Understanding the Social Context of Eating with Multimodal Smartphone Sensing: The Role of Country Diversity

Nathan Kammoun
Idiap Research Institute & EPFL
Switzerland

Lakmal Meegahapola
Idiap Research Institute & EPFL
Switzerland

Daniel Gatica-Perez
Idiap Research Institute & EPFL
Switzerland

ABSTRACT

Understanding the social context of eating is crucial for promoting healthy eating behaviors. Multimodal smartphone sensor data could provide valuable insights into eating behavior, particularly in mobile food diaries and mobile health apps. However, research on the social context of eating with smartphone sensor data is limited, despite extensive studies in nutrition and behavioral science. Moreover, the impact of country differences on the social context of eating, as measured by multimodal phone sensor data and self-reports, remains under-explored. To address this research gap, our study focuses on a dataset of approximately 24K self-reports on eating events provided by 678 college students in eight countries to investigate the country diversity that emerges from smartphone sensors during eating events for different social contexts (alone or with others). Our analysis revealed that while some smartphone usage features during eating events were similar across countries, others exhibited unique trends in each country. We further studied how user and country-specific factors impact social context inference by developing machine learning models with population-level (non-personalized) and hybrid (partially personalized) experimental setups. We showed that models based on the hybrid approach achieve AUC scores up to 0.75 with XGBoost models. These findings emphasize the importance of considering country differences in building and deploying machine learning models to minimize biases and improve generalization across different populations.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in ubiquitous and mobile computing**.

KEYWORDS

smartphone sensing, passive sensing, multimodal sensor data, eating behavior, social context, mobile food diary, mobile health

1 INTRODUCTION

The eating behavior of college students has significant implications for their overall well-being. As a result, behavioral and nutrition scientists have conducted extensive research on the interplay between food consumption and health [27, 34]. To this end, prior research has emphasized that eating is a holistic event with interconnected dimensions such as the type of food, context, and time, among other aspects [19]. The social context of eating is a factor that shapes people's food consumption, along with other contextual cues such as mood and location [12, 26, 32]. While the social context is multidimensional, prior research has conceptualized eating alone or with others as two basic types of social contexts. Therefore, several studies have investigated the impact of eating alone versus eating with others on food consumption [18, 42]. Moreover, the social context of eating is associated with key aspects of eating episodes, such as mood, physical condition, time, and location.

"Food and Nutrition" is among the most common categories of mobile health applications, but most dietary pattern assessment methodologies rely on Ecological Momentary Assessments (EMA) and survey questionnaires [8]. However, measuring eating behavior with such techniques is hardly ever done consistently, as self-reports can be burdensome for users [20]. Moreover, typical dietary assessment surveys do not always capture the social context and other contextual cues associated with an individual's food consumption [36, 38]. As an alternative, longitudinal behavior modeling with multimodal sensing can address the challenges of dietary assessment methods by capturing the interplay between food consumption and contextual characteristics while replacing burdensome questionnaires [6, 28]. Inference models relying on features derived using phone sensors (activity types, step count, location, app usage, typing and touch events, screen on and off episodes, etc.) can be used to understand key contextual aspects affecting food consumption [25, 26, 30]. However, the social context of eating, especially in the case of multiple countries, has not systematically been studied with mobile technologies.

Diversifying the data for training models by considering more countries is fundamental to improving the performance of the models and their ability to generalize to a broader range of populations. Generalizing research results by building diversity-aware models can help overcome issues related to biases and reproducibility and aim toward better generalizability of mobile sensing research across different contexts [2, 3, 31, 33]. However, even though computer vision and natural language processing research have studied cross-country or cross-dataset generalization of models, mobile sensing studies focused on longitudinal behavior modeling have started doing so only recently [2, 23, 41]. Considering these aspects, this paper aims to answer the following research questions:

RQ1: What behavioral and contextual characteristics can be extracted from analyzing the social context surrounding eating events among college students in eight countries, using multimodal smartphone sensing data and self-reports?

RQ2: Can multimodal phone sensor data be utilized to infer the social context of eating events, and if so, how does the geographical diversity affect the inference performance?

In answering these research questions, we provide the following contributions.

Contribution 1: We conducted a study using a large dataset (24K eating event self-reports and corresponding multimodal sensor data from ten modalities) collected from college students in eight countries (UK, Italy, Denmark, Paraguay, Mongolia, Mexico, India, and China) with the goal of identifying behavioral differences among countries. We first conducted a descriptive analysis that demonstrated the natural time-dependency of eating patterns and then examined the hourly distribution of eating-alone versus eating-with-others events across the eight countries to provide initial evidence of cross-country diversity. We found that the hourly pattern of the eating social context varies across countries as both the time of eating and the social context are tied to social practices. Furthermore, we statistically analyzed the interplay between smartphone features and the social context during eating events to demonstrate possible differences across countries. We found that while some features are used consistently across all settings, others are only relevant in some countries.

Contribution 2: Using smartphone features and binary target data on the social context of eating (alone or with others), we developed inference models with various methods. We trained a number of model configurations: *i)* separate models for each country; *ii)* aggregate models based on geographical proximity, such as Asia, Latin America, and Europe, and *iii)* a model that pools data from all countries. For each case, we evaluated population-level (non-personalized) and hybrid (partially personalized) models using AUC as the performance metric. The experimental results showed that the AUC of population-level models was 0.58, while the hybrid models achieved AUC scores of up to 0.75. We also analyzed the feature importance extracted from the hybrid models to obtain further insight into the cross-country differences of inference models. By displaying all feature importances based on their common use across all countries, we show that the same subset of features is used consistently across all countries with varying degrees of relevance. These findings provide a comprehensive understanding of country diversity by revealing more complex relationships in the data that could not be identified through statistical analysis alone.

2 BACKGROUND AND RELATED WORK

2.1 Eating as a Holistic Event. Studies in nutrition and behavioral sciences have shown that eating behavior is influenced by various factors, including situational and behavioral ones. Bisogni et al. [19] proposed a contextual framework for eating and drinking events, which includes eight interconnected dimensions such as "food and drink, time, location, activities, social context, mental processes, physical condition, and recurrence". Following this framework, eating is considered a complex event that is influenced by a range of internal and external factors, including the social

context. Researchers have highlighted the social context of meals as a fundamental aspect of food consumption [12, 32]. The social context itself comprises several dimensions, such as the type of relationships or the number of participants. Specifically, whether an individual is alone or with others has often been used in behavioral studies as two key categories of social context [7, 18]. Prior research in nutrition and behavioral sciences has demonstrated that eating in highly social contexts is likely to influence the amount of food consumed. The presence of one or more people during meals can also lead to impression management or social facilitation, resulting in under-eating and overeating, respectively [12, 17]. Therefore, the longitudinal understanding of social context during eating events could enable timely feedback and interventions for individuals who use mobile health applications.

2.2 Multimodal Mobile Sensing for Eating Behavior Modeling. The advancement of mobile sensing technologies has allowed researchers to monitor eating behaviors in real-time using wearable sensors to detect eating events or characterize behavioral and contextual attributes around eating episodes. Such developments in ubiquitous computing have provided researchers with additional insight into eating behaviors [4, 5, 15, 24, 29, 43]. Prior research has emphasized the importance of studying the context surrounding eating events to better understand an individual's eating social context [13, 30]. However, most of these studies have primarily used self-reports instead of sensor data. Moreover, the datasets used in these studies were not collected from multiple countries, limiting the ability to investigate country differences, as seen through both self-reports and sensor data. In studies that use sensing for studying eating behavior, Biel et al. [6] demonstrated that sensor data from phones could be used to distinguish between meal and snack events. Meegahapola et al. [28] showed that overeating events could be inferred with multimodal sensor data with personalized accuracy of up to 87%. Similarly, another study [25] used sensor data to infer the social context during drinking events with accuracies in the range of 0.75-0.86. However, despite being briefly discussed in one previous study [26], no other study has extensively examined the social context of eating with sensor data collected from multiple countries using the same study protocol, thus allowing for intuitive comparisons.

2.3 Diversity Awareness. Previous studies in the field of machine learning have highlighted the need to ensure model accuracy and fairness across diverse populations [1]. The concept of data diversity has been applied to various domains, including computer vision and natural language processing [35]. However, applying data diversification to mobile sensing is challenging due to the lack of large-scale multimodal datasets collected using a consistent protocol across multiple countries. Therefore, it is crucial to develop diversity-aware approaches to machine learning-based modeling of sensor data to examine model generalization across countries. Mobile sensing data can exhibit diversity due to various factors such as user behavior, differences in devices, and mobile networks in the country of interest. Smartphone usage can also vary significantly between countries, even within a given population, such as students. While being used consistently in mobile sensing studies without considering generalization to other countries, passive sensing features from WiFi or location need to be considered with care in deployment settings [2]. Recently, several research groups have explored this direction of research by collecting multimodal sensing

Table 1: Mobile Sensing Data Collection Summary.

Country (N)	μ Age (σ)	% Women	# Self Reports	# Eating Reports	% Eating Alone
China (41)	26.2 (4.2)	51	30,406	1548	45
Denmark (24)	30.2 (6.3)	58	12,354	613	41
India (39)	23.7 (3.2)	53	4478	340	42
Italy (240)	24.1 (3.3)	58	176,135	12,697	28
Mexico (20)	24.1 (5.3)	55	11,908	773	40
Mongolia (214)	22.0 (3.1)	65	121,809	5,674	15
Paraguay (28)	25.3 (5.1)	60	11,790	729	29
UK (72)	26.6 (5.0)	66	31,989	1,599	51
Total (678)	24.2 (4.2)	58	400,829	23,973	36

datasets for multiple countries for longitudinal behavior modeling [21, 23]. These studies have investigated the effect of distributional shifts across countries. However, studies focusing on social context inference during specific events such as eating and drinking have only used data from one or two countries [22, 25]. Therefore, no comparable multimodal smartphone sensing datasets exist for similar analysis, as previous work relied on data collected with contrasting protocols and features, hence restricting comparisons across countries [26]. Therefore, the impact of country diversity on social context inference during eating events, or eating events in general, has not been thoroughly discussed and requires further investigation.

3 DATA, FEATURES AND TARGET CLASSES

3.1 Dataset Information

To examine the research questions, we rely on a dataset used in our previous work [14, 23]. This dataset was collected during an in-the-wild study that was conducted over a four-week period in November 2020, aimed to explore the everyday behavior of college students through multimodal mobile phone sensors. Deployed in eight countries across various regions of the world with the same protocol, the study aimed to investigate the effect of country diversity on key aspects of mobile sensing.

The first data collection phase gathered demographic general data about each participant. The second phase involved collecting data through a smartphone application. Participants were asked to fill out time diaries throughout the day, indicating what they were doing (from 34 activities), location (from 26 semantic locations), social context (from 8 configurations), and mood (valence with a five-point scale).

In this study, we focus on eating events from the larger dataset, which contains reports for a wide range of activities. Therefore, we only considered the samples for which the reported activity is eating, which drastically reduced the number of data points, with eating events representing 5% to 7% of the data, depending on the country. Table 1 presents statistics about the participants, including the total number of reports and the number of eating reports, the sample sizes are fairly uneven between countries ranging from 4478 to 176,135 reports. The participants were asked to choose from a variety of categories when filling out self-reports; possible answers for social context include: alone, with relatives, with friends, with classmates, and so on. These answers were translated into a binary target Alone or With Others in line with previous studies [18, 25, 42]. Table 1 also provides the class distributions for each country.
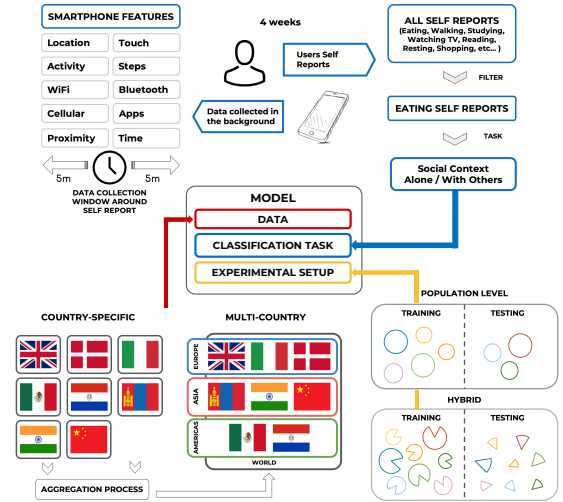
Figure 1: High-Level Overview of the Study

3.2 Smartphone Sensing Features

The app continuously collected data from more than thirty smartphone sensors, divided into two categories: continuous sensing modalities (activity type, step count, location, phone signal, WiFi, Bluetooth, battery, and proximity) and interaction sensing modalities (notifications clicking, application usage, screen episodes, and user presence). A time-window-based approach was used to process the feature data. All sensor measurements were aggregated with self-reports to create features characterizing the time window during which the report occurred. The length of the window was a ten-minute window around eating event self-report, during which statistics about each smartphone modality were computed [23], obtaining over 100 features (more details in the appendix). Two additional features were computed from timestamps: a binary feature stating whether the current day is a weekday or weekend and the hour of report collection. Further, in the context of smartphone sensing, missing data can occur for multiple reasons: device in low-consumption mode, sensor failure, user privacy settings, airplane mode, or simply the type of phone used, not including necessary hardware. To deal with this issue, the feature modalities for which more than 70% of the data was missing were dropped, namely: Bluetooth low energy, Bluetooth normal, Cellular GSM, and Cellular WCDMA. While most of the tools used in this study were robust against missing values, an imputed version of the dataset was still needed in some cases. A k-Nearest Neighbor (k-NN) imputation [39] was used to impute missing values from the remaining features in training sets, whenever needed.

4 BEHAVIOR & CONTEXT AROUND EATING SOCIAL CONTEXT (RQ1)

4.1 Methodology

To answer the first research question, a descriptive and statistical analysis was conducted in three separate components. First, we analyzed the hourly distribution of eating events with respect to the social context using self-reports. Next, we performed a statistical

analysis to explore the relationship between smartphone sensor features and the two social contexts of eating. Lastly, we created a two-dimensional embedding for the extracted sensor features to visually showcase the cross-country differences to better understand the relationship between sensor data and eating social contexts.

4.1.1 Descriptive Analysis. As eating events are inherently time-dependent, we analyzed the hourly distribution of eating reports for all countries. We provided two separate density plots to investigate how eating events were distributed during the day for different social contexts (alone vs. with others) in Figure 2. We then examined the hourly distribution of eating events to identify the relationship between time and social context across countries.

4.1.2 Statistical Analysis. To understand the cross-country differences, the class distributions of each eating context outcome were statistically analyzed to uncover features that define eating social context in each country. Hence, t-tests were performed to understand class distributions with mobile sensing features. The top five features with the highest t-statistic [37] and p-values lesser than 0.05 were reported in Table 2 alongside the effect size (Cohen's-d) [11] [1]. In order to gain insight into feature distributions across countries, in Figure 3, we plotted the effect sizes with features grouped by sensor types such as location, WiFi, cellular data, notifications, proximity, applications, activity, steps, and touch. The purpose of this plot is to display which feature groups are the most important and to provide a means of effectively analyzing differences between countries.

4.1.3 Visualizing Diversity. We also present a method for visualizing individual self-reports in the high-dimensional feature space. The t-Stochastic Neighbor Embedding (t-SNE) [40] algorithm was used for this purpose. t-SNE is a dimensionality reduction technique that maps high-dimensional features to a two-dimensional space for visualization purposes. First, t-SNE constructs a probability distribution of high-dimensional features by assigning higher probabilities to similar data points. Next, it defines a similar probability distribution over points in the low-dimensional space and optimizes the embedding by minimizing the Kullback-Leibler (KL) divergence between both distributions. In summary, t-SNE maps each eating event in the high-dimensional feature space to a two-dimensional point such that similar events are mapped to nearby points, and dissimilar events are mapped to distant points with high probability. This method allows us to gain insight into individual self-reports and their relationships in the feature space. In order to treat all dimensions with equal importance when performing the dimensionality reduction, we re-scaled features with the Z-score. *Country Specific Embedding.* A separate t-SNE analysis was conducted on each country's dataset to obtain a country-specific feature space mapping. This analysis aimed to show how user-specific eating events are sensed by plotting each user's eating events in a different color. The first step was to standardize the data, as t-SNE relies on a measure of similarity between data points. The embedding was then fitted using the hyper-parameters: $perplexity = 30.0$, $early\_exaggeration = 12.0$, and $learning\_rate = 200.0$. However, for most countries, the number of users was too high to plot all the

(a) Alone   (b) With Others
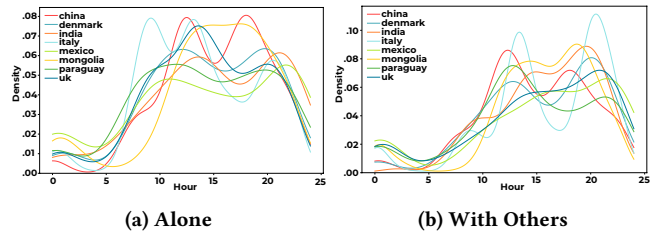
Figure 2: Hourly Distribution of Eating Events

sensed eating events for each user, which would have resulted in a more difficult reading. So, for each country, only twenty users were randomly selected for plotting after fitting the embedding on the entire country's dataset. Results are presented in Figure 4.
Multi-Country Embedding. A single t-SNE was also applied on the whole dataset, comprising an aggregation of all countries, to investigate country-specific behaviors by assigning a color to each country and plotting the reports for each country in separate plots for clarity. To avoid allowing the country with more data to imbalance the stochastic embedding, a balanced aggregation of countries was created by randomly selecting a subset of reports in each country. Then, data were standardized, and the embedding was fitted using the same parameters as before. While this approach employs individual points for each report to reveal clusters, overlapping points can make it challenging to observe how eating events are distributed in the 2D plots. To compensate for this, an additional density plot was presented where the 2D space was partitioned based on polar coordinates, and the density of events in each partition was calculated, as given in Figure 5 and Figure 6.

4.2 Results

4.2.1 Descriptive Analysis. Figure 2a and Figure 2b display the hourly distributions of eating alone and with others events. As these outcomes are measured during eating events, they naturally follow the trend of the general eating schedule. It is worth noting that as the classes are imbalanced, the dominant class pattern is always close to the general eating schedule. Therefore, it is more interesting to explore the hourly distribution of the minority class. These plots enable us to better understand the eating context's time dependency across countries. In Figure 2a, we can observe that all countries show a higher distribution of 'alone' reports during the morning in comparison with the 'with others' distribution, suggesting that breakfast might have been taken alone. This finding is consistent with previous research in nutrition sciences, showing that eating alone most commonly takes place in the morning and midday [42]. Please note that the data was collected during the covid19 pandemic year of 2020, so this influences typical socialization patterns. Moreover, the peak of 'alone' reports during the evening for Italy significantly decreases, indicating that dinner is more likely to happen with other people. So, the eating context is time-dependent, and while this holds for all countries with varying degrees, the patterns are not the same, which provides additional evidence of the cross-country diversity regarding eating behaviors.

4.2.2 Statistical Analysis with Hypothesis Testing. The results of the hypothesis testing are presented in Table 2. While some features

---

[1]As a rule of thumb, a value below 0.2: small effect size, a value of 0.5: medium effect size, and a value of 0.8 or higher: large effect size.

Table 2: t-statistic (TS) and Cohen's-d (CD). TS with p-values above 0.05 and CD with a 95% confidence interval overlapping with 0 are marked with a star (*). p-values are reported after the Bonferroni correction.

Country	Feature	TS	CD	Country	Feature	TS	CD
China	proximity mean	3.15	0.19	Mexico	app tools	3.09	0.23
	app tools	2.48	0.13		location speed mean	2.60	0.17
	app health & fitness	1.97	0.10*		location radius of gyration	2.58	0.17
	location_speed_max	1.80*	0.09*		weekday	2.56	0.18
	app_not_found	1.68*	0.13*		hour	2.42	0.18
Denmark	app tools	5.20	0.46	Mongolia	screen time total	6.25	0.24
	weekday	4.77	0.38		screen max episodes	5.77	0.22
	cellular lte std	3.52	0.35		screen # episodes	5.69	0.20
	app communication	3.16	0.28		touch # events	5.15	0.17
	activity tilting	3.14	0.34		screen time / episode	4.83	0.19
India	wifi max rssi	9.01	1.36	Paraguay	weekday	3.63	0.28
	wifi mean rssi	8.56	1.29		app tools	2.64	0.22
	wifi std rssi	7.27	1.13		location_altitude_min	2.32	0.21
	location altitude min	7.24	1.01		app board	2.15	0.13*
	proximity std	3.24	0.5		app strategy	1.90*	0.12*
Italy	hour	19.02	0.37	UK	app tools	6.06	0.32
	weekday	9.26	0.18		hour	3.40	0.17
	app tools	8.54	0.17		activity invehicle	3.39	0.22
	wifi std rssi	5.94	0.14		activity onbicycle	3.36	0.22
	wifi max rssi	5.77	0.14		activity running	3.33	0.21

Figure 3: Effect Sizes for Social Context of Eating

Figure 4: Country-Specific t-SNE: each color stands for a user.

4.2.3 Statistical Analysis on Effect Sizes and Feature Groups. Figure 3 presents the plot for effect sizes. It is observed that most features have an effect size below 0.25, but some features have a large effect size, and their Cohen's-d 95% interval does not overlap with zero, indicating higher reliability in results. The results show that India has above-average effect sizes on most features, except for activity-type features. Larger effect sizes were found for activity features in China and UK, while Paraguay has low Cohen's-d for cellular and proximity features. Mongolia also had smaller effect sizes for WiFi and notifications compared to other countries. These differences may be attributed to cultural aspects or mobile networks and hardware. In conclusion, analyzing effect sizes among feature groups can help identify differences between countries and discover which features could be useful for inference.

4.2.4 Visualizing Diversity with Country Specific Embeddings. Figure 4 shows the outcomes of the mappings, indicating that sensed eating events are highly personalized for each user. The presence of clusters of points of the same color across all countries indicates that user behavior is distinct and that phone features could be highly associated with the user's habits. However, some countries, such as China, exhibit less distinctive clusters, suggesting that sensor features during eating events are more similar across users in that country. The presence of clusters of the same color at different locations on the plot could indicate that the same user's behavior varies across different contexts. Despite the loss of information resulting from mapping a feature space from 96 dimensions to 2 dimensions, these plots demonstrate the variability across users in the same country and how that variability applies in different parts of the world.

4.2.5 Visualizing Diversity with Multi-Country Embedding. Figure 5 demonstrates the clustering of sensed eating events from each country in different areas of the plot, indicating the high degree of country-specificity. This highlights the necessity of developing diversity-aware models that take into account the cross-cultural variability in smartphone usage and eating behavior. The proximity of the points in the low-dimensional space represents the similarity between the sensed eating events. Although comparing distributions in two dimensions has its limitations, some general observations can still be made. For instance, in Figure 6, the clustering of points for Denmark, the UK, and Italy suggests that these European

are consistently used across countries, the vast majority of them differ, highlighting the diversity in smartphone usage. Time-related features rank among the most relevant for all countries except for China, Mexico, and Paraguay, confirming the hourly distributions' findings. Notably, the weekday feature is visible among features for Denmark, Italy, Mexico, and Paraguay, indicating a different social context depending on whether it is a weekend or a weekday. WiFi features play a significant role in India, which could be because, in the university where the study was conducted, students tend to gather in places with WiFi network-dense areas to eat. The most relevant features seem to point towards a measure of phone usage. For example, in Mongolia, screen features indicate an increase in smartphone usage, which could mean that the user is alone when eating. The presence of proximity features in China is also noteworthy. The proximity sensor is a good indicator of whether a phone is in the user's pocket/bag or hand, which could have a relation with being with people. It is important to note that the results are not all easily interpretable, and the table only shows a selection of features. However, the variety of features across countries provides a good display of cross-country diversity in the sensor data, for the two eating social contexts.
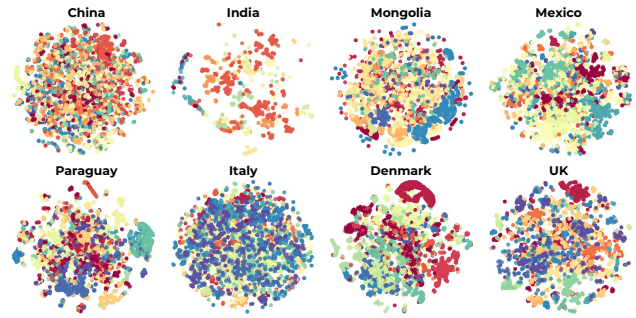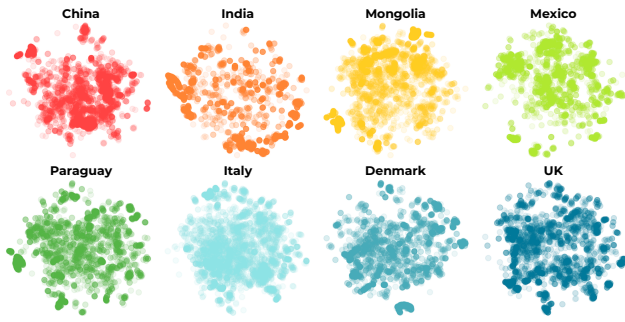
Figure 5: t-SNE fitted on a worldwide aggregation of countries, each country's eating events are plotted separately
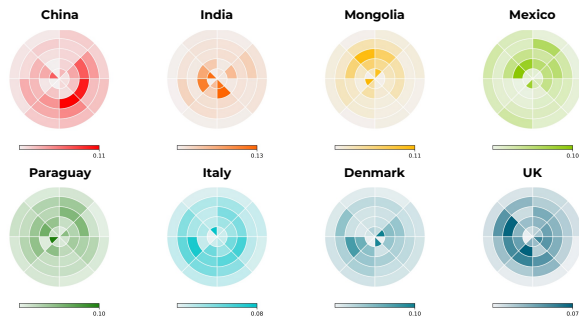

Figure 6: t-SNE fitted on the aggregation of all countries, the country-specific density of eating-reports in the 2D space

countries have similar phone usage patterns during eating events. In contrast, the sensed eating events from China, India, and Mongolia are located in opposite areas of the low-dimensional space, indicating differences in smartphone use and cultural practices.

5 INFERRING THE SOCIAL CONTEXT OF EATING (RQ2)

5.1 Methodology

The objective of this section of our study is to assess the capability of eating social context inference models to generalize across users and countries. We used smartphone sensor features collected within a time interval around eating events as the input space and the social context of eating as the target variable. We analyzed country-specific models as well as multi-country models. A series of experiments were carried out to compare the use of different inference methods such as multi-layer perceptron neural networks (MLPs), Random Forests, and Support Vector Classifiers. eXtreme Gradient Boosting (XGboost), a regularizing gradient boosting framework, [9] was chosen based on better performance and ease of adaptability and tuning across experimental setups.

Despite having observed a significant time-dependency in the eating behaviors, it was decided not to train separate models, e.g.,

weekend/weekdays, due to lack of data in weekends. However, models trained on time-related features only were used as a baseline to assess the added benefits of using smartphone sensing in comparison with a naive prediction solely based on time features. Finally, with a sample size ranging from 340 (India) to 12,667 (Italy) eating reports, and given the class imbalance, special care was given to addressing these issues when creating aggregations, performing data splits, and evaluating the model performance.

Population-Level models are based on the leave-k-participants-out cross-validation [16], which involves splitting data into training and testing sets based on individual users, creating disjoint sets between the two. The splitting is stratified to ensure that the percentage of Alone and With Others samples is maintained in both sets. This allows for testing the model's robustness against user-specific behaviors, simulating a scenario where the model is trained on a population, and new users join the system.

Hybrid models are based on K-fold cross-validation [16], and this approach involves randomly splitting the data into training and testing sets without considering individual users. The testing set contains a fraction of the eating social context reports from each user, and the splitting is done in a stratified manner to preserve the percentage of Alone and With Others samples in each split. This approach simulates a scenario where the model is trained on a population, and the same participants provide additional reports to the system, resulting in partially personalized models.

In all inference tasks, an interesting property of XGBoost was utilized, which is the sparsity-aware split finding technique [10]. This technique allows training on sparse data by letting tree branches handle missing values. Initially, models were trained with both imputed and sparse data. However, it was observed that the results were slightly better when using the sparse method. Therefore, the sparse method was chosen for all further training and testing.

The hyper-parameters of XGBoost have a large impact on its performance and generalizability. To optimize these parameters, nested cross-validation was used. The hyper-parameters considered were the step size shrinkage (*eta*), the minimum loss reduction required to make a split (*gamma*), the maximum tree depth (*max tree depth*), and the minimum child weight (*min child weight*). Grid search was used to find the best combination of hyper-parameters. Five train/test folds were used to evaluate the models' performance. Each train set was further divided into five nested cross-validation splits to determine the best parameters. The final hyper-parameters were selected by choosing the most frequently picked parameters among the five folds. To validate the performance of the models, a random binary vector was generated and used as a baseline to compare against the model's prediction. In addition to the time-based baseline, this baseline was used to verify that the prediction outperformed a random guess.

The countries were aggregated based on their continental region, clearly being aware of this oversimplifying assumption of similarity. Thus, Europe was formed by combining Italy, the UK, and Denmark; Latin America by combining Paraguay and Mexico; and Asia by combining Mongolia, India, and China. All countries were then aggregated to form a multi-country world dataset. An equal number of users were selected from each country to ensure that the number of eating reports in each country equaled the number of reports in the country with the least reported eating events. Stratified group K

Table 3: Mean ($\overline{A}$) and Standard Deviation (A_σ) of inference AUC for population-level and hybrid models, calculated with XGBoost and Baseline (trained on time based features)

	Population Level			Hybrid		
	Model	Baseline	Random	Model	Baseline	Random
China	.48 (.05)	.47 (.03)	0.5	.68 (.02)	.50 (.02)	0.5
Denmark	.52 (.08)	.50 (.06)	0.5	.74 (.04)	.57 (.03)	0.5
India	.50 (.04)	.49 (.03)	0.5	.75 (.06)	.58 (.05)	0.5
Italy	.58 (.01)	.57 (.02)	0.5	.64 (.01)	.57 (.01)	0.5
Mexico	.53 (.05)	.51 (.06)	0.5	.65 (.03)	.57 (.01)	0.5
Mongolia	.49 (.01)	.50 (.01)	0.5	.57 (.01)	.50 (.01)	0.5
Paraguay	.49 (.06)	.55 (.05)	0.5	.61 (.04)	.55 (.03)	0.5
UK	.55 (.04)	.53 (.03)	0.5	.75 (.04)	.56 (.02)	0.5
Latin America	.53 (.02)	.51 (.04)	0.5	.63 (.03)	.54 (.02)	0.5
Asia	.51 (.06)	.52 (.04)	0.5	.72 (.03)	.55 (.04)	0.5
Europe	.56 (.05)	.54 (.03)	0.5	.72 (.04)	.56 (.02)	0.5
World	.53 (.03)	.53 (.02)	0.5	.69 (.02)	.53 (.02)	0.5

Figure 7: XGBoost Feature Importance Values

fold splits were performed in each separate country to perform the cross-validation splits for population-level models before aggregating them into one training and one testing set. The same procedure was applied to hybrid models to maintain an equal amount of data from each country in train and test sets. To ensure the generalizability of the results on country aggregations, selected users for the majority of countries were resampled at each repetition.

The chosen metric for evaluating the results is the Area Under the Receiver Operating Characteristic (AUC). Due to the limited amount of data available, it was not feasible to perform more than five folds on the countries with less data. However, evaluating models on just five folds can lead to uneven performances between each run. Therefore, repeated stratified group K folds and repeated stratified K folds were conducted for the population level and hybrid experimental setups, respectively. The number of repetitions was set to 5, and the folds were randomized each time to obtain different data splits. To summarize, the models were evaluated on five repetitions of 5-fold validation, and the model performances were averaged over the resulting 25 folds. One of the objectives of building inference models is to leverage the feature importance as additional information about cross-country diversity. The feature importance coefficients were also averaged over the 25 folds.

5.2 Results

5.2.1 Population Level Results. In general, the country-specific models trained with the population-level approach were unsatisfactory, as shown in Table 3. The models exhibited uneven performance across countries, where social context seems more easily inferred in European countries than in Asia and Latin America. For instance, the models achieved AUC scores of 0.52 (0.08) for Denmark, 0.55 (0.04) for the UK, 0.58 (0.01) for Italy, and 0.56 (0.05) for Europe aggregate. The lower model variance observed in Italy could be explained by the high number of users, which provided the model with a more accurate picture of the population, leading to a comparatively robust model. Conversely, Denmark exhibited high model variance (0.08), indicating that the model either performed well or poorly, depending on the user split. This situation arises when the social context is reflected differently across users, causing the model to learn a trend that does not always apply to the remaining

users in the test set, depending on the split. This finding is consistent with the tight clusters observed for Denmark in Figure 4, pointing towards a variety of phone sensor features. Interestingly, the world model provided good results comparatively, whereas Asia and Latin America consistently outperformed the performance of their respective countries. This could be attributed to the increase in training samples being creating robust models to counter the diversity in data. However, this conclusion should not hold in the case where countries have sufficient data individually.

5.2.2 Hybrid Results. The inclusion of a fraction of each user's data in the models resulted in improvements in the results. For instance, the models for China, Mongolia, and Paraguay improved from showing no predictive power to showing reasonable results, with AUCs of 0.68 (0.02), 0.57 (0.01), and 0.61 (0.04), respectively. Generally, the country-specific models showed reasonable performance with an AUC of 0.75 (0.06) for India, 0.75 (0.04) for the UK, and 0.74 (0.04) for Denmark. The aggregations also provided good results with AUC of 0.72 (0.03), 0.72 (0.04), and 0.69 (0.02) for Asia, Europe, and World, respectively. These findings support the conclusions made for population-level setups. However, Latin America suffered from the country's aggregation. In summary, the hybrid models performed well and consistently outperformed the baseline based on time features. The improvements observed with hybrid model results are in line with the diversity of user practices observed in Figure 4.

5.2.3 Feature Importance. The use of XGBoost results in a loss of the natural explanatory power of models like Random Forests, but it is still possible to extract feature importances. XGBoost outputs a weight measure that reflects the number of times a feature is used to split the data across all trees. After extracting the feature importances at each evaluation split and averaging over the 25 folds, the values are displayed in Figure 7 for hybrid models. The features are ranked left to right, based on the mean across all configurations, to visualize better which features are shared across countries and continents. Higher feature importances in countries aggregations are not necessarily ones used in any of the separate countries, which leads us to think that feature importance cannot be approached as the sum of individual ones in a cross-country context. India and Asia had very strong feature importance on their most important features, which imbalanced the color scale and reduced the contrast in the rest of the plot. Therefore, the color scale was limited to a

certain threshold visible at the top of the color bar, which means all values beyond the threshold take a single color. It is worth noting that the countries with the most data, Italy and Mongolia, have a more spread-out use of features. The overall fading of the plots towards less important features shows that the different models tend to be using the same features, and the clear and dark tiles that stand out highlight features that are specific to one country. In summary, the feature importances resonate with the effect sizes listed in Table 2 with the difference that the model has the ability to leverage complex relations in the data that may not be visible otherwise.

6 DISCUSSION

6.1 Implications

This study provides insights into the contextual characteristics associated with eating events in different countries, which can aid in understanding the differences between eating events from a mobile sensing perspective. Our findings highlight the need to account for country biases when building machine learning models based on mobile sensing data, as these biases can arise from differences in demographic attributes, lifestyle, culture, and other factors. Further, the study shows how features associate with the social context of eating events in different countries and shows the importance of country diversity in how these events are sensed. The identification of different user practices associated with eating events in different countries provides useful insights for future data collection and highlights the impact of country diversity on the sensing of eating events. Therefore, we recommend that researchers consider relevant contextual factors and country differences during the planning phase of their studies for training models with multimodal data that captures a range of situational and behavioral contexts.

Our study also reveals that some countries exhibit a degree of similarity in terms of sensed eating events, but geographical proximity cannot be used as a proxy for similar user behaviors. While clear country-wise patterns and clusters arise when analyzing the countries, the difference in scores for population-level versus hybrid models, as well as the clusters observed in user behavior, highlight the fact that social context differs greatly across users, even in the same country. Moreover, the amount of user diversity among countries is subject to variation, with user practices being more or less spread out, and researchers should pay specific attention to countries showing a higher level of diversity among users when gathering data or designing studies. Finally, our study demonstrates that feature importance can be used to extract insights into user practices associated with eating events, highlighting the importance of accounting for the complex relations in the data that might not be visible otherwise.

6.2 Limitations and Future Work

This study focuses on the impact of geographical diversity on the eating social context, but it acknowledges that diversity is a multifaceted concept that encompasses more than just geographical shifts. The results suggest that assuming similar cultural and social behavior from people living in the same country only holds to a certain extent and that considering more granular aspects of diversity, such as socioeconomic background or cultural norms, could improve inference models' accuracy. Future work can focus on these aspects. Further, the use of additional APIs providing information about location, weather, and behavior could be considered in future studies for additional input of contextual information.

The study uses data that were collected at a single university in each country, which has implications on the representativeness of the samples of students that are considered. In addition, the number of participants is uneven between universities leading to a fairly high data imbalance between countries.

In addition, the data were collected during a period coinciding with the surge of the Covid-19 pandemic in Europe, Paraguay, and Mongolia, during which social distancing measures were imposed or recommended. While the pandemic likely influenced and altered participants' behavior and social practices, the study's findings are expected to hold in the future as remote work/study settings become more prevalent.

The analysis is limited to the two-class social context inference task of determining if an individual is alone or with others, while the dataset includes eight social contexts. While the variable offers valuable insight for mobile food diaries, a more granular approach to social context inference could help gain a better understanding of eating events. The decision to use the two-class inference was motivated by class imbalance, and the fact that the variable is well populated across countries, which is not the case for other class social contexts, making a similar analysis across countries more challenging. Further, the mobile sensing data used in this study can reveal sensitive information about the social context of people, which has ethical implications. Therefore, researchers must carefully consider and address ethical issues to maintain user confidentiality and safeguard against any potential ethical issues.

7 CONCLUSION

This study used a mobile sensing dataset and over 24K self-reports from 678 participants in eight countries, collected over a period of four weeks, to investigate how geographic diversity affects the eating social context inference. We performed an analysis and extracted key aspects of the behavioral and contextual differences that emerge from sensing eating events in different countries. We also evaluated country-specific and multi-country approaches trained on multimodal mobile sensing data to infer eating alone vs. with others, with population-level (non-personalized) and hybrid (partially personalized) models. In addition, the study of feature importance across different models provided additional insight into geographical diversity. Overall, we highlight the potential for mobile sensing-based machine learning models to generalize across geographically diverse settings, for inferring the social context of eating.

REFERENCES

[1] Saeed Abdullah, Nicholas Lane, and Tanzeem Choudhury. 2021. Towards Population Scale Activity Recognition: A Framework for Handling Data Diversity. *Proceedings of the AAAI Conference on Artificial Intelligence* 26, 1 (Sep. 2021).

[2] Karim Assi, Lakmal Meegahapola, William Droz, et al. 2023. Complex Daily Activities, Country-Level Diversity, and Smartphone Sensing: A Study in Denmark, Italy, Mongolia, Paraguay, and UK. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–23.

[3] Wageesha Bangamuarachchi, Anju Chamantha, Lakmal Meegahapola, Haeeun Kim, Salvador Ruiz-Correa, Indika Perera, and Daniel Gatica-Perez. 2023. Inferring Mood-While-Eating with Smartphone Sensing and Community-Based Model Personalization. *arXiv preprint arXiv:2306.00723* (2023).

[4] Wageesha Bangamuarachchi, Anju Chamantha, Lakmal Meegahapola, Salvador Ruiz-Correa, Indika Perera, and Daniel Gatica-Perez. 2022. Sensing Eating Events in Context: A Smartphone-Only Approach. *IEEE Access* 10 (2022), 61249–61264.

[5] Abdelkareem Bedri, Richard Li, Malcolm Haynes, Raj Prateek Kosaraju, Ishaan Grover, Temiloluwa Prioleau, et al. 2017. EarBit: Using Wearable Sensors to Detect Eating Episodes in Unconstrained Environments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 37 (sep 2017), 20 pages.

[6] Joan-Isaac Biel, Nathalie Martin, David Labbe, and Daniel Gatica-Perez. 2018. Bites'N'Bits: Inferring Eating Behavior from Contextual Mobile Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 125 (jan 2018), 33 pages.

[7] Mimi Bocanegra, Mailin Lemke, Roelof A.J. de Vries, and Geke D.S. Ludden. 2022. Commensality or Reverie in Eating? Exploring the Solo Dining Experience. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (Bengaluru, India) *(ICMI '22)*. ACM, 25–35.

[8] Meg Bruening, Irene Van Woerden, Michael Todd, Stephanie Brennhofer, Melissa N Laska, and Genevieve Dunton. 2016. A mobile ecological momentary assessment tool (devilSPARC) for nutrition and physical activity behaviors in college students: a validation study. *Journal of medical Internet research* 18, 7 (2016), e209.

[9] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. ACM, 785–794.

[10] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. ACM, 785–794.

[11] Jack Cohen. 1988. *Statistical Power Analysis for the behavioral sciences.*

[12] Tegan Cruwys, K.E. Bevelander, and Roel Hermans. 2015. Social modeling of eating: A review of when and why social influence affects food intake and choice. *Appetite* 86 (01 2015), 3–18.

[13] Roelof A. J. de Vries, Juliet A. M. Haarman, Emiel C. Harmsen, Dirk K. J. Heylen, and Hermie J. Hermens. 2020. The Sensory Interactive Table: Exploring the Social Space of Eating. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20)*. ACM, 689–693.

[14] Fausto Giunchiglia, Ivano Bison, Matteo Busso, Ronald Chenu-Abente, Marcelo Rodas, Mattia Zeni, Can Gunel, Giuseppe Veltri, Amalia De Götzen, Peter Kun, Amarsanaa Ganbold, Altangerel Chagnaa, George Gaskell, Sally Stares, Miriam Bidoglia, Luca Cernuzzi, Alethia Hume, Jose Luis Zarza, Hao Xu, Donglei Song, Shyam Diwakar, Chaitanya Nutakki, Salvador Ruiz Correa, Andrea-Rebeca Mendoza, Lakmal Meegahapola, and Daniel Gatica-Perez. 2022. A worldwide diversity pilot on daily routines and social practices (2020-2021). University of Trento Technical Report - DataScientia dataset descriptors. https://iris.unitn.it/handle/11572/338382.

[15] Simone Hantke, Maximilian Schmitt, Panagiotis Tzirakis, and Björn Schuller. 2018. EAT -: The ICMI 2018 Eating Analysis and Tracking Challenge. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO, USA) *(ICMI '18)*. ACM, 559–563.

[16] Trevor Hastie, Jerome Friedman, and Robert Tisbshirani. 2017. *The elements of Statistical Learning: Data Mining, Inference, and prediction.* Springer.

[17] C.P. Herman, J. Polivy, and Tullia Leone. 2005. The psychology of overeating. *Food, Diet and Obesity* (10 2005), 115–136.

[18] Marion M Hetherington, Annie S Anderson, Geraldine NM Norton, and Lisa Newson. 2006. Situational effects on meal intake: A comparison of eating alone and eating with others. *Physiology & behavior* 88, 4-5 (2006), 498–505.

[19] Margaret Jastran, Carole Bisogni, Jeffery Sobal, Christine Blake, and Carol Devine. 2008. Eating routines. Embedded, value based, modifiable, and reflective. *Appetite* 52 (10 2008), 127–36.

[20] Jisu Jung, Lyndal Wellard-Cole, Colin Cai, Irena Koprinska, Kalina Yacef, Margaret Allman-Farinelli, and Judy Kay. 2020. Foundations for systematic evaluation and benchmarking of a mobile food logger in a large-scale nutrition study. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (2020).

[21] Mohammed Khwaja, Sumer S. Vaid, Sara Zannone, Gabriella M. Harari, A. Aldo Faisal, and Aleksandar Matic. 2019. Modeling Personality vs. Modeling Personalidad. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (sep 2019), 1–24.

[22] Florian Labhart, Skanda Muralidhar, Benoit Massé, Lakmal Meegahapola, Emmanuel Kuntsche, and Daniel Gatica-Perez. 2021. Ten seconds of my nights: Exploring methods to measure brightness, loudness and attendance and their associations with alcohol use from video clips. *PLoS one* 16, 4 (2021), e0250443.

[23] Lakmal Meegahapola, William Droz, Peter Kun, et al. 2023. Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 176 (jan 2023), 32 pages.

[24] Lakmal Meegahapola and Daniel Gatica-Perez. 2020. Smartphone sensing for the well-being of young adults: A review. *IEEE Access* 9 (2020), 3374–3399.

[25] Lakmal Meegahapola, Florian Labhart, Thanh-Trung Phan, and Daniel Gatica-Perez. 2021. Examining the Social Context of Alcohol Drinking in Young Adults with Smartphone Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (sep 2021), 1–26.

[26] Lakmal Meegahapola, Salvador Ruiz-Correa, and Daniel Gatica-Perez. 2020. Alone or with others? understanding eating episodes of college students with mobile sensing. In *19th International Conference on Mobile and Ubiquitous Multimedia*. 162–166.

[27] Lakmal Meegahapola, Salvador Ruiz-Correa, and Daniel Gatica-Perez. 2020. Protecting mobile food diaries from getting too personal. In *Proceedings of the 19th International Conference on Mobile and Ubiquitous Multimedia*. 212–222.

[28] Lakmal Meegahapola, Salvador Ruiz-Correa, Viridiana del Carmen Robledo-Valero, Emilio Ernesto Hernandez-Huerfano, Leonardo Alvarez-Rivera, Ronald Chenu-Abente, and Daniel Gatica-Perez. 2021. One More Bite? Inferring Food Consumption Level of College Students Using Smartphone Sensing and Self-Reports. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 26 (mar 2021), 28 pages.

[29] Mehrab Bin Morshed, Samruddhi Shreeram Kulkarni, Richard Li, Koustuv Saha, Leah Galante Roper, Lama Nachman, Hong Lu, Lucia Mirabella, Sanjeev Srivastava, Munmun De Choudhury, Kaya de Barbaro, Thomas Ploetz, and Gregory D Abowd. 2020. A Real-Time Eating Detection System for Capturing Eating Moments and Triggering Ecological Momentary Assessments to Obtain Further Context: System Development and Validation Study. *JMIR mHealth and uHealth* 8, 12 (dec 2020), e20625.

[30] Mehrab Bin Morshed, Samruddhi Shreeram Kulkarni, Koustuv Saha, Richard Li, Leah G Roper, Lama Nachman, Hong Lu, Lucia Mirabella, et al. 2022. Food, Mood, Context: Examining College Students' Eating Context and Mental Well-Being. *ACM Transactions on Computing for Healthcare* (2022).

[31] Alexandre Nanchen, Lakmal Meegahapola, William Droz, and Daniel Gatica-Perez. 2023. Keep Sensors in Check: Disentangling Country-Level Generalization Issues in Mobile Sensor-Based Models with Diversity Scores. (2023).

[32] Kushal Patel and David Schlundt. 2001. Impact of moods and social context on eating behavior. *Appetite* 36 (05 2001), 111–8.

[33] Laura Schelenz, Ivano Bison, Matteo Busso, Amalia De Götzen, Daniel Gatica-Perez, Fausto Giunchiglia, Lakmal Meegahapola, and Salvador Ruiz-Correa. 2021. The theory, practice, and ethical challenges of designing a diversity-aware platform for social relations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. 905–915.

[34] Horace S Schwerin, John L Stanton, Jack L Smith, AM Riley Jr, and Barbara E Brett. 1982. Food, eating habits, and health: a further examination of the relationship between food eating patterns and nutritional health. *The American journal of clinical nutrition* 35, 5 (1982), 1319–1325.

[35] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. 2017. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. (11 2017).

[36] Jee-Seon Shim, Hai-Rim Shin, Kyungwon Oh, and Cherian Varghese. 2014. Dietary assessment methods in epidemiologic studies. *Epidemiology and health* 36 (07 2014), e2014009.

[37] Student. 1908. The probable error of a mean. *Biometrika* (1908), 1–25.

[38] Frances E. Thompson and Amy F. Subar. 2017. Dietary Assessment Methodology.

[39] Olga Troyanskaya, Mike Cantor, Gavin Sherlock, Trevor Hastie, Rob Tibshirani, David Botstein, and Russ Altman. 2001. Missing Value Estimation Methods for DNA Microarrays. *Bioinformatics* 17 (07 2001), 520–525.

[40] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605.

[41] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S. Kuehn, Jeremy F. Huckins, Margaret E. Morris, Paula S. Nurius, Eve A. Riskin, Shwetak Patel, Tim Althoff, Andrew Campbell, Anind K. Dey, and Jennifer Mankoff. 2023. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 190 (jan 2023), 34 pages.

[42] Luke Yates and Alan Warde. 2017. Eating together and eating alone: meal arrangements in British households. *The British Journal of Sociology* 68, 1 (2017).

[43] Shibo Zhang, Yuqi Zhao, Dzung Tri Nguyen, Runsheng Xu, Sougata Sen, Josiah Hester, and Nabil Alshurafa. 2020. NeckSense: A Multi-Sensor Necklace for Detecting Eating Activities in Free-Living Conditions. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 2, Article 72 (jun 2020), 26 pages.