Sensing Eating Events in Context: A Smartphone-Only Approach

WAGEESHA BANGAMUARACHCHI[1,*], **ANJU CHAMANTHA**[1,*], **LAKMAL MEEGAHAPOLA**[2,3,*], **SALVADOR RUIZ-CORREA**[4], **INDIKA PERERA**[1], (Senior Member, IEEE), **AND DANIEL GATICA-PEREZ**[2,3], (Member, IEEE)

[1]Department of Computer Science and Engineering, University of Moratuwa, Moratuwa 10400, Sri Lanka
[2]Social Computing Group, Idiap Research Institute, 1920 Martigny, Switzerland
[3]School of Engineering, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland
[4]Youth Innovation Laboratory, IPICYT, San Luis Potosí 78216, Mexico

Corresponding author: Lakmal Meegahapola (lakmal.meegahapola@epfl.ch)

ABSTRACT 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 paper 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 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 AUROC (area under the receiver operating characteristics curve) of 0.65 using subject-independent machine learning models, which can be further improved up to 0.81 for subject-dependent 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.

INDEX TERMS Smartphone sensing, mobile sensing, eating behavior, food diary, mobile health, automatic dietary monitoring, diet monitoring, eating event, eating episode, machine learning, personalization.

I. 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 [1]–[3]. 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 [4]. On an

The associate editor coordinating the review of this manuscript and approving it for publication was Giuseppe Desolda.

individual level, food diaries help users with self-awareness, self-monitoring, and behavior change, and have also helped people with weight loss goals [4], [5]. At the population level, they help researchers to conduct large-scale studies to understand population-level food consumption [4]. Food diaries originated as a pencil-and-paper based technique [6], but in recent times mobile food diaries have become popular, and widely adopted commercial mobile health (mHealth) apps such as Samsung Health [7] and MyFitnessPal [8] 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) [4], [5], [9]. Such drawbacks have called for tools and techniques to automatically recognize eating events, as this would allow reminding users to report food intake on time. Prior studies in mobile sensing have used wrist wearables [5], [10], [11], jaw-bone wearables [12], earables [3], [13], necklaces [14], [15], and other sensing modalities [16], [17] 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 [18]. For example, 96% of young adults aged 18-29 in the United States own a smartphone [19], and Nutrition and Diet apps has become the second most common app category among mHealth app users, just behind Fitness apps [18]. Prior studies in mobile sensing have looked into improving mobile food diaries with context-awareness [9], [20]–[22]. 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 [23]. 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 [20], social context [21], or food consumption level [9], a holistic food diary could be maintained with minimal user input. This vision has been discussed in prior dietary monitoring [23] and sensing [24] literature.

Self-Report Validation: A challenge in self-report-based questionnaires is under-reporting, i.e., participants failing to report eating events [23]. 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 [9], [20]–[22]. These inferences could help provide context-aware interventions and feedback to app users [23]. 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 [23]. Furthermore, a recent study discussed the detection of eating events using wearables to trigger further data capture on mobile food diaries [25]. 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 [21], food consumption level [9], food category and type [20]), the use of smartphones as standalone devices to infer eating events remains as an open issue. In this paper, we examine whether smartphone sensing features could be used to classify time windows as corresponding to eating vs. non-eating events using subject-dependent, hybrid, and subject-independent machine learning models [26].

We pose two research questions:

RQ1: What situational contexts and behaviors around eating and non-eating events can be observed by analyzing 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 paper 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 dataset of over 12000 events provided by 58 college students in Mexico is used. We show 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 define and evaluate the task of inferring eating and non-eating events using passive smartphone

sensing data, obtaining an AUROC of 0.65 with subject-independent models, which can be increased to 0.81 with subject-dependent models. Moreover, we show that feature selection plays a key role when training subject-dependent 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 subject-dependent models. The results illustrate the potential of using passive smartphone sensing for building context-aware and automated mobile food diaries.

This paper is organized as follows. Section II, describes the background and discusses related work. In Section III, the study design, data collection procedure, and feature extraction techniques are provided. Section IV, presents a descriptive analysis and a statistical analysis of the dataset. The inference task is defined and evaluated in Section V. A number of important issues are discussed in Section VI. Finally, the paper is concluded in Section VII.

II. BACKGROUND AND RELATED WORK
A. NUTRITION SCIENCE PERSPECTIVE
1) EATING AS A HOLISTIC EVENT
Guided by how and why people eat, Bisogni *et al.* [27] 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.* [28] 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 the eating behavior [29], [30]. Ma *et al.* [31] 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.* [32] 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 [23], going with the terminology regarding holistic eating behavior, throughout this paper, the term *eating episode* is used as the actual time period of food intake. Moreover, the time period around the food intake that 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 captured with a time window. This definition is in line with prior work in mobile sensing that looked into characterizing eating and drinking events [9], [20], [21], [33]. More details regarding how these terms are operationalized can be found in Section III-C.

2) 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 [24]: 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) [34], [35], activity levels and types (behavior) [7], [36], sociability and social context (person, context) [37]. In the context of eating behavior, these pillars of data can be mapped to the eight dimensions proposed by Bisogni *et al.* [27]. Gatica-Perez *et al.* [38] 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 [1], [9], [20], [33], [38]–[40]. Leveraging these relationships, this study 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 1. Hence, we hypothesize that date and time, application usage, screen usage, activity level, and movements are indicative of eating and non-eating.

B. MOBILE FOOD DIARIES

Food diaries are essential to understanding individual and population-level eating practices [4]. 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 [4]. Some recent mobile food diaries have also looked into easing the process of logging using speech and photos [41]. Going a step further, some studies looked into estimating the calorie intake [42] and also calorie deficit by combining information from food diaries and passively detected physical activities [43]. Some popular mHealth apps that provide food journaling functionalities include Samsung Health, Lose It!, MyFitnessPal, EasyDietDiary, and SparkPeople [4], [9]. 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 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 [4], [20]. Hence, to overcome this barrier, researchers have come up with automated eating detection systems using different mobile sensing modalities [10], [14], [15], [20], [25]. 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 study, too, is similar. However, the main difference is that this study only use smartphones to make the inference, while previous studies used wearables.

C. MOBILE SENSING FOR EATING BEHAVIOR MONITORING

Mobile sensing studies for eating behavior monitoring could be segregated into two main categories [24]: (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.* [14] 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, and precision of 78.2%, and a recall of 72.5% in a free-living study. Bedri *et al.* [3] 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.* [25] 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 assessments (EMA) in the smartphone to capture additional contextual and behavioral information about the eating episode. Thomaz *et al.* [10] 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.* [2] 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 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.

D. 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.* [48] 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.* [49] 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.* [20] 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.* [9] showed that self-reported food consumption levels (eating as usual, overeating, undereating) could be inferred with an accuracy of 83.49% in a three-class inference task using only smartphone sensing features with
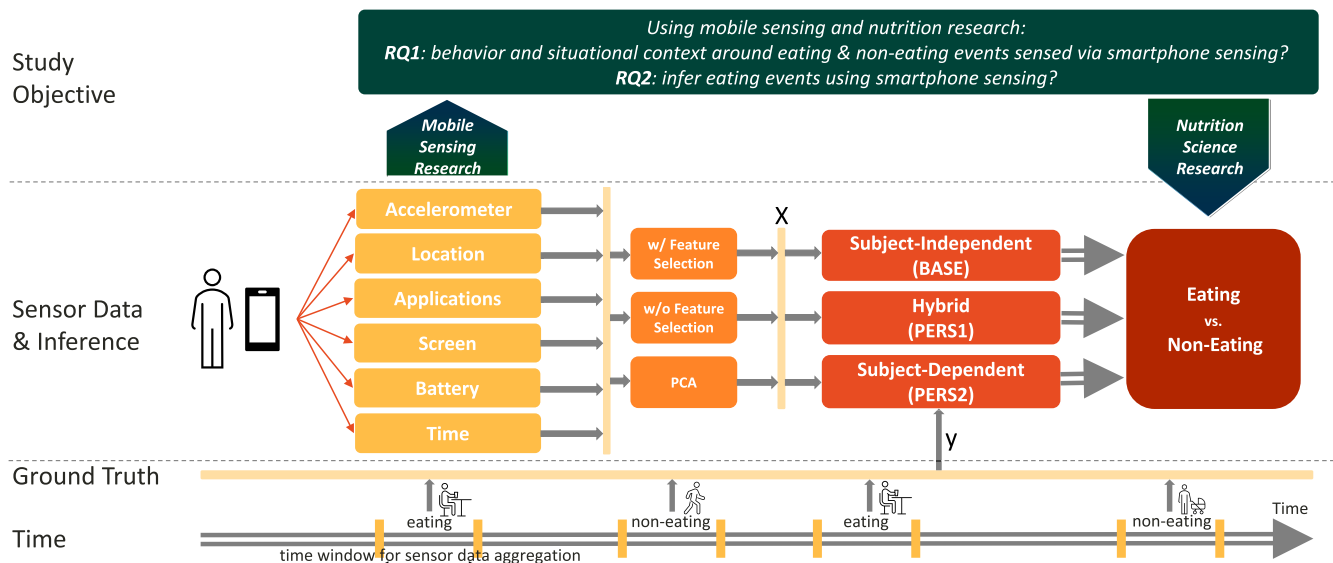
FIGURE 1. Objective of the study.

TABLE 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. [25]	28	✓	✓	Smartwatch (Wrist)	Accelerometer	Hand Movement
Kadomura et al. [44]	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. [45]	28	✓	✓	Bluetooth headset (Ear)	Microphone	Mastication
						Deglutition
Chun et al. [12]	23	✓	✓	Jawbone-mounted (Jawbone)	Accelerometer & Gyroscope	Mastication
Rahman et al. [16]	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. [10]	28	✓	✓	Smartwatch (Wrist)	Accelerometer	Hand Movement
Bi et al. [13]	24	✓	✓	Ear-mounted Sensor (Ear)	Microphone	Chewing Sound
Ye et al. [5]	7		✓	Pebble Smartwatch (Wrist)	Accelerometer	Hand to mouth eating gestures
Mondol et al. [46]	74		✓	Smartwatch (Wrist)	Accelerometer	Hand Movement
					EMG sensor	Motion & EMG activities
Chun et al. [14]	17	✓	✓	Necklace (Nect)	Proximity sensor	Head & Jawbone movements
Thomaz et al. [11]	14	✓		Wrist-wearable (Wrist)	Accelerometer & Gyroscope	Wrist Movements
Zhang, et al. [15]	20		✓	Necklace (Neck)	Accelerometer, Gyroscope, Magnetometer, Proximity, Ambient light	Head Movement and Chewing
Bedri et al. [3]	10	✓	✓	Ear wearable (Ear)	Microphone & Proximity & IMU	Chewing
Merck et al. [17]	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. [47]	43	✓		Ear wearable (Ear)	Microphone & Proximity & IMU	Chewing
Our Study	58		✓	Smartphone (Any)	Multimodal	Behavior and Context

a one hour time window around eating episodes. In another study, Meegahapola *et al.* [21] 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.* [27], 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 study, 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 study

complements them well in showing the potential of using only smartphones for detecting eating events.

Finally, as shown in Figure 1, the methodology was rigorously evaluated using different sensor features, feature selection techniques, model types, and personalization techniques.

III. DATA, SMARTPHONE FEATURES, AND DEFINITION OF EATING/NON-EATING EPISODES

A. DATASET

We used a dataset regarding the eating behavior of young adults from our previous work [9]. This dataset consists of self-reports and smartphone sensing data regarding eating and non-eating events from college students (mean age 23.4, 44% men) in Mexico. A mobile app called i-Log [50] was used to collect data from participants. An entry questionnaire was given to participants to collect demographic information. During the deployment, time diaries captured details regarding food intake similar to a mobile food diary and end-of-the-day surveys that captured additional details regarding activity, sleep, and mood. The users were sent notifications three times a day to fill out a food intake questionnaire which asked about the number of eating episodes they had within the last four hours. Then the users were asked to provide data regarding the last food intake which includes how long ago the eating episode had occurred, food categories (meat, fish, bread etc.), social context of eating (alone, with a date, with a group of friends etc.), semantic eating location (home, university, restaurant etc.), concurrent activities (reading, socializing, watching TV etc.), mood and stress level at the time of eating (5-point scale) [9].

Experiments were carried out in two phases. The first phase was from September to October 2019, with the participation of 29 young adults. The second phase was from November to December 2019, where 55 additional young adults took part. Participant inclusion criteria meant that they owned an Android smartphone and did not have eating disorders like bulimia or anorexia. After data cleaning and pre-processing, the final dataset used for this study consists of 12016 self-reports from 58 participants, out of which 1837 are eating events, and 10179 are non-eating events, for an average of 207 total events (eating or non-eating) per participant.

B. COLLECTING GROUND TRUTH AND PASSIVE SENSING DATA

The mobile app was designed to capture retrospective self-reports. Capturing retrospective reports is important because in-situ self-reports might alter the normal behavior of participants during eating episodes, which would add noise to sensed data. In addition, dietary recall techniques are common in eating behavior studies (e.g. 24H dietary recall [55], [56]). 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 a food intake (note that a far lesser four-hour window was used in this study to capture
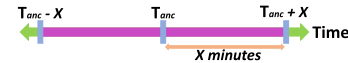

FIGURE 2. Time window for sensor data aggregation.

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 2.

C. 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 2. According to Figure 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 for ubicomp studies regarding eating and drinking behavior that considers the context and behavior of participants in addition to the actual eating/drinking episode [9], [20], [33], [40]. This does not mean that the time of inference should be 30 minutes after the end of eating. In reality, it could be much lesser 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 VI 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

TABLE 2. Summary of 40 features extracted from smartphone sensors.

Sensor	Sensor Description
– Acronym (# of features)	Example Features
Location	Using location data, the radius of gyration [51], [52] 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 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 [9].
– 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 [1]. 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 [33], [53]. Hence, the average battery level during the one hour was calculated. In addition, whether the phone is charging or not 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 [33], [53], 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 [54].
– TIME (3)	hours_elapsed, minutes_elapsed, weekend

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 >= T_{anc} and T_{anc} >= T_{sr}-210). This is to ensure that there are sufficient sensor 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 [23]. Therefore, for clarity, throughout this study, the one-hour time window reported by the participant as they had food during that time is referred 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 [9], [20], [21], [33], [40]. Finally, a one-hour time window in which food is not consumed by participants is referred as a ***non-eating*** event.

IV. DESCRIPTIVE AND STATISTICAL ANALYSIS OF SENSOR DATA AND EATING EVENTS (RQ1)

A. ACCELEROMETER FEATURES

Figure 3 shows accelerometer data distributions for all users and two randomly selected users. In Figure 3a, accelerometer data distributions for eating and non-eating events look
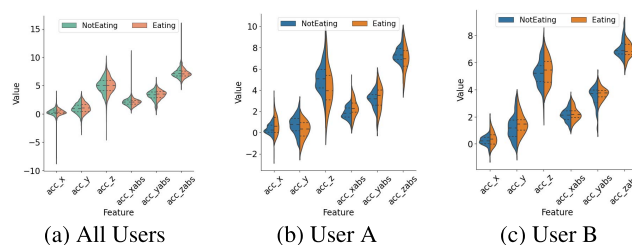

(a) All Users   (b) User A   (c) User B

FIGURE 3. Violin plots of six selected accelerometer features for all the users and for two randomly selected users.

similar when all the users are considered in general, with minimal differences between means in all six features. However, for individual users in Figure 3b and Figure 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.

B. LOCATION FEATURE

Figure 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 corresponding to eating events. This finding is coherent with prior work that said people might travel a
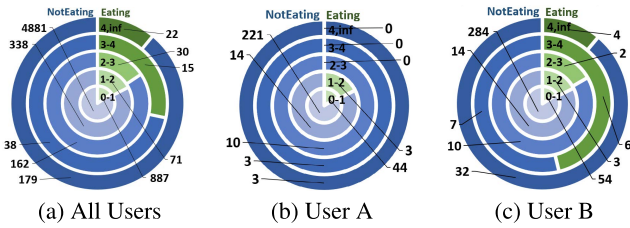
FIGURE 4. Distribution of eating and non-eating events for different radius of gyration values for all the users and for two randomly selected users.
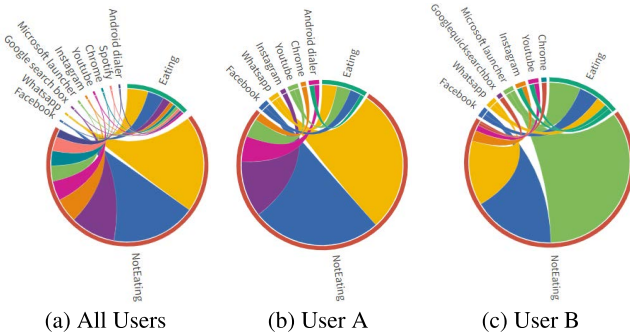

FIGURE 5. Distribution of app usage for eating and non-eating events for all users and for two randomly selected users.

sizable distance at noon in search of food, regardless of the geographical location [31], [57]. However, when distributions of two randomly selected users are considered (Figure 4b and Figure 4c), even though the distribution of *user B* is almost the same as *all-users distribution* in Figure 4a, *user A* has a different distribution. That user has eating events only from 0 to 2 radius of gyration values, suggesting 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.

C. APPLICATION USAGE FEATURES

Figure 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 5b and Figure 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 [58]. Furthermore, app usage behavior has shown good performance for other eating and drinking behavior-related tasks [1]. In addition, prior work has proposed that app usage can be used to gain an understanding of user context [59]. 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.

D. STATISTICAL SIGNIFICANCE OF FEATURES

Table 3 shows the top ten features that help discriminate eating and non-eating, when sorted in the descending order according to the t-statistic [60]. In addition, the p-values [61] are also given. Further, since prior work has highlighted the lack of informativeness in p-values [62], we also provide Cohen's-d (effect size) [63] with 95% confidence interval (CI) [64] to quantify the statistical significance of features. The rule of thumb for interpreting Cohen's-d values is as follows: $0.2 =$ small effect size, $0.5 =$ medium effect size, and $0.8 =$ large effect size. In addition, if the confidence interval does not overlap with zero, the difference between the two groups could be considered significant [62]. Hence, for the two-class problem of detecting eating and non-eating, this analysis aims to examine features that help discern between the two classes.

The top feature by t-statistics is acc_y. Mean values for eating (1.04) and non-eating (0.92) for acc_y, show that the mobile phone activity around the y-axis is higher when eating than during non-eating events. This observation is moderately supported by the t-statistic (5.02) and effect size (0.12) values. Moreover, similar behavior can be observed in the next top five features, all of which had t-statistics above three. However, all of them had below small effect sizes. The WhatsApp app usage and the feature that captures whether the day is a weekend or not too had t-statistic values above two. Furthermore, all features included in Table 3 had 95% confidence intervals not overlapping zero. These results suggest that smartphone sensing features from modalities such as ACC, TIME, and APP could be useful to discern between eating and non-eating events, albeit with limited statistical significance when considering the full participant cohort.

TABLE 3. t-statistic, p-value (p-value $\leq 10^{-4}$:****; p-value $\leq 10^{-3}$:***; p-value $\leq 10^{-2}$:**), and Cohen's-d with 95% confidence interval calculated using 12016 datapoints. Top 10 features are shown in the decreasing order of t-statistic.

| Feature | eating | | non-eating | | t-statistic | p | cohen's-d | Feature | eating | | non-eating | | t-statistic | p | cohen's-d |
	mean	std	mean	std			[95% CI]		mean	std	mean	std			[95% CI]
acc_y	1.04	0.89	0.92	0.92	(+) 5.0199	****	0.1276 [0.08, 0.18]	$acc_y abs\_aft$	3.39	0.92	3.30	0.98	(+) 3.4838	***	0.0897 [0.04, 0.14]
$acc_y abs\_bef$	3.37	0.93	3.24	1.01	(+) 4.9662	****	0.1292 [0.08, 0.18]	$acc_z abs\_bef$	7.15	0.83	7.23	0.95	(-) 3.4267	***	0.0904 [0.04, 0.14]
$acc_y abs$	3.38	0.83	3.27	0.91	(+) 4.5727	****	0.1189 [0.07, 0.17]	whatsapp	0.13	0.34	0.11	0.31	(+) 2.8487	**	0.0698 [0.02, 0.12]
$acc_y aft$	1.06	1.06	0.94	1.06	(+) 4.5035	****	0.1135 [0.06, 0.16]	weekend	0.03	0.04	0.30	0.46	(-) 2.7505	**	0.0703 [0.02, 0.12]
$acc_y bef$	1.01	0.99	0.90	1.03	(+) 4.3225	****	0.1103 [0.06, 0.16]	$acc_z abs$	7.15	0.75	7.21	0.08	(-) 2.7364	**	0.0709 [0.02, 0.12]
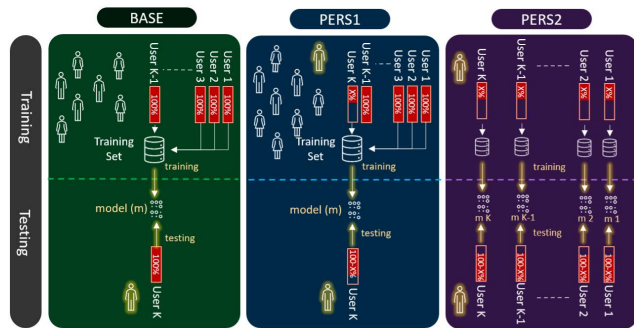
FIGURE 6. Three phases of the study.

V. DETECTING EATING EVENTS AND IMPORTANT FEATURES (RQ2)

A. 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 [65] and python is used to conduct experiments in three phases using different model types: (1) Random Forest (RF) [66], (2) Naive Bayes (NB) [67], (3) Gradient Boosting (GB) [68], and (4) AdaBoost (AB) [69]. These models were chosen by considering the tabular nature of the dataset, interpretability of results (e.g., getting feature importance values), and the small size of the dataset. Further, similar to recent ubicomp work [70], Synthetic Minority Over-sampling Technique (SMOTE) [71] was used to prepare training sets for each inference task. In addition, F1-Score and AUROC score, both with macro averaging is reported. This would give equal emphasize 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 as subject dependent, subject independent, and hybrid in line with prior work [26]) and summarized in Figure 6.

1) SUBJECT INDEPENDENT (BASE)

These models are also called population-level models. The leave-one-out cross-validation strategy that is commonly used in mobile sensing research [9], [35] is used in this phase. The objective is to train on a set of users and test on 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.

2) 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.

3) SUBJECT DEPENDENT (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 [35] 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 and 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 [1]. 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

TABLE 4. Averaged F1-score (F1) and AUROC 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. [12], [10], etc.) indicate the number of principal components used for the inference. For ALL-FS, the notation in square brackets (e.g. [F3], [F7], etc.) indicate the name of the feature group. More details about feature groups, including the list of features in each group are given in Table 7

Feature Group (# of Features)	RF F1, AUROC	NB F1, AUROC	GB F1, AUROC	AB F1, AUROC
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)	73.0, 0.52	72.7, 0.54	72.7, 0.53	72.7, 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. Averaged F1-score (F1) and AUROC 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]) indicate the name of the feature group. More details about the feature group, including the list of features in the group is given in Table 7.

Feature Group (# of Features)	BASE F1, AUROC	PERS1 F1, AUROC	PERS2 F1, AUROC
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**

TABLE 6. Personalized eating event detection accuracy breakdown for random five users in PERS2 with ALL-FS. F1-score (F1) and AUROC are shown. Top performing feature group, and modalities included in the feature group are shown with ✓. 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, AUROC	LOC	ACC	APP	BAT	SCR	TIME	F1-score Bump
1	0.92, 0.82		✓	✓	✓	✓	✓	
2	0.84, 0.59		✓	✓	✓	✓	✓	
3	0.88, 0.91	✓	✓	✓	✓	✓	✓	
4	0.81, 0.48		✓		✓			
5	0.98, 0.99	✓	✓	✓	✓		✓	
								70% 100%

BASE in most cases. In addition, for ALL, ALL-PCA and ALL-FS, F1 scores increased by over 6%. ALL-FS with F1 feature 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 AUROC 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 V-B discusses more details about these feature sets. This suggests that while subject-dependent 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 subject-dependent models for most feature groups.

B. FEATURE IMPORTANCE FOR EATING EVENT DETECTION (RQ2)

In order to study individual differences further, Figure 7 gives BASE feature importance values (Figure 7a) and PERS2 feature importance values for three randomly selected users (Figure 7b, Figure 7c, and Figure 7d) based on Gini importance of RF classifiers. These figures illustrate how 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 time of the day (minutes_elapsed, hours_elapsed). However, for *user C* (0.027, 0.019) the value 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 capture user-specific behaviors regarding eating and non-eating events.
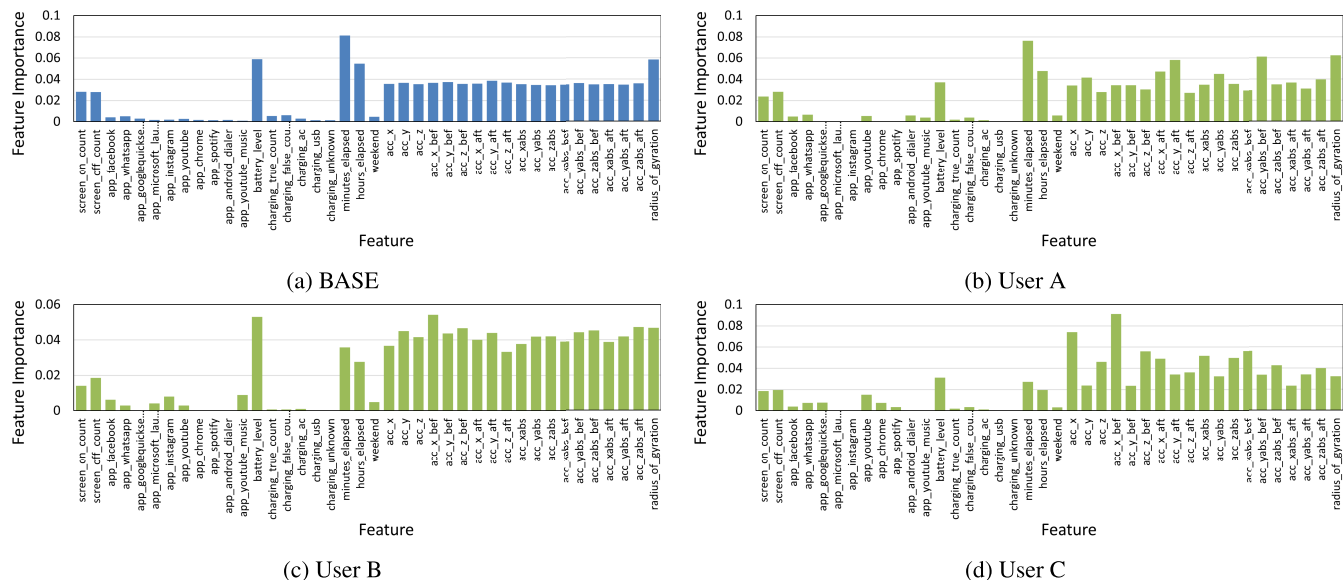
experiments were conducted for feature groups LOC, SCR, TIME, BAT, APP, and ACC separately.

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 [72] 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 [73] 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 ?? and Table ?? summarize the results of experiments. First, in Table ??, 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 AUROC 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 ??. 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 AUROC scores for PERS1 compared to

FIGURE 7. Feature importance values from RFs for (a) BASE and (b)–(d) PERS2 models from three randomly selected users.

C. EFFECT OF PERSONALIZATION ON INDIVIDUALS

Table 6 shows 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 that 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, 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*, 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 F1-score and AUROC decreased when going from BASE to PERS1 to PERS2.

VI. DISCUSSION AND LIMITATIONS

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

A. 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 [9], [20], [33], [34], [40], 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 ubicomp work, Bae *et al.* [33] used thirty-minute, one-hour, and two-hour time windows for drinking event detection. Meegahapola *et al.* [9] used a one-hour time window to detect subjective food consumption level inferences. This research is in line with Bisogni's contextual framework [27], which not only considers eating/drinking events but the whole context around them. Similarly, the models discussed in this paper 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} = 2.15pm at 2.45pm, and run again for T_{anc} = 2.17pm at 2.47pm, the granularity would be higher, compared to running the second inference after another hour for T_{anc} = 3.15pm 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 pm, the inference could be done 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 paper. However, as mentioned earlier, different time windows could be used in future work depending on the dataset and features.

B. 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 [23]. This could have a significant impact on 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 [23].

C. INTERPRETATION OF EATING EVENT DETECTION

There are many ways in which an eating event can be interpreted. In this paper, 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 [9], [20], [27], [33], [40], [40]. 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 our paper 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 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.

D. 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.* [2] 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 [3], [10], [14]. 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 [20], [24]. 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.

E. 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 [24].

F. IMPORTANCE OF DIVERSITY-AWARENESS

Depending on demographic attributes, lifestyle, and culture, eating behaviors can significantly vary [31], [74]. Early work documented differences between men and women related to eating behavior [74]. 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 [75]. 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 [76], [77]. Hence, future work needs to be carried out for different age groups and countries.

G. 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 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 study and is out of the scope of our paper.

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 [1], [20], [34], [35].

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 [1], [20], [33], [34]. Whether other decision-level

TABLE 7. Feature groups used in different inference models.

Table	Feature Group	Features
Table 4	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
	F7	googlequicksearchbox, microsoft_launcher, instagram, youtube, charging_false_count
Table 5	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

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 [78] 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.

VII. CONCLUSION

In this paper, we examined the eating behavior of 58 college students in Mexico using self-reports and passive smartphone sensing data. First, it was shown that 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 AUROC of 0.65 (F1-score of 0.75) using a subject-independent model, which can be further improved up to AUROC of 0.81 (F1-score of 0.85 for subject-dependent 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, subject-dependent models for different users rely on different feature groups. These findings are encouraging towards future mobile food diary apps that are context-aware for both user- and population-level use cases.
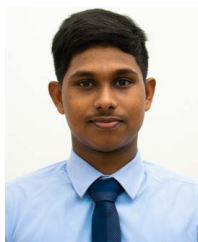
ACKNOWLEDGMENT

REFERENCES

[1] D. Santani, T.-M.-T. Do, F. Labhart, S. Landolt, E. Kuntsche, and D. Gatica-Perez, "DrinkSense: Characterizing youth drinking behavior using smartphones," *IEEE Trans. Mobile Comput.*, vol. 17, no. 10, pp. 2279–2292, Oct. 2018.

[2] T. Rahman, M. Czerwinski, R. Gilad-Bachrach, and P. Johns, "Predicting 'about-to-eat' moments for just-in-time eating intervention," in *Proc. 6th Int. Conf. Digit. Health Conf.*, New York, NY, USA, 2016, pp. 141–150.

[3] 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," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 3, pp. 1–20, Sep. 2017.

[4] 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," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 4, no. 2, pp. 1–25, Jun. 2020.

[5] X. Ye, G. Chen, Y. Gao, H. Wang, and Y. Cao, "Assisting food journaling with automatic eating detection," in *Proc. CHI Conf. Extended Abstr. Hum. Factors Comput. Syst.*, May 2016, pp. 3255–3262.

[6] J. Alexander, "Can journaling help you heal from your eating disorder?" Diary Healer, Australia, 2017. [Online]. Available: https://thediaryhealer.com/2017/11/28/can-journaling-help-you-heal-from-your-eating-disorder/

[7] (2021). *Samsung Health App.* [Online]. Available: https://www.samsung.com/us/support/owners/app/samsung-health

[8] (2021). *Myfitnesspal.* [Online]. Available: https://www.myfitnesspal.com/

[9] L. Meegahapola, S. Ruiz-Correa, V. del Carmen 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 *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 5, no. 1. New York, NY, USA: ACM, 2021, pp. 1–28.

[10] E. Thomaz, I. Essa, and G. D. Abowd, "A practical approach for recognizing eating moments with wrist-mounted inertial sensing," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2015, pp. 1029–1040.

[11] 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 *Proc. 1st Workshop Digit. Biomarkers*, Jun. 2017, pp. 21–26.

[12] K. S. Chun, H. Jeong, R. Adaimi, and E. Thomaz, "Eating episode detection with jawbone-mounted inertial sensing," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 4361–4364.

[13] S. Bi, T. Wang, N. Tobias, J. Nordrum, S. Wang, G. Halvorsen, S. Sen, R. Peterson, K. Odame, K. Caine, R. Halter, J. Sorber, and D. Kotz, "Auracle: Detecting eating episodes with an ear-mounted sensor," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 3, pp. 1–27, Sep. 2018.

[14] K. S. Chun, S. Bhattacharya, and E. Thomaz, "Detecting eating episodes by tracking jawbone movements with a non-contact wearable sensor," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 1–21, Mar. 2018.

[15] 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," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 4, no. 2, pp. 1–26, Jun. 2020.

[16] T. Rahman, M. Czerwinski, R. Gilad-Bachrach, and P. Johns, "Predicting 'about-to-eat' moments for just-in-time eating intervention," in *Proc. 6th Int. Conf. Digit. Health Conf.*, 2016, pp. 141–150.

[17] C. Merck, C. Maher, M. Mirtchouk, M. Zheng, Y. Huang, and S. Kleinberg, "Multimodality sensing for eating recognition," in *Proc. 10th EAI Int. Conf. Pervasive Comput. Technol. Healthcare*, 2016, pp. 130–137.

[18] Mobius. (2019). *11 Surprising Mobile Health Statistics.* [Online]. Available: https://www.statista.com/chart/13226/where-people-spend-the-most-time-eating-drinking/

[19] (2021). *Mobile Fact Sheet.* [Online]. Available: https://www.pewinternet.org/fact-sheet/mobile/

[20] J.-I. Biel, N. Martin, D. Labbe, and D. Gatica-Perez, "Bites'n'bits: Inferring eating behavior from contextual mobile data," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 4, pp. 125:1–125:33, Jan. 2018.

[21] L. Meegahapola, S. Ruiz-Correa, and D. Gatica-Perez, "Alone or with others? Understanding eating episodes of college students with mobile sensing," in *Proc. 19th Int. Conf. Mobile Ubiquitous Multimedia*, New York, NY, USA, Nov. 2020, pp. 162–166.

[22] L. Meegahapola, S. Ruiz-Correa, and D. Gatica-Perez, "Protecting mobile food diaries from getting too personal," in *Proc. 19th Int. Conf. Mobile Ubiquitous Multimedia*, New York, NY, USA, Nov. 2020, pp. 212–222.

[23] 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," *npj Digit. Med.*, vol. 3, no. 1, pp. 1–14, Dec. 2020.

[24] L. Meegahapola and D. Gatica-Perez, "Smartphone sensing for the well-being of young adults: A review," *IEEE Access*, vol. 9, pp. 3374–3399, 2021.

[25] M. Bin Morshed, S. S. Kulkarni, R. Li, K. Saha, L. G. Roper, L. Nachman, H. Lu, L. Mirabella, S. Srivastava, M. De Choudhury, K. de Barbaro, T. Ploetz, and G. D. Abowd, "A real-time eating detection system for capturing eating moments and triggering ecological momentary assessments to obtain further context: System development and validation study," *JMIR mHealth uHealth*, vol. 8, no. 12, Dec. 2020, Art. no. e20625.

[26] A. Ferrari, D. Micucci, M. Mobilio, and P. Napoletano, "On the personalization of classification models for human activity recognition," *IEEE Access*, vol. 8, pp. 32066–32079, 2020.

[27] 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," *Appetite*, vol. 48, no. 2, pp. 218–231, Mar. 2007.

[28] M. M. Jastran, C. A. Bisogni, J. Sobal, C. Blake, and C. M. Devine, "Eating routines. Embedded, value based, modifiable, and reflective," *Appetite*, vol. 52, no. 1, pp. 127–136, Feb. 2009.

[29] M. Singh, "Mood, food, and obesity," *Frontiers Psychol.*, vol. 5, p. 925, Sep. 2014.

[30] C. Herman, J. Polivy, and T. Leone, "6—The psychology of overeating," in *Food, Diet and Obesity* (Woodhead Publishing Series in Food Science, Technology and Nutrition), D. J. Mela, Ed. Sawston, U.K.: Woodhead Publishing, 2005, pp. 115–136.

[31] G. Ma, "Food, eating behavior, and culture in Chinese society," *J. Ethnic Foods*, vol. 2, no. 4, pp. 195–199, Dec. 2015.

[32] P. G. Turner and C. E. Lefevre, "Instagram use is linked to increased symptoms of orthorexia nervosa," *Eating Weight Disorders-Stud. Anorexia, Bulimia Obesity*, vol. 22, no. 2, pp. 277–284, Jun. 2017.

[33] 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," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 2, pp. 1–36, Jun. 2017.

[34] 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 *Proc. 26th Int. Conf. World Wide Web*, Apr. 2017, pp. 103–112.

[35] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong, "MoodScope: Building a mood sensor from smartphone usage patterns," in *Proc. 11th Annu. Int. Conf. Mobile Syst., Appl., Services (MobiSys)*, 2013, pp. 389–402.

[36] M. Rabbi, M. H. Aung, M. Zhang, and T. Choudhury, "MyBehavior: Automatic personalized health feedback from user behaviors and preferences using smartphones," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2015, pp. 707–718.

[37] E. M. Berke, T. Choudhury, S. Ali, and M. Rabbi, "Objective measurement of sociability and activity: Mobile sensing in the community," *Ann. Family Med.*, vol. 9, no. 4, pp. 344–350, Jul. 2011.

[38] D. Gatica-Perez, J.-I. Biel, D. Labbe, and N. Martin, "Discovering eating routines in context with a smartphone app," in *Proc. Int. Joint Conf. Pervasive Ubiquitous Comput., ACM Int. Symp. Wearable Comput.*, Sep. 2019, pp. 422–429.

[39] 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," *PLoS ONE*, vol. 16, no. 4, Apr. 2021, Art. no. e0250443.

[40] L. Meegahapola, F. Labhart, T.-T. Phan, and D. Gatica-Perez, "Examining the social context of alcohol drinking in young adults with smartphone sensing," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 5, no. 3, pp. 1–26, Sep. 2021.

[41] 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 *Proc. Designing Interact. Syst. Conf.*, Jun. 2021, pp. 606–618.

[42] K. Okamoto and K. Yanai, "An automatic calorie estimation system of food images on a smartphone," in *Proc. 2nd Int. Workshop Multimedia Assist. Dietary Manage.*, Oct. 2016, pp. 63–70.

[43] T. Denning, A. Andrew, R. Chaudhri, C. Hartung, J. Lester, G. Borriello, and G. Duncan, "BALANCE: Towards a usable pervasive wellness application with accurate activity inference," in *Proc. HotMobile*, Santa Cruz, CA, USA. New York, NY, USA: Association for Computing Machinery, 2009, Art. no. 5, doi: 10.1145/1514411.1514416.

[44] A. Kadomura, C.-Y. Li, K. Tsukada, H.-H. Chu, and I. Siio, "Persuasive technology to improve eating behavior using a sensor-embedded fork," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Sep. 2014, pp. 319–329.

[45] Y. Gao, N. Zhang, H. Wang, X. Ding, X. Ye, G. Chen, and Y. Cao, "IHear food: Eating detection using commodity Bluetooth headsets," in *Proc. IEEE 1st Int. Conf. Connected Health, Appl., Syst. Eng. Technol. (CHASE)*, Jun. 2016, pp. 163–172.

[46] 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," 2020, *arXiv:2007.05831*.

[47] Y. Dong, J. Scisco, M. Wilson, E. Muth, and A. Hoover, "Detecting periods of eating during free-living by tracking wrist motion," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 4, pp. 1253–1260, Jul. 2013.

[48] A. Madan, S. T. Moturu, D. Lazer, and A. S. Pentland, "Social sensing: Obesity, unhealthy eating and exercise in face-to-face networks," in *Wireless Health 2010*. New York, NY, USA: ACM, 2010, pp. 104–110.

[49] 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," *PLoS ONE*, vol. 11, no. 4, pp. 1–16, 2016.

[50] M. Zeni, I. Zaihrayeu, and F. Giunchiglia, "Multi-device activity logging," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., Adjunct Publication*, New York, NY, USA, Sep. 2014, pp. 299–302.

[51] Y. Yue, T. Lan, A. G. O. Yeh, and Q.-Q. Li, "Zooming into individuals to understand the collective: A review of trajectory-based travel behaviour studies," *Travel Behav. Soc.*, vol. 1, no. 2, pp. 69–78, May 2014.

[52] G. Barlacchi, C. Perentis, A. Mehrotra, M. Musolesi, and B. Lepri, "Are you getting sick? Predicting influenza-like symptoms using human mobility behaviors," *EPJ Data Sci.*, vol. 6, no. 1, p. 27, 2017.

[53] 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 *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, New York, NY, USA, Sep. 2016, pp. 178–189.

[54] 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 *Proc. 17th Annu. Int. Conf. Mobile Syst., Appl., Services*, 2019, pp. 41–53.

[55] 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," *Public Health Nutrition*, vol. 14, no. 11, pp. 1998–2005, Nov. 2011.

[56] G. S. Castell, L. Serra-Majem, and L. Ribas-Barba, "What and how much do we eat? 24-hour dietary recall method," *Nutricion hospitalaria*, vol. 31, no. 3, pp. 46–48, 2015.

[57] J. Kerr, L. Frank, J. F. Sallis, B. Saelens, K. Glanz, and J. Chapman, "Predictors of trips to food destinations," *Int. J. Behav. Nutrition Phys. Activity*, vol. 9, no. 1, pp. 1–10, 2012.

[58] T.-M.-T. Do and D. Gatica-Perez, "By their apps you shall understand them: Mining large-scale patterns of mobile phone usage," in *Proc. 9th Int. Conf. Mobile Ubiquitous Multimedia*, 2010, pp. 1–10.

[59] 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 *Proc. 13th Int. Conf. Human Comput. Interact. Mobile Devices Services*, 2011, pp. 47–56.

[60] T. K. Kim, "T test as a parametric statistic," *Korean J. Anesthesiology*, vol. 68, no. 6, p. 540, 2015.

[61] S. Greenland, S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman, and D. G. Altman, "Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations," *Eur. J. Epidemiol.*, vol. 31, no. 4, pp. 337–350, Apr. 2016.

[62] D. K. Lee, "Alternatives to p value: Confidence interval and effect size," *Korean J. Anesthesiol.*, vol. 69, no. 6, p. 555, 2016.

[63] M. E. Rice and G. T. Harris, "Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r," *Law Hum. Behav.*, vol. 29, no. 5, pp. 615–620, 2005.

[64] D. Lakens, "Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs," *Frontiers Psychol.*, vol. 4, p. 863, Nov. 2013.

[65] 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," *J. Mach. Learn. Res.*, vol. 12, no. 10, pp. 2825–2830, Jul. 2017.

[66] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," in *Ensemble Machine Learning: Methods and Applications*, C. Zhang and Y. Ma, Eds. Boston, MA, USA: Springer, 2012, pp. 157–175, doi: 10.1007/978-1-4419-9326-7_5.

[67] I. Rish *et al.*, "An empirical study of the naive Bayes classifier," in *Proc. IJCAI Workshop Empirical Methods Artif. Intell.*, 2001, vol. 3, no. 22, pp. 41–46.

[68] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers Neurorobot.*, vol. 7, p. 21, Dec. 2013.

[69] R. E. Schapire, "Explaining adaboost," in *Empirical Inference*. Cham, Switzerland: Springer, 2013, pp. 37–52.

[70] 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," *Addictive Behaviors*, vol. 83, pp. 42–47, Aug. 2018.

[71] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[72] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, 1987.

[73] A. Marcano-Cedeño, J. Quintanilla-Domínguez, M. G. Cortina-Januchs, and D. Andina, "Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network," in *Proc. 36th Annu. Conf. IEEE Ind. Electron. Soc.*, Nov. 2010, pp. 2845–2850.

[74] B. J. Rolls, I. C. Fedoroff, and J. F. Guthrie, "Gender differences in eating behavior and body weight regulation," *Health Psychol.*, vol. 10, no. 2, p. 133, 1991.

[75] N. McCarthy. (2020). *Where People Spend the Most Time Eating & Drinking*. [Online]. Available: https://www.statista.com/chart/13226/where-people-spend-the-most-time-eating-drinking/

[76] 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 *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 3. New York, NY, USA: Association for Computing Machinery, Sep. 2019, doi: 10.1145/3351246.

[77] 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 *Proc. AAAI/ACM Conf. AI, Ethics, Soc.* New York, NY, USA: Association for Computing Machinery, 2021, pp. 905–915, doi: 10.1145/3461702.3462595.

[78] 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," *BMC Med. Res. Methodology*, vol. 19, no. 1, Dec. 2019.

LAKMAL MEEGAHAPOLA received the B.Sc. degree in computer science and engineering from the University of Moratuwa, Sri Lanka, in 2018. He is currently pursuing the Ph.D. degree with the École Polytechnique fédérale de Lausanne (EPFL), Switzerland. He is a Research Assistant with the Social Computing Group, Idiap Research Institute, Switzerland. He is interested and experienced in using machine learning and data mining techniques for research in the intersection of mobile health sensing, context-awareness, and diversity-aware machine learning.

SALVADOR RUIZ-CORREA received the Ph.D. degree in electrical engineering from the University of Washington. He is currently an Adjunct Professor and a Researcher with the Instituto Potosino de Investigación Científica y Tecnológica, where he leads the Youth Innovation Laboratory (You-i Laboratory). He co-directs the Center for Mobile Life (Ce Mobili), a Research Initiative in Mexico. His research interests include machine learning and computer vision applications to social computing in development contexts, data for social good, and citizen innovation.

WAGEESHA BANGAMUARACHCHI was born in Colombo, Sri Lanka, in 1997. She is currently pursuing the B.Sc. (Engineering) degree in computer science and engineering with the University of Moratuwa. She has worked as a Software Engineering Intern with the Sysco Laboratories Technologies, Sri Lanka. Her current research interests include ubiquitous computing, mobile sensing, and machine learning.

INDIKA PERERA (Senior Member, IEEE) received the B.Sc. (Engineering) (Hons.) and M.Sc. degrees from the University of Moratuwa, Sri Lanka, the M.B.S. and P.G.D.B.M. degrees from the University of Colombo, Sri Lanka, and the Ph.D. degree from the University of St Andrews, U.K. He is currently a Professor and the Head of the Department of Computer Science and Engineering, University of Moratuwa. His research interests include software architecture, software engineering, technology enhanced learning, and user experience and application development for bio-health research.

ANJU CHAMANTHA is currently pursuing the B.Sc. (Engineering) degree with the Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka. He has completed his internship at WSO2 Lanka Ltd., a reputed open source software development company. His research interests include machine learning, mobile sensing, and human–computer interaction.

DANIEL GATICA-PEREZ (Member, IEEE) directs the Social Computing Group, Idiap Research Institute, and also a Professor at EPFL, Switzerland. His research interests include social computing, ubiquitous computing, and crowdsourcing for social good.

• • •