, N. Männikkö, M. Kääriänen, Z. Demetrovics, O. Király, M. Chóliz, J. J. Zacarés, E. Serra, M. D. Griffiths, H. M. Pontes, B. Lelonek-Kuleta, J. Chwaszcz, D. Zullino, L. Rochat, S. Achab, and J. Billieux. “Self-reported dependence on mobile phones in young adults: A European cross-cultural empirical survey”. In: *Journal of Behavioral Addictions* 6.2 (2017), pp. 168–177.
- [297] *Lose It!*, <https://www.loseit.com/>. 2019. URL: <https://www.loseit.com/>.
- [298] A. D. Lotz. “In between the global and the local: Mapping the geographies of Netflix as a multinational service”. In: *International Journal of Cultural Studies* 24.2 (2021), pp. 195–215.
- [299] B. Löwe, K. Kroenke, W. Herzog, and K. Gräfe. “Measuring depression outcome with a brief self-report instrument: sensitivity to change of the Patient Health Questionnaire (PHQ-9)”. In: *Journal of affective disorders* 81.1 (2004), pp. 61–66.

Bibliography

- [300] H. Lu, D. Fraundorfer, M. Rabbi, M. S. Mast, G. T. Chittaranjan, A. T. Campbell, D. Gatica-Perez, and T. Choudhury. “StressSense: Detecting Stress in Unconstrained Acoustic Environments Using Smartphones”. In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. UbiComp '12. Pittsburgh, Pennsylvania: ACM, 2012, pp. 351–360.
- [301] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang. “Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2507–2516.
- [302] Y. Luo, Y.-H. Kim, B. Lee, N. Hassan, and E. K. Choe. “FoodScrap: Promoting Rich Data Capture and Reflective Food Journaling Through Speech Input”. In: *Designing Interactive Systems Conference 2021*. 2021, pp. 606–618.
- [303] H. T. Luomala, R. Kumar, V. Worm, and J. Singh. “Cross-cultural differences in mood-regulation: An empirical comparison of individualistic and collectivistic cultures”. In: *Journal of International Consumer Marketing* 16.4 (2004), pp. 39–62.
- [304] G. Ma. “Food, eating behavior, and culture in Chinese society”. In: *Journal of Ethnic Foods* 2.4 (2015), pp. 195–199.
- [305] Y. Ma, B. Xu, Y. Bai, G. Sun, and R. Zhu. “Daily Mood Assessment Based on Mobile Phone Sensing”. In: *2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks*. IEEE. 2012, pp. 142–147.
- [306] M. Macht. “How emotions affect eating: a five-way model”. In: *Appetite* 50.1 (2008), pp. 1–11.
- [307] A. Madan, M. Cebrian, D. Lazer, and A. Pentland. “Social Sensing for Epidemiological Behavior Change”. In: *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*. UbiComp '10. Copenhagen, Denmark: Association for Computing Machinery, 2010, 291–300.
- [308] N. Mairittha, T. Mairittha, and S. Inoue. “Improving activity data collection with on-device personalization using fine-tuning”. In: *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*. 2020, pp. 255–260.
- [309] S. Makonin, L. Bartram, and F. Popowich. “A Smarter Smart Home: Case Studies of Ambient Intelligence”. In: *IEEE Pervasive Computing* 12.1 (2013), pp. 58–66.
- [310] K. Malhotra and A. Khosla. “Automatic identification of gender & accent in spoken Hindi utterances with regional Indian accents”. In: *2008 IEEE Spoken Language Technology Workshop*. IEEE. 2008, pp. 309–312.
- [311] Q. Mao, M. Jay, J. L. Hoffman, J. Calvert, C. Barton, D. Shimabukuro, L. Shieh, U. Chettipally, G. Fletcher, Y. Kerem, et al. “Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU”. In: *BMJ open* 8.1 (2018), e017833.
- [312] A. Marcano-Cedeño, J. Quintanilla-Domínguez, M. Cortina-Januchs, and D. Andina. “Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network”. In: *IECON 2010-36th annual conference on IEEE industrial electronics society*. IEEE. 2010, pp. 2845–2850.
- [313] E. J. Marshall. “Adolescent alcohol use: risks and consequences”. In: *Alcohol and alcoholism* 49.2 (2014), pp. 160–164.
- [314] M. Mathie, B. Celler, N. Lovell, and A. Coster. “Classification of basic daily movements using a triaxial accelerometer”. In: *Medical & biological engineering & computing* 42 (Oct. 2004), pp. 679–87.

- [315] A. Mathur, A. Isopoussu, N. Berthouze, N. D. Lane, and F. Kawsar. “Unsupervised domain adaptation for robust sensory systems”. In: *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 2019, pp. 505–509.
- [316] A. Mathur, L. M. Kalanadhabhatta, R. Majethia, and F. Kawsar. “Moving Beyond Market Research: Demystifying Smartphone User Behavior in India”. en. In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1.3 (Sept. 2017), 82:1–82:27.
- [317] M. Matthews, G. Doherty, J. Sharry, and C. Fitzpatrick. “Mobile phone mood charting for adolescents”. In: *British Journal of Guidance & Counselling* 36.2 (2008), pp. 113–129.
- [318] K. Matton, R. Lewis, J. Guttag, and R. Picard. “Contrastive Learning of Electrodermal Activity Representations for Stress Detection”. In: *Conference on Health, Inference, and Learning*. PMLR. 2023, pp. 410–426.
- [319] N. McCarthy. *Where People Spend The Most Time Eating & Drinking*. 2020. URL: <https://www.statista.com/chart/13226/where-people-spend-the-most-time-eating-drinking/>.
- [320] D. McCarty. “Environmental factors in substance abuse: The microsetting”. In: *Determinants of substance abuse: Biological, psychological, and environmental factors*. Springer, 1985, pp. 247–281.
- [321] F. J. McClernon and R. Roy Choudhury. “I am your smartphone, and I know you are about to smoke: the application of mobile sensing and computing approaches to smoking research and treatment”. In: *Nicotine & tobacco research* 15.10 (2013), pp. 1651–1654.
- [322] M. A. McCrory, P. J. Fuss, N. P. Hays, A. G. Vinken, A. S. Greenberg, and S. B. Roberts. “Overeating in America: Association between Restaurant Food Consumption and Body Fatness in Healthy Adult Men and Women Ages 19 to 80”. In: *Obesity Research* 7.6 (1999), pp. 564–571.
- [323] L. Meegahapola, W. Bangamuarachchi, A. Chamantha, S. Ruiz-Correa, I. Perera, and D. Gatica-Perez. “Sensing Eating Events in Context: A Smartphone-Only Approach”. In: *IEEE Access* 10 (2022), pp. 61249–61264.
- [324] L. Meegahapola, W. Droz, P. Kun, A. de Götzen, C. Nutakki, S. Diwakar, S. R. Correa, D. Song, H. Xu, M. Bidoglia, et al. “Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6.4 (2023), pp. 1–32.
- [325] L. Meegahapola and D. Gatica-Perez. “Smartphone sensing for the well-being of young adults: A review”. In: *IEEE Access* 9 (2020), pp. 3374–3399.
- [326] L. Meegahapola, T. Kandappu, K. Jayarajah, L. Akoglu, S. Xiang, and A. Misra. “BuScope: Fusing Individual & Aggregated Mobility Behavior for Live Smart City Services”. In: *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. MobiSys ’19. Seoul, Republic of Korea: ACM, 2019, pp. 41–53.
- [327] L. Meegahapola, F. Labhart, T.-T. Phan, and D. Gatica-Perez. “Examining the social context of alcohol drinking in young adults with smartphone sensing”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5.3 (2021), pp. 1–26.
- [328] L. Meegahapola, S. Ruiz-Correa, and D. Gatica-Perez. “Alone or with others? understanding eating episodes of college students with mobile sensing”. In: *Proceedings of the 19th International Conference on Mobile and Ubiquitous Multimedia*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 162–166.

Bibliography

- [329] L. Meegahapola, S. Ruiz-Correa, and D. Gatica-Perez. “Protecting mobile food diaries from getting too personal”. en. In: *Proceedings of the 19th International Conference on Mobile and Ubiquitous Multimedia*. MUM 2020. Essen, Germany: Association for Computing Machinery, Nov. 2020, pp. 212–222.
- [330] L. Meegahapola, S. Ruiz-Correa, V. d. C. Robledo-Valero, E. E. Hernandez-Huerfano, L. Alvarez-Rivera, R. Chenu-Abente, and D. Gatica-Perez. “One more bite? Inferring food consumption level of college students using smartphone sensing and self-reports”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5.1 (2021), pp. 1–28.
- [331] A. Mehrotra and M. Musolesi. *Intelligent Notification Systems: A Survey of the State of the Art and Research Challenges*. 2017.
- [332] K. Mercer, L. Giangregorio, E. Schneider, P. Chilana, M. Li, and K. Grindrod. “Acceptance of Commercially Available Wearable Activity Trackers Among Adults Aged Over 50 and With Chronic Illness: A Mixed-Methods Evaluation”. en. In: *JMIR mHealth and uHealth* 4.1 (Jan. 2016), e7.
- [333] 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.
- [334] M. A. Merrill and T. Althoff. “Self-supervised Pretraining and Transfer Learning Enable Flu and COVID-19 Predictions in Small Mobile Sensing Datasets”. In: *arXiv preprint arXiv:2205.13607* (2022).
- [335] J. Meyer, J. Kay, D. A. Epstein, P. Eslambolchilar, and L. M. Tang. “A life of data: characteristics and challenges of very long term self-tracking for health and wellness”. In: *ACM Transactions on Computing for Healthcare* 1.2 (2020), pp. 1–4.
- [336] D. Miller, L. A. Rabho, P. Awondo, M. de Vries, M. Duque, P. Garvey, L. Haapio-Kirk, C. Hawkins, A. Otaegui, S. Walton, and X. Wang. *The Global Smartphone: Beyond a youth technology*. UCL Press, 2021.
- [337] W. Min, S. Jiang, L. Liu, Y. Rui, and R. Jain. “A Survey on Food Computing”. In: *ACM Comput. Surv.* 52.5 (Sept. 2019).
- [338] W. Min, L. Liu, Z. Luo, and S. Jiang. “Ingredient-Guided Cascaded Multi-Attention Network for Food Recognition”. In: *Proceedings of the 27th ACM International Conference on Multimedia*. MM '19. Nice, France: Association for Computing Machinery, 2019, 1331–1339.
- [339] M. Mirtchouk, D. McGuire, A. Deierlein, and S. Kleinberg. “Automated Estimation of Food Type from Body-worn Audio and Motion Sensors in Free-Living Environments”. In: *Proceedings of machine learning research* 106 (Aug. 2019), pp. 641–662.
- [340] V. Mishra, S. Sen, G. Chen, T. Hao, J. Rogers, C.-H. Chen, and D. Kotz. “Evaluating the reproducibility of physiological stress detection models”. In: *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 4.4 (2020), pp. 1–29.
- [341] A. Mislove. “Pulse of the nation: US mood throughout the day inferred from twitter”. In: <http://www.ccs.neu.edu/home/amislove/twittermood/> (2010).
- [342] Y. Mitani and Y. Hamamoto. “A local mean-based nonparametric classifier”. In: *Pattern Recognition Letters* 27.10 (2006), pp. 1151–1159.
- [343] *Mobile Fact Sheet*. 2021. URL: <https://www.pewinternet.org/fact-sheet/mobile/>.
- [344] *Mobile Fact Sheet*, <https://www.pewinternet.org/fact-sheet/mobile/>. 2019. URL: <https://www.pewinternet.org/fact-sheet/mobile/>.

- [345] Mobius. *11 surprising mobile health statistics*. 2019. URL: <https://www.mobius.md/blog/2019/03/11-mobile-health-statistics/>.
- [346] S. M. Mohammad. "Ethics sheet for automatic emotion recognition and sentiment analysis". In: *arXiv preprint arXiv:2109.08256* (2021).
- [347] C. D. Mohr, S. Aversa, D. A. Kenny, and F. K. Del Boca. "' Getting by (or getting high) with a little help from my friends": an examination of adult alcoholics' friendships." In: *Journal of Studies on Alcohol* 62.5 (2001), pp. 637–645.
- [348] D. C. Mohr, M. Zhang, and S. M. Schueller. "Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning". In: *Annual Review of Clinical Psychology* 13.1 (2017). PMID: 28375728, pp. 23–47.
- [349] M. A. S. Mondol, B. Bell, M. Ma, R. Alam, I. Emi, S. M. Preum, K. de la Haye, D. Spruijt-Metz, J. C. Lach, and J. A. Stankovic. "MFED: A System for Monitoring Family Eating Dynamics". In: *arXiv preprint arXiv:2007.05831* (2020).
- [350] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. Pentland. "Unique in the shopping mall: On the reidentifiability of credit card metadata". In: *Science* 347.6221 (2015), pp. 536–539.
- [351] M. Moor, N. Bennett, D. Plečko, M. Horn, B. Rieck, N. Meinshausen, P. Bühlmann, and K. Borgwardt. "Predicting sepsis using deep learning across international sites: a retrospective development and validation study". In: *EClinicalMedicine* 62 (2023).
- [352] J. Morales and D. Akopian. "Physical activity recognition by smartphones, a survey". In: *Biocybernetics and Biomedical Engineering* 37.3 (2017), pp. 388–400.
- [353] M. B. Morshed, J. Hernandez, D. McDuff, J. Suh, E. Howe, K. Rowan, M. Abdin, G. Ramos, T. Tran, and M. Czerwinski. "Advancing the Understanding and Measurement of Workplace Stress in Remote Information Workers from Passive Sensors and Behavioral Data". In: *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2022, pp. 1–8.
- [354] M. B. Morshed, S. S. Kulkarni, R. Li, K. Saha, L. G. Roper, L. Nachman, H. Lu, L. Mirabella, S. Srivastava, M. De Choudhury, et al. "A real-time eating detection system for capturing eating moments and triggering ecological momentary assessments to obtain further context: System development and validation study". In: *JMIR mHealth and uHealth* 8.12 (2020), e20625.
- [355] M. B. Morshed, S. S. Kulkarni, K. Saha, R. Li, L. G. Roper, L. Nachman, H. Lu, L. Mirabella, S. Srivastava, K. de Barbaro, et al. "Food, mood, context: Examining college students' eating context and mental well-being". In: *ACM Transactions on Computing for Healthcare* 3.4 (2022), pp. 1–26.
- [356] M. B. Morshed, K. Saha, R. Li, S. K. D'Mello, M. De Choudhury, G. D. Abowd, and T. Plötz. "Prediction of Mood Instability with Passive Sensing". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3.3 (Sept. 2019), pp. 1–21.
- [357] S. T. Moturu, I. Khayal, N. Aharony, W. Pan, and A. S. Pentland. "Using Social Sensing to Understand the Links Between Sleep, Mood, and Sociability". In: *Proceedings of IEEE International Conference on Social Computing*. 2011.
- [358] S. R. Müller, X. L. Chen, H. Peters, A. Chaintreau, and S. C. Matz. "Depression predictions from GPS-based mobility do not generalize well to large demographically heterogeneous samples". In: *Scientific Reports* 11.1 (2021), pp. 1–10.

Bibliography

- [359] E. L. Murnane, S. Abdullah, M. Matthews, M. Kay, J. A. Kientz, T. Choudhury, G. Gay, and D. Cosley. "Mobile Manifestations of Alertness: Connecting Biological Rhythms with Patterns of Smartphone App Use". In: *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. MobileHCI '16. Florence, Italy: ACM, 2016, pp. 465–477.
- [360] *MyFitnessPal*. 2019. URL: <https://www.myfitnesspal.com/>.
- [361] M. Nahum, T. M. Van Vleet, V. S. Sohal, J. J. Mirzabekov, V. R. Rao, D. L. Wallace, M. B. Lee, H. Dawes, A. Stark-Inbar, J. T. Jordan, B. Biagianti, M. Merzenich, and E. F. Chang. "Immediate Mood Scaler: Tracking Symptoms of Depression and Anxiety Using a Novel Mobile Mood Scale". In: *JMIR Mhealth Uhealth* 5.4 (2017), e44.
- [362] I. Nahum-Shani, S. N. Smith, B. J. Spring, L. M. Collins, K. Witkiewitz, A. Tewari, and S. A. Murphy. "Just-in-time adaptive interventions (JITAs) in mobile health: key components and design principles for ongoing health behavior support". In: *Annals of Behavioral Medicine* 52.6 (2018), pp. 446–462.
- [363] A. Nanchen, L. Meegahapola, W. Droz, and D. Gatica-Perez. "Keep Sensors in Check: Disentangling Country-Level Generalization Issues in Mobile Sensor-Based Models with Diversity Scores". In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES)* (2023).
- [364] C. Naseeb and B. A. Saeedi. "Activity recognition for locomotion and transportation dataset using deep learning". en. In: *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*. Virtual Event Mexico: ACM, Sept. 2020, pp. 329–334.
- [365] A. Natekin and A. Knoll. "Gradient boosting machines, a tutorial". In: *Frontiers in Neurorobotics* 7 (2013), p. 21.
- [366] nationsonline. *The Continents of the World*. 2022. URL: <https://www.nationsonline.org/oneworld/continents.htm>.
- [367] T. Nef, P. Urwyler, M. Büchler, I. Tarnanas, R. Stucki, D. Cazzoli, R. Müri, and U. Mosimann. "Evaluation of Three State-of-the-Art Classifiers for Recognition of Activities of Daily Living from Smart Home Ambient Data". en. In: *Sensors* 15.5 (May 2015), pp. 11725–11740.
- [368] J. J. Nelson and C. M. Pieper. "'Maladies of Infinite Aspiration': Smartphones, Meaning-Seeking, and Anomigenesis". en. In: *Sociological Perspectives* (2022), p. 073112142211142.
- [369] S. Nepal, G. J. Martinez, S. Mirjafari, K. Saha, V. D. Swain, X. Xu, P. G. Audia, M. De Choudhury, A. K. Dey, A. Striegel, et al. "A Survey of Passive Sensing in the Workplace". In: *arXiv preprint arXiv:2201.03074* (2022).
- [370] M. D. Newcomb and P. M. Bentler. "Impact of adolescent drug use and social support on problems of young adults: A longitudinal study." In: *Journal of Abnormal Psychology* (1988).
- [371] J. Nhan, K. Bowen, and A. Bartula. "A comparison of a public and private university of the effects of low-cost streaming services and income on movie piracy". In: *Technology in Society* 60 (2020), p. 101213.
- [372] C. E. Nihan. "Healthier? More efficient? Fairer? An overview of the main ethical issues raised by the use of ubicomp in the workplace". In: *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal* 2.1 (2013), p. 29.
- [373] W. S. Noble. "What is a support vector machine?" In: *Nature biotechnology* 24.12 (2006), pp. 1565–1567.

- [374] J. Northcote and M. Livingston. "Accuracy of self-reported drinking: observational verification of 'last occasion' drink estimates of young adults". In: *Alcohol and Alcoholism* 46.6 (2011), pp. 709–713.
- [375] M. Obuchi, J. F. Huckins, W. Wang, A. daSilva, C. Rogers, E. Murphy, E. Hedlund, P. Holtzheimer, S. Mirjafari, and A. Campbell. "Predicting brain functional connectivity using mobile sensing". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.1 (2020), pp. 1–22.
- [376] H. Oh, L. Jalali, and R. Jain. "An intelligent notification system using context from real-time personal activity monitoring". In: *2015 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2015, pp. 1–6.
- [377] K. Okamoto and K. Yanai. "An automatic calorie estimation system of food images on a smartphone". In: *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*. 2016, pp. 63–70.
- [378] J. A. Olson, D. A. Sandra, Élissa S. Colucci, A. Al Bikaii, D. Chmoulevitch, J. Nahas, A. Raz, and S. P. Veissière. "Smartphone addiction is increasing across the world: A meta-analysis of 24 countries". In: *Computers in Human Behavior* 129 (2022), p. 107138.
- [379] A. Olteanu, C. Castillo, F. Diaz, and E. KÄ±cÄ±man. "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries". In: *Frontiers in Big Data* 2 (2019), p. 13.
- [380] D. W. Osgood, D. T. Ragan, L. Wallace, S. D. Gest, M. E. Feinberg, and J. Moody. "Peers and the emergence of alcohol use: Influence and selection processes in adolescent friendship networks". In: *Journal of research on adolescence* 23.3 (2013), pp. 500–512.
- [381] R. O'Donnell, B. Richardson, M. Fuller-Tyszkiewicz, and P. K. Staiger. "Delivering personalized protective behavioral drinking strategies via a smartphone intervention: a pilot study". In: *International journal of behavioral medicine* 26 (2019), pp. 401–414.
- [382] R. O'Driscoll, J. Turicchi, K. Beaulieu, S. Scott, J. Matu, K. Deighton, G. Finlayson, and J. Stubbs. "How well do activity monitors estimate energy expenditure? A systematic review and meta-analysis of the validity of current technologies". In: *British Journal of Sports Medicine* 54.6 (2020), pp. 332–340.
- [383] N. Palmius, K. E. Saunders, O. Carr, J. R. Geddes, G. M. Goodwin, and M. De Vos. "Group-personalized regression models for predicting mental health scores from objective mobile phone data streams: observational study". In: *Journal of medical Internet research* 20.10 (2018), e10194.
- [384] S. J. Pan and Q. Yang. "A Survey on Transfer Learning". In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (2010), pp. 1345–1359.
- [385] H. Pape and E. K. Bye. "Drinking with parents: Different measures, different associations with underage heavy drinking?" In: *Nordic studies on alcohol and drugs* 34.6 (2017), pp. 445–455.
- [386] G. S. Parcel, L. D. Muraskin, and C. M. Endert. "Community education: Study group report". In: *Journal of Adolescent Health Care* 9.6, Supplement (1988), S41 –S45.
- [387] S. Park, C. Gopalsamy, R. Rajamanickam, and S. Jayaraman. "The Wearable Motherboard: a flexible information infrastructure or sensate liner for medical applications". eng. In: *Studies in Health Technology and Informatics* 62 (1999), pp. 252–258.
- [388] *Partnership-on-AI, The Ethics of AI and Emotional Intelligence*. 2022. URL: <https://partnershiponai.org/paper/the-ethics-of-ai-and-emotional-intelligence/>.

Bibliography

- [389] K. Patel and D. Schlundt. "Impact of moods and social context on eating behavior". In: *Appetite* 36.2 (2001), pp. 111–118.
- [390] V. Patel, A. J. Flisher, S. Hetrick, and P. McGorry. "Mental health of young people: a global public-health challenge". In: *The Lancet* 369.9569 (2007), pp. 1302–1313.
- [391] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [392] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. "Scikit-Learn: Machine learning in Python". In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.
- [393] V. Pejovic, N. Lathia, C. Mascolo, and M. Musolesi. "Mobile-Based Experience Sampling for Behaviour Research". In: *Emotions and Personality in Personalized Services* (2016), 141–161.
- [394] Pekka and A. Kouvo. "LINKED OR DIVIDED BY THE WEB?: Internet use and sociability in four European countries". In: *Information, Communication & Society* 10.2 (2007), pp. 219–241.
- [395] J. E. Pelletier, D. J. Graham, and M. N. Laska. "Social Norms and Dietary Behaviors among Young Adults". In: *American Journal of Health Behavior* 38.1 (2014), pp. 144–152.
- [396] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang. "Moment matching for multi-source domain adaptation". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1406–1415.
- [397] I. Pentina, L. Zhang, H. Bata, and Y. Chen. "Exploring privacy paradox in information-sensitive mobile app adoption: A cross-cultural comparison". In: *Computers in Human Behavior* 65 (2016), pp. 409–419.
- [398] A. Pentland. "Looking at people: sensing for ubiquitous and wearable computing". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.1 (2000), pp. 107–119.
- [399] C. S. Perone, P. Ballester, R. C. Barros, and J. Cohen-Adad. "Unsupervised domain adaptation for medical imaging segmentation with self-ensembling". In: *NeuroImage* 194 (2019), pp. 1–11.
- [400] P. L. Peterson, J. D. Hawkins, R. D. Abbott, and R. F. Catalano. "Disentangling the effects of parental drinking, family management, and parental alcohol norms on current drinking by black and white adolescents". In: *Alcohol problems among adolescents*. Psychology Press, 2013, pp. 33–57.
- [401] PewResearchCentre. *Adapt your app by understanding what users are doing*. 2017. URL: <https://www.pewresearch.org/fact-tank/2017/09/13/about-6-in-10-young-adults-in-us-primarily-use-online-streaming-to-watch-tv/>.
- [402] L. V. Phan, N. Modersitzki, K. K. Gloystein, and S. Müller. *Mobile Sensing Around the Globe: Considerations for Cross-Cultural Research*. 2022.
- [403] T.-T. Phan and D. Gatica-Perez. "Healthy fondue dinner: Analysis and Inference of Food and Drink Consumption Patterns on Instagram". In: *Proceedings of the 16th International Conference on Mobile and Ubiquitous Multimedia*. MUM '17. Stuttgart, Germany: Association for Computing Machinery, 2017, 327–338.
- [404] T.-T. Phan, F. Labhart, S. Muralidhar, and D. Gatica-Perez. "Understanding heavy drinking at night through smartphone sensing and active human engagement". In: *Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 2020, pp. 211–222.

- [405] T.-T. Phan, S. Muralidhar, and D. Gatica-Perez. “Drinks & crowds: Characterizing alcohol consumption through crowdsensing and social media”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3.2 (2019), pp. 1–30.
- [406] R. W. Picard. “Measuring affect in the wild”. In: *Affective Computing and Intelligent Interaction: 4th International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part I 4*. Springer. 2011, pp. 3–3.
- [407] E. A. Poelen, R. H. Scholte, G. Willemsen, D. I. Boomsma, and R. C. Engels. “Drinking by parents, siblings, and friends as predictors of regular alcohol use in adolescents and young adults: a longitudinal twin-family study”. In: *Alcohol & Alcoholism* 42.4 (2007), pp. 362–369.
- [408] J. Polivy and C. P. Herman. “Overeating in Restrained and Unrestrained Eaters”. In: *Frontiers in Nutrition* 7 (2020), p. 30.
- [409] J. Polivy, C. P. Herman, and R. Deo. “Getting a bigger slice of the pie. Effects on eating and emotion in restrained and unrestrained eaters”. In: *Appetite* 55.3 (2010), pp. 426–430.
- [410] V. Prabhakaran, R. Qadri, and B. Hutchinson. “Cultural Incongruencies in Artificial Intelligence”. In: *arXiv preprint arXiv:2211.13069* (2022).
- [411] A. Pratap, D. C. Atkins, B. N. Renn, M. J. Tanana, S. D. Mooney, J. A. Anguera, and P. A. Areán. “The accuracy of passive phone sensors in predicting daily mood”. In: *Depression and anxiety* 36.1 (2019), pp. 72–81.
- [412] J. Pucher and R. Buehler. “At the frontiers of cycling. Policy innovations in the Netherlands, Denmark, and Germany.” In: (2007).
- [413] *Put down that mobile phone - it's making you fat! Scientists warn using a smartphone at meal-times can lead to an expanding waistline*. 2019. URL: <https://www.dailymail.co.uk/news/article-6738357/Scientists-warn-using-smartphone-mealtimes-lead-expanding-waistline.html>.
- [414] R. L. Pyle, J. E. Mitchell, and E. D. Eckert. “Bulimia: a report of 34 cases.” In: *The Journal of Clinical Psychiatry* (1981).
- [415] *Qualcomm's new chip brings A.I. smarts (and 5G!) to 2019 flagship Android phones*. 2019. URL: <https://www.digitaltrends.com/mobile/qualcomm-snapdragon-855-news/>.
- [416] V. M. Quick and C. Byrd-Bredbenner. “Disturbed eating behaviours and associated psychographic characteristics of college students”. In: *Journal of Human Nutrition and Dietetics* 26.s1 (2013), pp. 53–63.
- [417] M. Rabbi, M. H. Aung, M. Zhang, and T. Choudhury. “MyBehavior: Automatic Personalized Health Feedback from User Behaviors and Preferences Using Smartphones”. In: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp '15. Osaka, Japan: ACM, 2015, pp. 707–718.
- [418] K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C. Longworth, and A. Aucinas. “Emotion-Sense: A Mobile Phones Based Adaptive Platform for Experimental Social Psychology Research”. In: *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*. UbiComp '10. Copenhagen, Denmark: ACM, 2010, pp. 281–290.
- [419] M. Raento, A. Oulasvirta, and N. Eagle. “Smartphones: An Emerging Tool for Social Scientists”. In: *Sociological Methods & Research* 37.3 (2009), pp. 426–454.
- [420] T. Rahman, M. Czerwinski, R. Gilad-Bachrach, and P. Johns. “Predicting “About-to-Eat” Moments for Just-in-Time Eating Intervention”. In: *Proceedings of the 6th International Conference on Digital Health Conference*. DH '16. Montréal, Québec, Canada: Association for Computing Machinery, 2016, 141–150.

Bibliography

- [421] A. Rai, Z. Yan, D. Chakraborty, T. K. Wijaya, and K. Aberer. "Mining complex activities in the wild via a single smartphone accelerometer". In: *Proceedings of the sixth international workshop on knowledge discovery from sensor data*. 2012, pp. 43–51.
- [422] I. D. Raji and J. Buolamwini. "Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products". en. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '19. Honolulu, HI, USA: ACM, Jan. 2019, pp. 429–435.
- [423] N. Ramirez-Esparza, M. R. Mehl, J. Alvarez-Bermudez, and J. W. Pennebaker. "Are Mexicans more or less sociable than Americans? Insights from a naturalistic observation study". In: *Journal of Research in Personality* 43.1 (2009), pp. 1–7.
- [424] A. Rapp and F. Cena. "Self-monitoring and technology: challenges and open issues in personal informatics". In: *International Conference on Universal Access in Human-Computer Interaction*. Springer. 2014, pp. 613–622.
- [425] H. Rashid, S. Mendu, K. E. Daniel, M. L. Beltzer, B. A. Teachman, M. Boukhechba, and L. E. Barnes. "Predicting Subjective Measures of Social Anxiety from Sparsely Collected Mobile Sensor Data". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.3 (2020), pp. 1–24.
- [426] K. Raspopov, K. Matheson, A. Abizaid, and H. Anisman. "Unsupportive social interactions influence emotional eating behaviors. The role of coping styles as mediators". In: *Appetite* 62 (2013). Marketing to Children - Implications for Eating Behaviour and Obesity: A special issue with the UK Association for the Study of Obesity (ASO), pp. 143–149.
- [427] Z. A. Ratan, S. B. Zaman, S. M. S. Islam, and H. Hosseinzadeh. "Smartphone overuse: A hidden crisis in COVID-19". en. In: *Health Policy and Technology* 10.1 (Mar. 2021), pp. 21–22.
- [428] A. S. Rathod, A. Ingole, A. Gaidhane, and S. G. Choudhari. "Psychological Morbidities Associated With Excessive Usage of Smartphones Among Adolescents and Young Adults: A Review". en. In: *Cureus* (2022).
- [429] M. E. Rice and G. T. Harris. "Comparing Effect Sizes in Follow-Up Studies: ROC Area, Cohen's d, and r". In: *Law and Human Behavior* 29.5 (2005), pp. 615–620.
- [430] A. Richards, K. K. Kattelman, and C. Ren. "Motivating 18- to 24-Year-Olds to Increase Their Fruit and Vegetable Consumption". In: *Journal of the American Dietetic Association* 106.9 (2006), pp. 1405–1411.
- [431] D. J. Rickwood, F. P. Deane, and C. J. Wilson. "When and how do young people seek professional help for mental health problems?" In: *Medical journal of Australia* 187.S7 (2007), S35–S39.
- [432] I. Rish et al. "An empirical study of the naive Bayes classifier". In: *IJCAI 2001 workshop on empirical methods in artificial intelligence*. Vol. 3. 22. 2001, pp. 41–46.
- [433] D. Rizzuto, N. Orsini, C. Qiu, H.-X. Wang, and L. Fratiglioni. "Lifestyle, social factors, and survival after age 75: population based study". In: *BMJ* 345 (2012).
- [434] J. A. Roberts and M. E. David. "My life has become a major distraction from my cell phone: Partner phubbing and relationship satisfaction among romantic partners". In: *Computers in human behavior* 54 (2016), pp. 134–141.
- [435] E. Robinson, P. Aveyard, A. Daley, K. Jolly, A. Lewis, D. Lycett, and S. Higgs. "Eating attentively: a systematic review and meta-analysis of the effect of food intake memory and awareness on eating". In: *The American Journal of Clinical Nutrition* 97.4 (Feb. 2013), pp. 728–742.

- [436] E. Robinson, E. Harris, J. Thomas, P. Aveyard, and S. Higgs. "Reducing high calorie snack food in young adults: A role for social norms and health based messages". In: *The international journal of behavioral nutrition and physical activity* 10 (June 2013), p. 73.
- [437] Y. Rogers, N. Yuill, and P. Marshall. "Contrasting lab-based and in-the-wild studies for evaluating multi-user technologies". In: *The SAGE handbook of Digital Technology Research*, SAGE, London (2013), pp. 359–373.
- [438] B. J. Rolls, I. C. Fedoroff, and J. F. Guthrie. "Gender differences in eating behavior and body weight regulation." In: *Health Psychology* 10.2 (1991), p. 133.
- [439] N. Ronel and G. Libman. "Eating Disorders and Recovery: Lessons from Overeaters Anonymous". In: *Clinical Social Work Journal* 31 (June 2003), pp. 155–171.
- [440] M. P. Rothney, M. Neumann, A. Béziat, and K. Y. Chen. "An artificial neural network model of energy expenditure using nonintegrated acceleration signals". In: *Journal of applied physiology* (2007).
- [441] C. L. Rowe and H. A. Liddle. "Substance abuse". In: *Journal of Marital and Family Therapy* 29.1 (2003), pp. 97–120.
- [442] H. K. Ruddock and C. A. Hardman. "Guilty pleasures: The effect of perceived overeating on food addiction attributions and snack choice". In: *Appetite* 121 (2018), pp. 9–17.
- [443] S. Rump. "What kind of thief are you? Linking perceptions, personality traits and music taste to illegal downloading-how preferences, traits and notions affect online crime". B.S. thesis. University of Twente, 2011.
- [444] A. Russell, C. Hart, C. Robinson, and S. Olsen. "Children's sociable and aggressive behaviour with peers: A comparison of the US and Australia, and contributions of temperament and parenting styles". In: *International Journal of Behavioral Development* 27.1 (2003), pp. 74–86.
- [445] J. A. Russell. "A circumplex model of affect." In: *Journal of personality and social psychology* 39.6 (1980), p. 1161.
- [446] H. Saadeh, R. Q. Al Fayez, A. Al Refaei, N. Shewaikani, H. Khawaldah, S. Abu-Shanab, and M. Al-Hussaini. "Smartphone Use Among University Students During COVID-19 Quarantine: An Ethical Trigger". In: *Frontiers in Public Health* 9 (July 2021), p. 600134.
- [447] S. Saguna, A. Zaslavsky, and D. Chakraborty. "Complex activity recognition using context-driven activity theory and activity signatures". In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 20.6 (2013), pp. 1–34.
- [448] D. Sahoo, W. Hao, S. Ke, W. Xiongwei, H. Le, P. Achananuparp, E.-P. Lim, and S. C. H. Hoi. "FoodAI: Food Image Recognition via Deep Learning for Smart Food Logging". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. Anchorage, AK, USA: ACM, 2019, pp. 2260–2268.
- [449] B. Y. Salazar Vazquez, M. A. Salazar Vazquez, G. Lopez Gutierrez, K. Acosta Rosales, P. Cabrales, F. Vadillo-Ortega, M. Intaglietta, R. Perez Tamayo, and G. W. Schmid-Schoenbein. "Control of overweight and obesity in childhood through education in meal time habits. The 'good manners for a healthy future' programme". In: *Pediatric Obesity* 11.6 (2016), pp. 484–490.
- [450] S.-J. Salvy, K. de la Haye, J. C. Bowker, and R. C. Hermans. "Influence of peers and friends on children's and adolescents' eating and activity behaviors". In: *Physiology & Behavior* 106.3 (2012). Proceedings from the 2011 meeting of the Society for the Study of Ingestive Behavior (SSIB), pp. 369–378.

Bibliography

- [451] *Samsung Health App*. 2020. URL: <https://www.samsung.com/us/support/owners/app/samsung-health>.
- [452] K. San Chun, H. Jeong, R. Adaimi, and E. Thomaz. “Eating episode detection with jawbone-mounted inertial sensing”. In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2020, pp. 4361–4364.
- [453] A. Sano and R. W. Picard. “Stress Recognition Using Wearable Sensors and Mobile Phones”. In: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE. 2013, pp. 671–676.
- [454] D. Santani, T. Do, F. Labhart, S. Landolt, E. Kuntsche, and D. Gatica-Perez. “DrinkSense: Characterizing Youth Drinking Behavior Using Smartphones”. In: *IEEE Transactions on Mobile Computing* 17.10 (2018), pp. 2279–2292.
- [455] D. Santani and D. Gatica-Perez. “Loud and trendy: Crowdsourcing impressions of social ambiance in popular indoor urban places”. In: *Proceedings of the 23rd ACM international conference on Multimedia*. 2015, pp. 211–220.
- [456] J. E. Sasaki, A. M. Hickey, J. W. Staudenmayer, D. John, J. A. Kent, and P. S. Freedson. “Performance of Activity Classification Algorithms in Free-Living Older Adults”. en. In: *Medicine & Science in Sports & Exercise* 48.5 (May 2016), pp. 941–950.
- [457] R. E. Schapire. “Explaining adaboost”. In: *Empirical inference*. Springer, 2013, pp. 37–52.
- [458] L. Schelenz, I. Bison, M. Busso, A. De Götzen, D. Gatica-Perez, F. Giunchiglia, L. Meegahapola, and S. Ruiz-Correa. “The Theory, Practice, and Ethical Challenges of Designing a Diversity-Aware Platform for Social Relations”. In: *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, AIES’21, May 19–21, 2021, Virtual Event, USA*. Association for Computing Machinery, 2021, pp. 905–915.
- [459] J. J. Schouteten, J. Verwaeren, X. Gellynck, and V. L. Almlı. “Comparing a standardized to a product-specific emoji list for evaluating food products by children”. In: *Food Quality and Preference* 72 (2019), pp. 86–97.
- [460] S. M. Schueller, M. Neary, J. Lai, and D. A. Epstein. “Understanding People’s Use of and Perspectives on Mood-Tracking Apps: Interview Study”. In: *JMIR mental health* 8.8 (2021), e29368.
- [461] S. H. Schwartz. “Are there universal aspects in the structure and contents of human values?” In: *Journal of social issues* 50.4 (1994), pp. 19–45.
- [462] S. H. Schwartz, G. Melech, A. Lehmann, S. Burgess, M. Harris, and V. Owens. “Extending the cross-cultural validity of the theory of basic human values with a different method of measurement”. In: *Journal of cross-cultural psychology* 32.5 (2001), pp. 519–542.
- [463] N. Schwarz. “Retrospective and Concurrent Self-Reports: The Rationale for Real-Time Data Capture”. In: Oct. 2012.
- [464] Scikit-Learn. *Scikit-Learn Metrics - AUROC*. 2022. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html.
- [465] Scikit-Learn. *Scikit-Learn Metrics - F1-Score*. 2022. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html.
- [466] H. Selye. “Stress without Distress”. In: *Psychopathology of Human Adaptation*. Ed. by G. Serban. Boston, MA: Springer US, 1976, pp. 137–146.
- [467] S. Sen, V. Subbaraju, A. Misra, R. Balan, and Y. Lee. “Annapurna: An automated smartwatch-based eating detection and food journaling system”. In: *Pervasive and Mobile Computing* 68 (2020), p. 101259.

- [468] S. Servia-Rodríguez, K. K. Rachuri, C. Mascolo, P. J. Rentfrow, N. Lathia, and G. M. Sandstrom. “Mobile sensing at the service of mental well-being: a large-scale longitudinal study”. In: *Proceedings of the 26th international conference on world wide web*. 2017, pp. 103–112.
- [469] E. Seto, J. Hua, L. Wu, V. Shia, S. Eom, M. Wang, and Y. Li. “Models of Individual Dietary Behavior Based on Smartphone Data: The Influence of Routine, Physical Activity, Emotion, and Food Environment”. In: *PLOS ONE* 11.4 (Apr. 2016), pp. 1–16.
- [470] C. J. Sewall, T. M. Bear, J. Merranko, and D. Rosen. “How psychosocial well-being and usage amount predict inaccuracies in retrospective estimates of digital technology use”. In: *Mobile Media & Communication* 8.3 (2020), pp. 379–399.
- [471] R. Sharma, V. I. Pavlovic, and T. S. Huang. “Toward multimodal human-computer interface”. In: *Proceedings of the IEEE* 86.5 (1998), pp. 853–869.
- [472] M. Shattell, Y. Apostolopoulos, S. Sönmez, and M. Griffin. “Occupational stressors and the mental health of truckers”. In: *Issues in mental health nursing* 31.9 (2010), pp. 561–568.
- [473] A. Shema and D. E. Acuna. “Show Me Your App Usage and I Will Tell Who Your Close Friends Are: Predicting User’s Context from Simple Cellphone Activity”. In: *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. CHI EA ’17. Denver, Colorado, USA: Association for Computing Machinery, 2017, 2929–2935.
- [474] C. Sheppard-Sawyer, R. McNally, and J. Fischer. “Film-induced sadness as a trigger for disinhibited eating”. In: *The International journal of eating disorders* 28 (Oct. 2000), pp. 215–20.
- [475] S. Shiffman, A. Stone, and H. MR. “Ecological momentary assessment.” In: vol. 4. *Annual Review of Clinical Psychology*. Annual Reviews, 2008, pp. 1–32.
- [476] P. Siirtola, P. Laurinen, J. Roning, and H. Kinnunen. “Efficient accelerometer-based swimming exercise tracking”. In: *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. Paris, France: IEEE, Apr. 2011, pp. 156–161.
- [477] M. Singh. “Mood, food, and obesity”. In: *Frontiers in psychology* 5 (Sept. 2014), p. 925.
- [478] C. J. Skrzynski and K. G. Creswell. “A systematic review and meta-analysis on the association between solitary drinking and alcohol problems in adults”. In: *Addiction* 116.9 (2021), pp. 2289–2303.
- [479] *Smartphone Addiction Facts & Phone Usage Statistics*. 2019. URL: <https://www.bankmycell.com/blog/smartphone-addiction/>.
- [480] K. Smit, M. Groefsema, M. Luijten, R. Engels, and E. Kuntsche. “Drinking motives moderate the effect of the social environment on alcohol use: An event-level study among young adults”. In: *Journal of Studies on Alcohol and Drugs* 76.6 (2015), pp. 971–980.
- [481] J. M. Smyth and A. A. Stone. “Ecological momentary assessment research in behavioral medicine”. In: *Journal of Happiness studies* 4 (2003), pp. 35–52.
- [482] M. Soleymani, M. Pantic, and T. Pun. “Multimodal emotion recognition in response to videos”. In: *IEEE transactions on affective computing* 3.2 (2011), pp. 211–223.
- [483] B. Sornpaisarn, K. Shield, J. Manthey, Y. Limmade, W. Y. Low, V. Van Thang, and J. Rehm. “Alcohol consumption and attributable harm in middle-income South-East Asian countries: Epidemiology and policy options”. In: *International Journal of Drug Policy* 83 (2020), p. 102856.
- [484] D. Southerton, C. Díaz-Méndez, A. Warde, et al. “Behavioural change and the temporal ordering of eating practices: A UK–Spain comparison”. In: *The International Journal of Sociology of Agriculture and Food* 19.1 (2012), pp. 19–36.

Bibliography

- [485] D. Spathis, S. Servia-Rodriguez, K. Farrahi, C. Mascolo, and J. Rentfrow. "Passive mobile sensing and psychological traits for large scale mood prediction". In: *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 2019, pp. 272–281.
- [486] D. Spruijt-Metz and W. Nilsen. "Dynamic models of behavior for just-in-time adaptive interventions". In: *IEEE Pervasive Computing* 13.3 (2014), pp. 13–17.
- [487] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.
- [488] A. K. Stevely, J. Holmes, S. McNamara, and P. S. Meier. "Drinking contexts and their association with acute alcohol-related harm: A systematic review of event-level studies on adults' drinking occasions". In: *Drug and alcohol review* 39.4 (2020), pp. 309–320.
- [489] N. Stiglic and R. M. Viner. "Effects of screentime on the health and well-being of children and adolescents: a systematic review of reviews". In: *BMJ Open* 9.1 (2019).
- [490] E. J. Stinson, S. B. Votruba, C. A. Venti, M. Perez, J. Krakoff, and M. E. Gluck. "Food insecurity is associated with maladaptive eating behaviors and objectively measured overeating". In: *Obesity (Silver Spring, Md.)* 26 (2018), pp. 1841–1848.
- [491] S. Stockwell, M. Trott, M. Tully, J. Shin, Y. Barnett, L. Butler, D. McDermott, F. Schuch, and L. Smith. "Changes in physical activity and sedentary behaviours from before to during the COVID-19 pandemic lockdown: a systematic review". In: *BMJ Open Sport & Exercise Medicine* 7.1 (Jan. 2021), e000960.
- [492] M. Straczekiewicz, P. James, and J.-P. Onnela. "A systematic review of smartphone-based human activity recognition methods for health research". In: *NPJ Digital Medicine* 4.1 (2021), pp. 1–15.
- [493] A. J. Stunkard. "Eating patterns and obesity". In: *Psychiatric quarterly* 33.2 (1959), pp. 284–295.
- [494] X. Su, H. Tong, and P. Ji. "Activity recognition with smartphone sensors". In: *Tsinghua science and technology* 19.3 (2014), pp. 235–249.
- [495] Y. Suhara, Y. Xu, and A. Pentland. "Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks". In: *Proceedings of the 26th International Conference on World Wide Web*. 2017, pp. 715–724.
- [496] S. Sun, B. Zhang, L. Xie, and Y. Zhang. "An unsupervised deep domain adaptation approach for robust speech recognition". In: *Neurocomputing* 257 (2017), pp. 79–87.
- [497] H. Sung, R. L. Siegel, P. S. Rosenberg, and A. Jemal. *Emerging cancer trends among young adults in the USA: analysis of a population-based cancer registry*. 2019.
- [498] Q. Suo, F. Ma, Y. Yuan, M. Huai, W. Zhong, J. Gao, and A. Zhang. "Deep patient similarity learning for personalized healthcare". In: *IEEE transactions on nanobioscience* 17.3 (2018), pp. 219–227.
- [499] T. Szakmany, R. Pugh, M. Kopczyńska, R. M. Lundin, B. Sharif, P. Morgan, G. Ellis, J. Abreu, S. Kulikouskaya, K. Bashir, et al. "Defining sepsis on the wards: results of a multi-centre point-prevalence study comparing two sepsis definitions". In: *Anaesthesia* 73.2 (2018), pp. 195–204.
- [500] M. Tabuchi, T. Nakagawa, A. Miura, and Y. Gondo. "Generativity and Interaction Between the Old and Young: The Role of Perceived Respect and Perceived Rejection". In: *The Gerontologist* 55.4 (Nov. 2013), pp. 537–547.
- [501] R. Taylor. "Interpretation of the correlation coefficient: a basic review". In: *Journal of diagnostic medical sonography* 6.1 (1990), pp. 35–39.

- [502] S. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard. “Personalized multitask learning for predicting tomorrow’s mood, stress, and health”. In: *IEEE Transactions on Affective Computing* 11.2 (2017), pp. 200–213.
- [503] E. Teo, D. Goh, K. M. Vijayakumar, and J. C. J. Liu. “To Message or Browse? Exploring the Impact of Phone Use Patterns on Male Adolescents’ Consumption of Palatable Snacks”. In: *Frontiers in Psychology* 8 (2018), p. 2298.
- [504] 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.” In: *Obesity* 19 8 (2011), pp. 1574–9.
- [505] E. Thomaz, A. Bedri, T. Prioleau, I. Essa, and G. D. Abowd. “Exploring symmetric and asymmetric bimanual eating detection with inertial sensors on the wrist”. In: *Proceedings of the 1st Workshop on Digital Biomarkers*. 2017, pp. 21–26.
- [506] 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*. UbiComp ’15. Osaka, Japan: Association for Computing Machinery, 2015, 1029–1040.
- [507] 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.
- [508] J. Thrul and E. Kuntsche. “The impact of friends on young adults’ drinking over the course of the evening—an event-level analysis”. In: *Addiction* 110.4 (2015), pp. 619–626.
- [509] J. Thrul, F. Labhart, and E. Kuntsche. “Drinking with mixed-gender groups is associated with heavy weekend drinking among young adults”. In: *Addiction* 112.3 (2017), pp. 432–439.
- [510] J. Thrul, S. Lipperman-Kreda, and J. W. Grube. “Do associations between drinking event characteristics and underage drinking differ by drinking location?” In: *Journal of studies on alcohol and drugs* 79.3 (2018), pp. 417–422.
- [511] J. Thrul, S. Lipperman-Kreda, and J. W. Grube. “Do Associations Between Drinking Event Characteristics and Underage Drinking Differ by Drinking Location?” en. In: *Journal of Studies on Alcohol and Drugs* 79.3 (May 2018), pp. 417–422.
- [512] K. Tirri and P. Nokelainen. “Identification of multiple intelligences with the Multiple Intelligence Profiling Questionnaire III”. In: *Psychology Science* 50.2 (2008), p. 206.
- [513] T. Tobin. *What time people typically eat dinner in 12 different places around the world*. Nov. 2018.
- [514] S. G. Trost. “State of the Art Reviews: Measurement of Physical Activity in Children and Adolescents”. In: *American Journal of Lifestyle Medicine* 1.4 (2007), pp. 299–314.
- [515] B. Tulu, C. Ruiz, J. Allard, J. Acheson, A. Busch, A. Roskusku, G. Heeringa, V. Jaskula, J. Oleski, and S. Pagoto. “SlipBuddy: A Mobile Health Intervention to Prevent Overeating”. In: (Jan. 2017).
- [516] P. Turner and C. Lefevre. “Instagram use is linked to increased symptoms of orthorexia nervosa”. In: *Eating and Weight Disorders - Studies on Anorexia, Bulimia and Obesity* 22.2 (Mar. 2017).
- [517] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. “Adversarial discriminative domain adaptation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7167–7176.

Bibliography

- [518] H. Van Der Vorst, R. C. Engels, W. Meeus, and M. Deković. “The impact of alcohol-specific rules, parental norms about early drinking and parental alcohol use on adolescents’ drinking behavior”. In: *Journal of Child Psychology and Psychiatry* 47.12 (2006), pp. 1299–1306.
- [519] H. Van Der Vorst, R. C. Engels, W. Meeus, M. Deković, and J. Van Leeuwe. “The role of alcohol-specific socialization in adolescents’ drinking behaviour”. In: *Addiction* 100.10 (2005), pp. 1464–1476.
- [520] M. D. Van Der Zwaag, C. Dijksterhuis, D. De Waard, B. L. Mulder, J. H. Westerink, and K. A. Brookhuis. “The influence of music on mood and performance while driving”. In: *Ergonomics* 55.1 (2012), pp. 12–22.
- [521] A. J. van Deursen, C. L. Bolle, S. M. Hegner, and P. A. Kommers. “Modeling habitual and addictive smartphone behavior: The role of smartphone usage types, emotional intelligence, social stress, self-regulation, age, and gender”. In: *Computers in Human Behavior* 45 (2015), pp. 411–420.
- [522] T. van Strien, E. L. Gibson, R. Baños, A. Cebolla, and L. H. Winkens. “Is comfort food actually comforting for emotional eaters? A (moderated) mediation analysis”. In: *Physiology & Behavior* 211 (2019), p. 112671.
- [523] 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?” In: *Appetite* 52.2 (2009), pp. 380–387.
- [524] T. van Strien, C. Peter Herman, and M. W. Verheijden. “Eating style, overeating and weight gain. A prospective 2-year follow-up study in a representative Dutch sample”. In: *Appetite* 59.3 (2012), pp. 782–789.
- [525] E. A. Vandewater, M. suk Shim, and A. G. Caplovitz. “Linking obesity and activity level with children’s television and video game use”. In: *Journal of Adolescence* 27.1 (2004). Video Games and Public Health, pp. 71–85.
- [526] K. R. Varshney. *Trustworthy Machine Learning*. Chappaqua, NY, USA: Independently Published, 2022.
- [527] L. Vartanian, N. Reily, S. Spanos, C Herman, and J. Polivy. “Self-reported overeating and attributions for food intake”. In: *Psychology & health* 32 (Jan. 2017), pp. 1–10.
- [528] L. R. Vartanian, N. M. Reily, S. Spanos, L. C. McGuirk, C. P. Herman, and J. Polivy. “Hunger, taste, and normative cues in predictions about food intake”. In: *Appetite* 116 (2017), pp. 511–517.
- [529] C. A. Vereecken, J. Todd, C. Roberts, C. Mulvihill, and L. Maes. “Television viewing behaviour and associations with food habits in different countries”. In: *Public Health Nutrition* 9.2 (2006), 244–250.
- [530] Y. Verma. *A Complete Guide to Sequential Feature Selection*. 2021. URL: <https://analyticsindiamag.com/a-complete-guide-to-sequential-feature-selection/>.
- [531] V. Vickerstaff, R. Z. Omar, and G. Ambler. “Methods to adjust for multiple comparisons in the analysis and sample size calculation of randomised controlled trials with multiple primary outcomes”. In: *BMC medical research methodology* 19.1 (2019), pp. 1–13.
- [532] F. Van de Vijver and N. K. Tanzer. “Bias and equivalence in cross-cultural assessment: An overview”. In: *European Review of Applied Psychology* 54.2 (2004), pp. 119–135.
- [533] S. Visweswaran, G. F. Cooper, and M. Chickering. “Learning Instance-Specific Predictive Models.” In: *Journal of Machine Learning Research* 11.12 (2010).

- [534] P. Voigt and A. Von dem Bussche. “The eu general data protection regulation (gdpr)”. In: *A Practical Guide, 1st Ed., Cham: Springer International Publishing* 10.3152676 (2017), pp. 10–5555.
- [535] T. Vu, F. Lin, N. Alshurafa, and W. Xu. “Wearable Food Intake Monitoring Technologies: A Comprehensive Review”. In: *Computers* 6.1 (2017), p. 4.
- [536] F. Wahle, T. Kowatsch, E. Fleisch, M. Rufer, S. Weidt, et al. “Mobile sensing and support for people with depression: a pilot trial in the wild”. In: *JMIR mHealth and uHealth* 4.3 (2016), e5960.
- [537] M. Walker, L. Thornton, M. D. Choudhury, J. Teevan, C. M. Bulik, C. A. Levinson, and S. Zerwas. “Facebook Use and Disordered Eating in College-Aged Women”. In: *Journal of Adolescent Health* 57.2 (2015), pp. 157–163.
- [538] B. Wammes, S. French, and J. Brug. “What young Dutch adults say they do to keep from gaining weight: self-reported prevalence of overeating, compensatory behaviours and specific weight control behaviours”. In: *Public Health Nutrition* 10.8 (2007), 790–798.
- [539] B. Wammes, B. Breedveld, S. Kremers, and J. Brug. “The ‘balance intervention’ for promoting caloric compensatory behaviours in response to overeating: a formative evaluation”. In: *Health Education Research* 21.4 (Apr. 2006), pp. 527–537.
- [540] R. Wampfler, S. Klingler, B. Solenthaler, V. R. Schinazi, M. Gross, and C. Holz. “Affective State Prediction from Smartphone Touch and Sensor Data in the Wild”. In: *CHI Conference on Human Factors in Computing Systems*. CHI ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022.
- [541] J. Wang, B. Cao, P. S. Yu, L. Sun, W. Bao, and X. Zhu. “Deep Learning Towards Mobile Applications”. In: *CoRR* abs/1809.03559 (2018).
- [542] Q. Wang, Y. Ma, K. Zhao, and Y. Tian. “A comprehensive survey of loss functions in machine learning”. In: *Annals of Data Science* (2020), pp. 1–26.
- [543] R. Wang, M. S. H. Aung, S. Abdullah, R. Brian, A. T. Campbell, T. Choudhury, M. Hauser, J. Kane, M. Merrill, E. A. Scherer, V. W. S. Tseng, and D. Ben-Zeev. “CrossCheck: Toward Passive Sensing and Detection of Mental Health Changes in People with Schizophrenia”. In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp ’16. Heidelberg, Germany: ACM, 2016, pp. 886–897.
- [544] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell. “StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones”. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp ’14. Seattle, Washington: ACM, 2014, pp. 3–14.
- [545] R. Wang, W. Wang, A. daSilva, J. F. Huckins, W. M. Kelley, T. F. Heatherton, and A. T. Campbell. “Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing”. In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2.1 (Mar. 2018), 43:1–43:26.
- [546] S.-C. Wang. “Artificial neural network”. In: *Interdisciplinary computing in java programming*. Springer, 2003, pp. 81–100.
- [547] X. Wang, D. Rosenblum, and Y. Wang. “Context-Aware Mobile Music Recommendation for Daily Activities”. In: *Proceedings of the 20th ACM International Conference on Multimedia*. MM ’12. Nara, Japan: Association for Computing Machinery, 2012, 99–108.

Bibliography

- [548] Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, and O. Russakovsky. "Towards fairness in visual recognition: Effective strategies for bias mitigation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 8919–8928.
- [549] B. Wansink, M. M. Cheney, and N. Chan. "Exploring comfort food preferences across age and gender". In: *Physiology & Behavior* 79.4 (2003), pp. 739–747.
- [550] J. Ward. *How Many Continents In The World? 5,6,7?* 2020. URL: <https://onestep4ward.com/how-many-continents-in-the-world/>.
- [551] A. Warde. *The practice of eating*. John Wiley & Sons, 2016.
- [552] J. Wardle, A. Haase, A. Steptoe, M. Nillapun, K. Jonwutiwes, and F. Bellisle. "Gender Differences in Food Choice: The Contribution of Health Beliefs and Dieting". In: *Annals of behavioral medicine : a publication of the Society of Behavioral Medicine* 27 (May 2004), pp. 107–16.
- [553] D. Watson, L. A. Clark, and A. Tellegen. "Development and validation of brief measures of positive and negative affect: the PANAS scales." In: *Journal of personality and social psychology* 54.6 (1988), p. 1063.
- [554] [www.merriam webster.com. Mood – Definition](https://www.merriam-webster.com/dictionary/mood). 2022. URL: <https://www.merriam-webster.com/dictionary/mood>.
- [555] E. W. Weisstein. "Bonferroni correction". In: <https://mathworld.wolfram.com/> (2004).
- [556] K. Westerterp. "4 - Physical activity and obesity". In: *Food, Diet and Obesity*. Ed. by D. J. Mela. Woodhead Publishing Series in Food Science, Technology and Nutrition. Woodhead Publishing, 2005, pp. 76–85.
- [557] A. White and R. Hingson. "The burden of alcohol use: Excessive alcohol consumption and related consequences among college students". In: *Alcohol research: current reviews* 35.2 (2014), p. 201.
- [558] W. H. O. (WHO). *Obesity and overweight*. 2020. URL: <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight>.
- [559] *Why stress causes people to overeat*. 2018. URL: <https://www.health.harvard.edu/staying-healthy/why-stress-causes-people-to-overeat>.
- [560] J. M. Wiener, R. J. Hanley, R. Clark, and J. F. Van Nostrand. "Measuring the activities of daily living: Comparisons across national surveys". In: *Journal of gerontology* 45.6 (1990), S229–S237.
- [561] D. A. Williamson, D. H. Gleaves, and O. J. Lawson. "Biased perception of overeating in bulimia nervosa and compulsive binge eaters". In: *Journal of Psychopathology and Behavioral Assessment* 13 (1991), pp. 257–268.
- [562] G. Wilson, J. R. Doppa, and D. J. Cook. "Domain Adaptation Under Behavioral and Temporal Shifts for Natural Time Series Mobile Activity Recognition". In: *arXiv preprint arXiv:2207.04367* (2022).
- [563] J. B. de Wit, F. M. Stok, D. J. Smolenski, D. D. de Ridder, E. de Vet, T. Gaspar, F. Johnson, L. Nureeva, and A. Luszczynska. "Food Culture in the Home Environment: Family Meal Practices and Values Can Support Healthy Eating and Self-Regulation in Young People in Four European Countries". In: *Applied Psychology: Health and Well-Being* 7.1 (2015), pp. 22–40.
- [564] S. Wold, K. Esbensen, and P. Geladi. "Principal component analysis". In: *Chemometrics and intelligent laboratory systems* 2.1-3 (1987), pp. 37–52.
- [565] Worldometer. *7 Continents*. 2022. URL: <https://www.worldometers.info/geography/7-continents/>.

- [566] Y. Wu, D. Spathis, H. Jia, I. Perez-Pozuelo, T. Gonzales, S. Brage, N. Wareham, and C. Mascolo. "UDAMA: Unsupervised Domain Adaptation through Multi-discriminator Adversarial Training with Noisy Labels Improves Cardio-fitness Prediction". In: *arXiv preprint arXiv:2307.16651* (2023).
- [567] L. Xu, X. Hao, N. D. Lane, X. Liu, and T. Moscibroda. "More with less: Lowering user burden in mobile crowdsourcing through compressive sensing". In: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 2015, pp. 659–670.
- [568] X. Xu, P. Chikersal, J. M. Dutcher, Y. S. Sefidgar, W. Seo, M. J. Tumminia, D. K. Villalba, S. Cohen, K. G. Creswell, J. D. Creswell, et al. "Leveraging Collaborative-Filtering for Personalized Behavior Modeling: A Case Study of Depression Detection among College Students". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5.1 (2021), pp. 1–27.
- [569] X. Xu, X. Liu, H. Zhang, W. Wang, S. Nepal, Y. Sefidgar, W. Seo, K. S. Kuehn, J. F. Huckins, M. E. Morris, et al. "GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6.4 (2023), pp. 1–34.
- [570] X. Xu, J. Mankoff, and A. K. Dey. "Understanding practices and needs of researchers in human state modeling by passive mobile sensing". In: *CCF Transactions on Pervasive Computing and Interaction* 3.4 (2021), pp. 344–366.
- [571] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo. "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2272–2281.
- [572] T. Yan, D. Chu, D. Ganesan, A. Kansal, and J. Liu. "Fast App Launching for Mobile Devices Using Predictive User Context". In: *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*. MobiSys '12. Low Wood Bay, Lake District, UK: Association for Computing Machinery, 2012, 113–126.
- [573] L. Yang, Y. Balaji, S.-N. Lim, and A. Shrivastava. "Curriculum manager for source selection in multi-source domain adaptation". In: *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer. 2020, pp. 608–624.
- [574] R. Yang and B. Wang. "PACP: A Position-Independent Activity Recognition Method Using Smartphone Sensors". en. In: *Information* 7.4 (Dec. 2016), p. 72.
- [575] Y. Yang, H. Zhang, D. Katabi, and M. Ghassemi. "Change is hard: A closer look at subpopulation shift". In: *arXiv preprint arXiv:2302.12254* (2023).
- [576] W. Yao, Y. Liu, D. Zhou, Z. Pan, M. J. Till, J. Zhao, L. Zhu, L. Zhan, Q. Tang, and Y. Liu. "Impact of GPS signal loss and its mitigation in power system synchronized measurement devices". In: *IEEE Transactions on Smart Grid* 9.2 (2016), pp. 1141–1149.
- [577] K. Yatani. "Effect Sizes and Power Analysis in HCI". In: *Modern Statistical Methods for HCI*. Ed. by J. Robertson and M. Kaptein. Cham: Springer International Publishing, 2016, pp. 87–110.
- [578] X. Ye, G. Chen, Y. Gao, H. Wang, and Y. Cao. "Assisting food journaling with automatic eating detection". In: *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*. 2016, pp. 3255–3262.
- [579] C.-w. You, K.-C. Wang, M.-C. Huang, Y.-C. Chen, C.-L. Lin, P.-S. Ho, H.-C. Wang, P. Huang, and H.-H. Chu. "Soberdiary: A phone-based support system for assisting recovery from alcohol dependence". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2015, pp. 3839–3848.

Bibliography

- [580] *Youth@Night*, <http://www.youth-night.ch/>. 2014. URL: <http://www.youth-night.ch/>.
- [581] Y. Yue, T. Lan, A. G. Yeh, and Q.-Q. Li. “Zooming into individuals to understand the collective: A review of trajectory-based travel behaviour studies”. In: *Travel Behaviour and Society* 1.2 (2014), pp. 69–78.
- [582] T. C. Yun, S. R. Ahmad, and D. K. S. Quee. “Dietary Habits and Lifestyle Practices among University Students in Universiti Brunei Darussalam”. In: *The Malaysian Journal of Medical Sciences: MJMS* 25 (2018), pp. 56–66.
- [583] Ö. Yürür, C. H. Liu, Z. Sheng, V. C. Leung, W. Moreno, and K. K. Leung. “Context-awareness for mobile sensing: A survey and future directions”. In: *IEEE Communications Surveys & Tutorials* 18.1 (2014), pp. 68–93.
- [584] N. Yuval-Davis. *Gender and nation*. Routledge, 2004.
- [585] M. Zeni, I. Zaihrayeu, and F. Giunchiglia. “Multi-Device Activity Logging”. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. UbiComp ’14 Adjunct. Seattle, Washington: Association for Computing Machinery, 2014, 299–302.
- [586] H. Zhang, S. Gashi, H. Kimm, E. Hanci, and O. Matthews. “Moodbook: An Application for Continuous Monitoring of Social Media Usage and Mood”. In: *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. UbiComp ’18. Singapore, Singapore: Association for Computing Machinery, 2018, 1150–1155.
- [587] J. Zhang, D. Li, R. Dai, H. Cos, G. A. Williams, L. Raper, C. W. Hammill, and C. Lu. “Predicting Post-Operative Complications with Wearables: A Case Study with Patients Undergoing Pancreatic Surgery”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6.2 (2022), pp. 1–27.
- [588] S. Zhang, W. Stogin, and N. Alshurafa. “I sense overeating: Motif-based machine learning framework to detect overeating using wrist-worn sensing”. In: *Information Fusion* 41 (2018), pp. 37–47.
- [589] S. Zhang, Y. Zhao, D. T. Nguyen, R. Xu, S. Sen, J. Hester, and N. Alshurafa. “Necksense: A multi-sensor necklace for detecting eating activities in free-living conditions”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.2 (2020), pp. 1–26.
- [590] S. Zhang. “Nearest neighbor selection for iteratively kNN imputation”. In: *Journal of Systems and Software* 85.11 (2012), pp. 2541–2552.
- [591] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar. “Cornnet: Fine-grained emotion recognition for video watching using wearable physiological sensors”. In: *Sensors* 21.1 (2020), p. 52.
- [592] W. Zhang, Q. Shen, S. Teso, B. Lepri, A. Passerini, I. Bison, and F. Giunchiglia. “Putting human behavior predictability in context”. In: *EPJ Data Science* 10.1 (2021), p. 42.
- [593] X. Zhang, F. Zhuang, W. Li, H. Ying, H. Xiong, and S. Lu. “Inferring mood instability via smart-phone sensing: A multi-view learning approach”. In: *Proceedings of the 27th ACM International Conference on Multimedia*. 2019, pp. 1401–1409.
- [594] C. Zheng, W. Y. Huang, S. Sheridan, C. H.-P. Sit, X.-K. Chen, and S. H.-S. Wong. “COVID-19 Pandemic Brings a Sedentary Lifestyle in Young Adults: A Cross-Sectional and Longitudinal Study”. en. In: *International Journal of Environmental Research and Public Health* 17.17 (Aug. 2020), p. 6035.

-
- [595] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy. “Domain generalization: A survey”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).
- [596] Y. Zhou, S. De, W. Wang, R. Wang, and K. Moessner. “Missing data estimation in mobile sensing environments”. In: *IEEE Access* 6 (2018), pp. 69869–69882.
- [597] K. Zhu, X. Zhang, B. Xiang, and L. Zhang. “Exploiting user context and network information for mobile application usage prediction”. In: *Proceedings of the 7th International Workshop on Hot Topics in Planet-scale mObile computing and online Social neTworking*. 2015, pp. 25–30.
- [598] Z. Zhuang and Y. Xue. “Sport-Related Human Activity Detection and Recognition Using a Smartwatch”. en. In: *Sensors* 19.22 (Nov. 2019), p. 5001.
- [599] J. Zou and L. Schiebinger. *AI can be sexist and racist—it’s time to make it fair*. 2018.
- [600] J. Zulueta, A. Piscitello, M. Rasic, R. Easter, P. Babu, S. A. Langenecker, M. McInnis, O. Ajilore, P. C. Nelson, K. Ryan, and A. Leow. “Predicting Mood Disturbance Severity with Mobile Phone Keystroke Metadata: A BiAffect Digital Phenotyping Study”. In: *J Med Internet Res* 20.7 (2018), e241.
- [601] M. D. van der Zwaag, J. H. Janssen, C. Nass, J. H. Westerink, S. Chowdhury, and D. de Waard. “Using music to change mood while driving”. In: *Ergonomics* 56.10 (2013), pp. 1504–1514.

Lakmal Meegahapola

🏠 Martigny, Switzerland

✉ lakmalbuddikalucky@gmail.com [in](#) [🌐](#) [🎓](#) [🐦](#)

Summary

Mobile & Wearable Sensing, Digital Health, Machine Learning, Human-Computer Interaction

I work with **multimodal and high-dimensional real-world data** (smartphone and wearable sensors, electronic health records, self-reports), and techniques including **classic machine learning** and **deep learning** (multi-task learning, domain adaptation, adversarial learning) to examine hard-to-quantify aspects related to **mental wellbeing, health, and behavior** in both **non-clinical** and **clinical** contexts. I am particularly interested in the **domain shift, generalization, personalization, and other ethical aspects** of AI/ML models for human behavior modeling and health time series. To this end, my research involves the entire pipeline of user study design, deployments, data collection, signal processing, and ML-based modeling.

Education

EPFL – Switzerland

2019/06 - 2023/11

PhD, Electrical Engineering | Advisor: Prof. Daniel Gatica-Perez

Thesis: Generalization and Personalization of Machine Learning for Multimodal Mobile Sensing in Everyday Life

University of Moratuwa – Sri Lanka

2014/01 - 2018/01

BSc, Computer Science and Engineering (1st Class Honours) | Advisors: Prof. Dulani Meedeniya, Prof. Indika Perera, Prof. Sampath Jayarathna (ODU)

Thesis: Change Detection and Notification of Web Pages

Experience

IDIAP & EPFL – Switzerland

2019/06 - 2023/11

Doctoral Researcher | Advisor: Prof. Daniel Gatica-Perez | Collaborators: Prof. Fausto Giunchiglia (University of Trento)

Developed machine/deep learning models to sense behavior & health with multimodal sensor data, focusing on generalization, personalization, and domain shift.

Google Research – Mountain View, California, USA

2022/09 - 2022/12

Student Researcher | Managers: Dr. Venky Ramachandran, Dr. Kai Kohlhoff, Dr. Vidhya Navalpakkam

Conducted user studies and developed computer vision algorithms and machine learning models to understand digital well-being with mobile sensing data.

Nokia Bell Labs – Cambridge, UK

2022/06 - 2022/08

Research Intern | Managers: Dr. Michael Eggleston, Prof. Daniele Quercia

Worked on the well-being of humans (in-lab) and pet dogs (in-the-wild) using ML/AI and wearables with time series data from IMU, PPG & EDA sensors.

University of Cambridge – Cambridge, UK

2022/01 - 2022/08

Visiting Researcher | Host: Prof. Cecilia Mascolo

Worked on ensemble machine learning models based on personality diversity for mood inference with multimodal mobile sensor data.

Singapore Management University – Singapore

2018/01 - 2019/05

Research Engineer | Advisor: Prof. Archan Misra | Collaborators: Prof. Leman Akoglu (CMU), Prof. Mirco Musolesi (UCL)

Developed data analytics pipelines to process large-scale mobility traces of people/vehicles to generate real-time analytics.

Innoscripta AG – Munich, Germany

2016/08 - 2017/07

Software Engineering Intern

Developed and deployed web applications with technologies such as AngularJS, PHP, MySQL, and AWS.

Selected Honors and Awards (full list here)

ACM UbiComp Gaetano Borriello Outstanding Student Award Finalist (top 4) - ACM UbiComp

2023

Awarded to a student who is making outstanding research contributions in the field of ubiquitous, mobile, and wearable computing.

Distinguished Paper Award - ACM IMWUT/UbiComp

2023

IMWUT assigns this award to 3-4% of all papers published by the journal in the previous year. 8 papers were awarded in 2023.

President's Award for Scientific Research - Awarded by his Excellency the President of Sri Lanka

2023

A limited number of highly rated research papers across all disciplines published by Sri Lankan researchers in 2020 were awarded in 2023.

Outstanding Reviewer (x2) - ACM CHI

2024

Outstanding Reviewer - ACM ICMI

2023

Outstanding Reviewer (x3) - ACM IMWUT/UbiComp

2022-2023

IBM PhD Fellowship - EPFL Nominee

2022

Exemplary Reviewer (top 3%) - IEEE Wireless Communications Letters Journal

2020

Best Paper Presentation - ACM ICSCA

2018

Best Paper Finalist (top 1%) - ISCA CAINE

2018

Runner Up & 10,000 USD Prize - Microsoft Imagine Cup World Finals, Seattle, USA

2016

Second Runner Up & Bronze Award - 18th National Best Quality ICT Awards (NBQSA), Sri Lanka

229

2016

Winner - Microsoft Imagine Cup National Finals (Innovation), Sri Lanka

2016

Winner - Microsoft Imagine Cup National Finals (Games), Sri Lanka

2016

Winner - HackaTUM Hackathon (Media-Saturn track), TU Munich, Germany

2016

Winner - Google I/O Extended Hackathon, Sri Lanka	2015
National Merit Scholarship Worth 240,000 LKR for Undergraduate Studies (top 1%) - Government of Sri Lanka	2012
Bronze Medalist (top 1%) - National Physics Olympiad, Sri Lanka	2012
Sri Lanka Team Member & Team's Highest Scorer - International Olympiad on Astrophysics (IOAA), Katowice, Poland	2011
Winner & Gold Medalist - National Astrophysics Olympiad, Sri Lanka	2011

Selected Publications (full list [here](#))

C - Conference, J - Journal, * - equal contribution

[C] L. Meegahapola*, A. Ruben*, D. Gatica-Perez

Learning about Social Context from Smartphone Data: Generalization Across Countries and Daily Life Moments
Under Review

[J] L. Meegahapola, H. Hassoune, D. Gatica-Perez

M3BAT: Unsupervised Domain Adaptation for Multimodal Mobile Sensing with Multi-Branch Adversarial Training
Under Review

[J] L. Meegahapola*, W. Bangamuarachchi*, A. Chamantha*, H. Kim, S. Ruiz-Correa, D. Gatica-Perez

Inferring Mood-While-Eating with Smartphone Sensing and Community-Based Model Personalization [↗](#)
Under Review

[C] N. Kammoun, L. Meegahapola, D. Gatica-Perez

Understanding the Social Context of Eating with Multimodal Smartphone Sensing: The Role of Country Diversity [↗](#)
ACM ICMI — 25th ACM International Conference on Multimodal Interaction. Paris, France. 2023

[J] L. Meegahapola, W. Droz, D. Gatica-Perez et al.

Generalization & Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries [↗](#) [📄](#)
ACM IMWUT (UbiComp) — Proceedings of the ACM on Interactive, Mobile, Wearable, and Ubiquitous Technologies. Cancun, Mexico. 2023

🏆 **Distinguished Paper Award (1%)**

🔗 **Coverage:** [EPFL News](#) [↗](#), [Idiap News](#) [↗](#)

[C] A. Nanchen, L. Meegahapola, W. Droz, D. Gatica-Perez

Keep Sensors in Check: Disentangling Country-Level Generalization Issues in Mobile Sensor-Based Models with Diversity Scores [↗](#)
AAAI/ACM AIES — AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society. Vancouver, Canada. 2023

[C] Y. Amarasinghe, D. Sandaruwan, T. Madusanka, I. Perera, L. Meegahapola

Multimodal Earable Sensing for Human Energy Expenditure Estimation [↗](#)
IEEE EMBC — 45th International Conference of the IEEE Engineering in Medicine and Biology Society. Sydney, Australia. 2023

[C] L. Meegahapola, M. Constantinides, Z. Radivojevic, H. Li, M. Eggleston, D. Quercia

Quantified Canine: Inferring Dog Personality From Wearables [↗](#) [📄](#)
ACM CHI — ACM Conference on Human Factors in Computing Systems. Hamburg, Germany. 2023

🔗 **Coverage:** [Nokia Bell Labs](#) [↗](#) [🐦](#) [↗](#)

[C] L. Meegahapola*, A. Karim*, D. Gatica-Perez et al.

Complex Daily Activities, Country-Level Diversity and Smartphone Sensing: A Five-Country Study [↗](#) [📄](#)
ACM CHI — ACM Conference on Human Factors in Computing Systems. Hamburg, Germany. 2023

[C] E. Bouton-Bessac, L. Meegahapola, D. Gatica-Perez

Your Day in Your Pocket: Complex Activity Recognition from Smartphone Accelerometers [↗](#)
EAI PervasiveHealth — 16th EAI International Conference on Pervasive Computing Technologies for Healthcare. Thessaloniki, Greece. 2022

[J] L. Meegahapola, F. Labhart, T.T. Phan, D. Gatica-Perez

Examining the Social Context of Alcohol Drinking in Young Adults with Smartphone Sensing [↗](#) [📄](#)
ACM IMWUT (UbiComp) — Proceedings of the ACM on Interactive, Mobile, Wearable, and Ubiquitous Technologies. Virtual. 2021

[J] L. Meegahapola, S.R. Correa, V.C. Valero, E.H. Huerfano, L.A. Rivera, R. Acosta, D. Gatica-Perez

One More Bite? Inferring Food Consumption Level of College Students using Smartphone Sensing & Self-Reports [↗](#) [📄](#)
ACM IMWUT (UbiComp) — Proceedings of the ACM on Interactive, Mobile, Wearable, and Ubiquitous Technologies. Virtual. 2021

🔗 **Coverage:** [PYMNTS](#) [↗](#), [EI Universal](#) [↗](#), [Plano Informativo](#) [↗](#), [EI Portal](#) [↗](#), [Televisa San Luis Potosi](#) [↗](#)

[J] L. Meegahapola, D. Gatica-Perez

Smartphone Sensing for the Well-being of Young Adults: A Review [↗](#) [📄](#)
IEEE Access — 2021 [IF: 3.9, h5-index: 233]

[J] F. Labhart, L. Meegahapola*, S. Muralidhar*, B. Masse*, E. Kuntsche, D. Gatica-Perez

Exploring Different Methods to Measure Brightness, Loudness & Attendance & Their Associations w/ Alcohol Use using Short Video Clips [↗](#)
PLOS ONE — 2021 [IF: 3.75, h5-index: 212]

🔗 **Coverage:** [LeNouvelliste](#) [↗](#), [20min.ch](#) [↗](#), [Idiap News](#) [↗](#), [ictjournal.ch](#) [↗](#), [netzwoche.ch](#) [↗](#), [rhonelfm.ch](#) [↗](#)

[C] L. Schelenz, I. Bison*, M. Busso*, A. de Gotzen*, D. Gatica-Perez*, F. Giunchiglia*, L. Meegahapola*

The Theory, Practice, and Ethical Challenges of Designing a Diversity-Aware Platform for Social Relations [↗](#) [📄](#)
AAAI/ACM AIES — AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society. Virtual. 2021

[J] V.G. Mallawaarachchi, L. Meegahapola, R.M. Alwis, E.H. Nimalarathna, D. Meedeniya, S. Jayarathne
Change Detection and Notification of Webpages: A Survey [↗](#)

ACM Computing Surveys — 2020 [IF: 14.3, h5-index: 131]

🏆 **President's Award for Scientific Research, awarded by his excellency the President of Sri Lanka**

[C] L. Meegahapola, S.R. Correa, D. Gatica-Perez

Protecting Mobile Food Diaries From Getting Too Personal: A Multi-Task Learning Based Approach [↗](#)

ACM MUM — 19th International Conference on Mobile and Ubiquitous Multimedia. Virtual. 2020

[C] L. Meegahapola, S.R. Correa, D. Gatica-Perez

Alone or with Others? Understanding Eating Episodes of College Students with Mobile Sensing [↗](#)

ACM MUM — 19th International Conference on Mobile and Ubiquitous Multimedia. Virtual. 2020

[C] T. Kandappu, A. Mehrotra, A. Misra, M. Musolesi, S.F. Cheng, L. Meegahapola

PokeME: Applying Context-Driven Notifications to Increase Worker Engagement in Mobile Crowd-sourcing [↗](#)

ACM SIGIR CHIIR — 5th ACM SIGIR Conference on Human Information Interaction and Retrieval. Vancouver, Canada. 2020

[C] L. Meegahapola, T. Kandappu, K. Jayarajah, L. Akoglu, S. Xiang, A. Misra

BuSCOPE: Fusing Individual & Aggregated Mobility Behavior for "Live" Smart City Services [↗](#) [📄](#)

ACM Mobisys — 17th ACM International Conference on Mobile Systems, Applications, and Services. Seoul, South Korea. 2019

[C] L. Meegahapola, N. Athaide, K. Jayarajah, S. Xiang, A. Misra

Inferring Accurate Bus Trajectories from Noisy Estimated Arrival Time Records [↗](#)

IEEE ITSC — 22nd IEEE Intelligent Transportation Systems Conference. Auckland, New Zealand. 2019

[C] L. Meegahapola, V.G. Mallawaarachchi, R.M. Alwis, E.H. Nimalarathna, D. Meedeniya, S. Jayarathne

Random Forest Classifier based Scheduler Optimization for Search Engine Web Crawlers [↗](#)

ACM ICSCA — 7th ACM International Conference on Software and Computer Applications. Kuantan, Malaysia. 2018

🏆 **Best Presentation Award** [↗](#)

[C] L. Meegahapola, I. Perera,

Enhanced In-Store Shopping Experience through Smartphone based Mixed Reality Application [↗](#)

ICTer — 17th IEEE International Conference on Advances in ICT for Emerging Regions. Colombo, Sri Lanka. 2017

Selected Press Coverage (full list here)

Two Young Researchers Garner Prestigious International Awards

EPFL News [↗](#), Idiap News [↗](#)

2023/10

AI and an app to understand drinking habits among young adults

lenouvelliste [↗](#), 20min.ch [↗](#), planetesante [↗](#), Idiap News [↗](#), ictjournal.ch [↗](#), netzwoche.ch [↗](#), rhonefm.ch [↗](#)

2021/04

Evaluating effects of COVID-19 lockdown on well-being - Civique mobile app

Civique [↗](#), EPFL News [↗](#), 24heures [↗](#), lfm.ch [↗](#), MirageNews [↗](#)

2020/04

When Social Networks Become "Diversity-Aware" Research Platforms

PYMNTS [↗](#)

2020/02

WENET Behavioral Mobile Sensing App Mexico Pre-pilot

El Universal [↗](#), Plano Informativo [↗](#), El Portal [↗](#), Televisa San Luis Potosi [↗](#)

2019/12

Mr 360 | Lakmal Meegahapola

olivescript.com [↗](#)

2018/08

2016 Throwback: The Biggest Moments In Sri Lankan Tech

readme.lk [↗](#)

2017/01

Asia teams impress at Imagine Cup 2016

Microsoft News [↗](#)

2016/08

Bit Masters & Sri Lanka win big at the Imagine Cup Finals

readme.lk [↗](#), dailymirror [↗](#)

2016/07

AMPLUS - smart ad display software wins Jaffna-based YGC

sundaytimes [↗](#), ft.lk [↗](#)

2015/12

Invited Talks

Multimodal Mobile Sensing and Machine Learning for Healthcare: From Generalization to Personalization

ETH Zurich, Switzerland (x2)

2023/06

Nokia Bell Labs, Cambridge, UK

2023/06

Swiss Data Science Centre, Switzerland

2023/05

Fraunhofer, Germany

2023/05

University of Oxford, UK

2023/05

University of Lausanne, Switzerland | UNIL Digital Week

2022/02, 2023/02

University of Kelaniya, Sri Lanka | Planery Talk at ICAPS 2022 Conference

2022/10

University of Trento, Italy | Knowdive Seminars Series

231 2021/12

Sensing Humans and Animals with Wearable Devices

Nokia Bell Labs, Cambridge, UK | Science & Spritz Volume 14

2022/07

Low Cost, Easy To Use, Intelligent Digital Signage Platform for Targeted Advertising

Microsoft, Redmond, USA | Microsoft Imagine Cup World Finals
Google I/O Extended, Colombo, Sri Lanka

2016/07
2015/04

Teaching

EPFL – Lausanne, Switzerland

Teaching Assistant - DH500 Computational Social Media - Spring 2021, Spring 2022, Spring 2023
Teaching Assistant - Computational Social Science - EPFL Summer School - Summer 2021

WENET – Online

Lecturer - Mini Lecture Series on "Learning Context with Smartphone Sensing" - Spring 2021

Service

Program Committee

27th ACM International Symposium on Wearable Computers (ISWC)	2023
25th ACM International Conference on Multimodal Interaction (ICMI)	2023
ACM Ubicomp - FairComp Workshop	2023
ACM Ubicomp - WellComp Workshop	2022, 2023
ACM CHI - EmpathiCH Workshop	2023
21st International Conference on Mobile and Ubiquitous Multimedia (MUM) – Poster Track	2022

Reviewer

ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT, UbiComp)	2019, 2020, 2021, 2022, 2023
ACM Conference on Human Factors in Computing Systems (CHI)	2020, 2023, 2024
IEEE Transactions on Affective Computing	2023
ACM Transactions on Computing for Healthcare (HEALTH)	2022
ACM Computing Surveys (CSUR)	2022
IEEE Wireless Communications Letters	2019, 2020

Student Supervision

Y. Amarasinghe, D. Sandaruwan, T. Madusanka (IEEE EMBC)	BSc CSE (Thesis Project) at UoM, 2022/08 - 2023/06
H. Hassoune (Under Review)	MS Computer Science (Semester Project) at EPFL, 2023/02 - 2023/06
N. Kammoun (ACM ICMI)	MS Data Science (Semester Project) at EPFL, 2022/02 - 2022/06
M. Guido	BS Computer Science (Semester Project) at UZH, 2022/02 - 2022/06
W. Bangamuarachchi, A. Chamantha (IEEE Access)	BSc CSE (Thesis Project) at UoM, 2021/04 - 2022/04
E. Bouton-Bessac (EAI PervasiveHealth)	MS Microengineering (Thesis Project) at EPFL, 2021/09 - 2022/02
M. A. Ruben (Under Review)	MS Digital Humanities (Semester Project) at EPFL, 2021/09 - 2022/02
K. Assi (ACM CHI)	MS Data Science (Semester Project) at EPFL, 2021/09 - 2022/02
H. Kim (Under Review)	MS Digital Humanities (Semester Project) at EPFL, 2021/03 - 2021/08
D. Rathnayake, D. Pubudumal, A.K. De Silva, A. Manjitha (IEEE MASS)	BSc CSE (Thesis Project) at UoM, 2019/03 - 2020/03

UZH: University of Zurich – Switzerland, CSE: Computer Science and Engineering, UoM: University of Moratuwa – Sri Lanka, SMU: School of Information Systems – Singapore Management University – Singapore, NUS: National University of Singapore

Familiar Languages, Platforms, Technologies and Skills

Languages	Python, R, Java
ML	Scikit-Learn, Keras, Tensorflow, Numpy, Pandas, SciPy, matplotlib
Other	Signal Processing, Audio Processing, Azure, Google Cloud Platform, Biomedical, Classification, Regression, Clustering, Augmented Reality (AR), Virtual Reality (VR), Unity Game Engine, User Research, AB Testing, Experiment Design Time Series, Multimodal, SQL, Git