# Learning About Social Context from Smartphone Data: Generalization Across Countries and Daily Life Moments

Aurel Ruben Mäder[†]
Idiap Research Institute & EPFL
Switzerland

Lakmal Meegahapola[†§]
Idiap Research Institute & EPFL
Switzerland

Daniel Gatica-Perez
Idiap Research Institute & EPFL
Switzerland

## ABSTRACT

Understanding how social situations unfold in people's daily lives is relevant to designing mobile systems that can support users in their personal goals, well-being, and activities. As an alternative to questionnaires, some studies have used passively collected smartphone sensor data to infer social context (i.e., being alone or not) with machine learning models. However, the few existing studies have focused on specific daily life occasions and limited geographic cohorts in one or two countries. This limits the understanding of how inference models work in terms of generalization to everyday life occasions and multiple countries. In this paper, we used a novel, large-scale, and multimodal smartphone sensing dataset with over 216K self-reports collected from 581 young adults in five countries (Mongolia, Italy, Denmark, UK, Paraguay), first to understand whether social context inference is feasible with sensor data, and then, to know how behavioral and country-level diversity affects inferences. We found that several sensors are informative of social context, that partially personalized multi-country models (trained and tested with data from all countries) and country-specific models (trained and tested within countries) can achieve similar performance above 90% AUC, and that models do not generalize well to unseen countries regardless of geographic proximity. These findings confirm the importance of the diversity of mobile data, to better understand social context inference models in different countries.

## CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in ubiquitous and mobile computing**; **Smartphones**; **Empirical studies in HCI**; • **Social and professional topics** → **Geographic characteristics**; **Cultural characteristics**.

## KEYWORDS

mobile sensing, smartphone sensing, multimodal, social context, machine learning, generalization, personalization, context-awareness, digital health

[†]co-primary, listed alphabetically
[§]now at ETH Zurich, Switzerland

## 1 INTRODUCTION

Human beings inherently need social interactions, community, and interpersonal relationships [9]. As studies have shown, a fulfilling social life is particularly important for the mental well-being of young adults [14, 35, 90]. Regular interaction with friends, family, or colleagues during adolescence positively impacts long-term mental and physical health. Conversely, social isolation, such as living alone or prolonged solitude, increases the likelihood of depression and other mental and physical health issues [9, 14, 40, 53, 84]. Therefore, understanding an individual's social context over time is beneficial in mobile health applications, enabling timely interventions and feedback [57, 58]. Beyond that, "a system is context-aware if it uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task" [30], and social context is a key attribute in making mobile applications context-aware, enhancing their relevance and effectiveness [23, 31, 47, 54, 67, 71]. Hence, inferring social context could also benefit context-aware computing, enabling devices and applications, like smartphones, to adapt functionalities (i.e., displaying notifications, silencing the phone, recommendations, etc.) based on the user's social setting. Hence, this inference not only improves user engagement through personalized interactions, but also aids digital health applications to provide timely support and enhance the overall well-being of users.

While there is no single definition of social context, the notion of whether a person is alone or not has been a key construct in previous studies [19, 42, 57, 58, 89]. Social context, similar to mood, location, and food consumption, has traditionally been studied using surveys and ecological momentary assessments [10, 57]. However, these techniques often lead to user burden and sparse data. As an alternative, mobile sensing, which involves using smartphone and wearable sensor data-based inferences, offers a passive method for data collection [56]. This approach is a part of the broader fields of ubiquitous computing and human-computer interaction, focusing on passively sensing a wide range of individual contexts [2, 26, 29, 62, 70, 88, 94]. Nevertheless, accurately determining whether a person is alone or not with smartphone sensing is still challenging [3, 19, 41, 56, 58, 79, 93, 94]. Previous studies in this direction have generally focused on specific situations, such as during eating or drinking events, or within specific populations like

those with depression [19, 56, 58]. This leaves a research gap in understanding how social context interplays with various behavioral and contextual factors, including mood, location, time of day, and activities. Moreover, the effectiveness of smartphone sensor-based models in inferring social context in daily life has yet to be fully explored.

Data diversification in machine learning involves using a heterogeneous dataset for training models to enhance their representativeness and ability to generalize across different populations [37]. This concept, vital in domains like computer vision [18, 77] and natural language processing [15, 86], focuses on the country of origin of data as a crucial diversity factor. Such diversity not only enhances model robustness, but also ensures fairness in the outcomes, benefiting a wide range of users globally [32, 97]. However, applying data diversification to mobile sensing has been challenging in research, due to the scarcity of large-scale datasets from various countries, collected with similar protocols [5, 55]. The large cost and effort required to gather phone sensor data across multiple countries have been major obstacles. Hence, only a few studies have examined the effect of country diversity in mobile sensing-based models [44, 55, 61], and notably, none have focused on inferring social context [65]. Hence, we ask the following research questions.

• **RQ1:** Which situational and behavioral aspects are associated with different social contexts in a five-country dataset?

• **RQ2:** Can the social context of daily life moments be inferred using smartphone sensing with a non-personalized, one-size-fits-all, multi-country approach? Does the separation of sensing data by its origin countries aid the performance?

• **RQ3:** What is the effect of training partially personalized models on model performance?

• **RQ4:** Do social context inference models generalize well to unseen countries?

In answering the above RQs, in line with terms provided in Table 1, our work provides the following contributions:

• **Contribution 1:** We analyzed a large smartphone sensing dataset collected from 580 young adults across five countries: Denmark, Italy, Mongolia, Paraguay, and the UK. This dataset comprises passive sensing data from various modalities (Table 6) and over 216K social context self-reports gathered over a period of four weeks. A statistical analysis revealed that features from modalities such as the amount of app usage (i.e., tools, communication, productivity, social, etc.), time spent doing activities (i.e., in a vehicle, on foot, walking, etc.), step count, and screen episodes were among the top ten in terms of statistical significance for discerning alone versus not alone, based on mixed effects models. Furthermore, we observed that these features vary across countries, highlighting the diversity in individual behaviors and the corresponding contextual factors in different social contexts.

• **Contribution 2:** Our study compared three approaches to social context inference: country-specific, country-agnostic, and multi-country. The multi-country approach yielded a moderate AUC (67.61% without feature selection, 68.0% with it), while country-specific models showed similar performance, generally within a 5% performance difference. Notably, Paraguay and Mongolia exceeded the multi-country model's performance, unlike the other three countries. This finding challenges the notion common in previous multi-country sensing studies that country-specific models are

always superior, suggesting that for social context inference, this may not necessarily be the case.

• **Contribution 3:** We found that hybrid models yielded similar performance for both multi-country and country-specific approaches. While the multi-country model achieved an AUC of 95.3%, countries such as the UK, Denmark, and Mongolia slightly outperformed this performance, whereas Italy and Paraguay were marginally below it. These results show that with personalization, social context in everyday life could be inferred with high performance.

• **Contribution 4:** In the specific cases of mood inference and complex daily activity recognition, even though prior work found that models might generalize reasonably to geographically closer countries in Europe, we did not find such associations in social context inference performance. In fact, in all cases, models did not generalize well to unseen countries regardless of the geographical proximity. This raises concerns regarding where a trained social context inference models could be deployed.

## 2 BACKGROUND AND RELATED WORK

Next, we surveyed various lines of research that our work draws upon and grouped them into three main areas: *(i)* social context inference; *(ii)* context recognition and health sensing; and *(iii)* leveraging diversity-awareness.

### 2.1 Social Context Recognition

The social context of individuals is a multifaceted concept encompassing complex phenomena operating on various dimensions. Two dimensions that are of particular importance are *(i)* the number of people in the individual's surroundings, and *(ii)* the type of relationship associated with the individual's current social context, such as family, colleagues, partner, etc. [2, 12, 54, 74]. A fundamental consideration in this context is whether an individual is alone or not, an aspect that is often discussed in mobile sensing studies exploring topics like depression, mood, eating behavior, drinking behavior, and overall mental well-being [11, 19, 57, 58, 76]. Hence, as mentioned in Table 1, we use alone or not alone as a fundamental construct for social context[1].

Previous studies examining the task of social context sensing have frequently delved into ethical and technical considerations [3] or have aimed to contextualize the task within the broader scope of social networks and user behavior [2, 89]. An intriguing paradox related to the internet, particularly relevant in such analyses, is the way mobile phones serve as platforms to replicate and, to some extent, replace physical social contexts. This phenomenon is particularly notable among young adults who increasingly substitute face-to-face interactions with mobile-mediated communication [83]. It could be argued that the "connection technology" of the Internet, in many instances, has heightened social disconnection

---

[1]In this study, we chose the term "social context" instead of other possible terms such as "social encounters" or "social interactions". This decision was guided by the ambiguity inherent in these alternatives. "Social encounter" and "social interaction" imply a degree of engagement that we could not definitively ascertain given the study design [39, 41]. Our term aligns with previous ubicomp and human-computer interaction studies [42, 57, 58], which often use "social context" to differentiate between being alone versus not alone. We acknowledge this as a limitation and recognize that our use of "social context" carries a narrower definition specific to our study's design.

**Table 1: Terminology and definitions used in the paper.**

| Term | Description |
|---|---|
| Social Context | If a person is Alone or Not (i.e., with friends, family, or others). We borrow this definition from prior studies in human-computer interaction and ubiquitous computing domain [19, 42, 57, 58]. |
| Country-Specific Approach | Machine learning models are trained and tested only on data from one specific country. |
| Country-Agnostic Approach | Machine learning models are trained on one country or a set of countries and tested on another country not seen in training. Such models are usually trained with the assumption that the model could be deployed to other countries, hence the term agnostic. |
| Multi-Country Approach | Machine learning models are trained on a set of countries and tested again on the same set of countries. This is the generic way of training models when data from multiple countries are available. |
| Population-Level Model (PLM) | Training and Testing splits have a disjoint set of users. Represents a case where a machine learning model trained with a population is deployed to a mobile app that a new user uses. Hence, end-user data have not been used in model training, leading to non-personalized and generic one-size-fits-all models. |
| Hybrid Model (HM) | Training and testing splits do not have a disjoint set of users. Represent a case when a mobile app user uses a machine learning model for some time, and data from the user is used in re-training (or fine-tuning in the case of neural networks) models. Hence, this approach leads to partially personalized models. |

[2]. Another recent study [89] examined the social context, which they referred to as being alone vs. dyadic vs. in a group, during virtual conversations of a group of socially anxious undergraduate students. With a multi-task learning model, they showed that passive sensor data from Empatica E4 devices could detect contexts relevant to social anxiety. However, it is crucial to clarify that in the context of our study, the task of social context inference pertains to the presence of individuals in physical, in-person settings and does not encompass virtual connections formed through mobile phones, virtual conversations via video conferencing applications, or social networks.

Several studies have attempted to infer in-person social context, specifically whether individuals are alone or not, utilizing sensing data. In a study by Burns et al. [19], sensor data was collected from eight participants over eight weeks, involving a total of 38 different sensors. They employed regression trees to predict an individual's social context, achieving mean accuracies of 80% across participants. It is worth noting that this study focused on a small sample of individuals diagnosed with depressive symptoms, leaving questions regarding the generalizability of their findings to a broader population without such conditions. Other studies have concentrated on social context inference within specific contextual situations. For instance, in a study by Meegahapola et al. [58], the authors explored the social context of university students during eating episodes, emphasizing the significance of determining whether an individual is alone or not for providing timely interventions and feedback in mobile food diaries. They trained and evaluated their models with data collected during eating events. Further, they used population-level models and achieved accuracies ranging from 70% to 75% for Switzerland and Mexico, separately. However, a limitation of this study is that direct comparisons between the results of the two countries are challenging due to variations in data collection methods, sensor usage, time periods, and protocols. Additionally, their

sample sizes were relatively smaller compared to the current study. In a related work by Meegahapola et al. [57], the social context of young adults during alcohol drinking episodes was examined. Using population-level models, they attained accuracies ranging from 80% to 87% for various two-class social context inferences, including alone or not, as well as alone or with friends/family/partners. However, this study was based on a dataset collected in a single European country and exclusively revolved around the social context of drinking events. Furthermore, while both of these studies involved healthy participants without eating, drinking, or mental well-being-related disorders, they were primarily centered on specific contextual scenarios like eating and drinking. Therefore, the social context of young adults in daily life moments has remained relatively unexplored in prior research employing smartphone sensing data. Additionally, prior work has not delved deeply into analyzing country-level diversity and model personalization approaches, which are also focal points of our study.

## 2.2 Context Recognition and Health Sensing

In the mobile sensing literature, social context inference can be seen as a context recognition task. Following the framework that categorizes mobile sensing-based inferences into three pillars: behavior, person, and context, social context inference falls within the third pillar [56]. Furthermore, context awareness in devices offers significant benefits, such as battery life optimization. For example, reducing sensor sampling rates like GPS and cellular sensors during low battery usage periods, like sleep or social events, can conserve battery [94]. Devices can also adapt behaviors to specific situations, like disabling calls or switching to silent mode during meetings [50]. Additionally, mobile sensing has been extensively used to detect health-related contexts, including stress, mood, depression, energy expenditure, and eating and drinking behaviors, along with their social implications [4, 6, 19, 20, 48, 58]. These studies aim to identify,

analyze, and intervene in health-related situations. Accurate context identification is crucial for further analysis or intervention [11]. Once a health-related context is identified, it enables the exploration of associated factors, like peer pressure in heavy drinking episodes among college students [57], or depression linked to prolonged solitude [53]. Applications developed for health interventions, such as those for depression [19, 87] or mood tracking [76], are based on accurately identifying these contexts. Hence, social context recognition, given its direct impact on mental and physical well-being, is a crucial subset of health sensing [40, 53]. However, social context recognition in broad daily life situations, from smartphone data, has rarely been explored.

## 2.3 Leveraging Diversity Awareness

Systems that rely on learning patterns from data, such as classification and regression models, may inadvertently encode the internal biases present in the training data [97]. Consequently, these biases can manifest in the form of biased predictions, particularly concerning gender, race, socio-economic status, or the psychological profile of new users, when applying pre-trained models. In the realm of computer vision research, the impact of biased training data has been extensively examined. For example, one study evaluated gender classification models across various subgroups defined by gender and race [18]. This study revealed considerable disparities in model performance, with dark-skinned women experiencing error rates as high as 34.7%, while error rates for light-skinned men were only around 0.8%. These biases frequently stem from the under-representation of certain groups within the datasets used for model training.

Popular image classification datasets like ImageNet [27] and Open Images [46] have faced criticism for under-representing certain regions, with only a small fraction of images from countries like China and India [77]. This leads to biased predictions when models trained on these datasets are used in these countries [77]. In natural language processing, similar biases are evident, such as gender stereotypes in word embeddings and gender-biased translations in software like Google Translate [15, 68]. Addressing these imbalances is crucial for improving both the performance and fairness of machine learning systems [97]. In mobile sensing, diversity across datasets poses unique challenges, with biases not as apparent as in other fields. Strategies like partitioning datasets into smaller subsets [1] and considering wearables' usage diversity [21] have been proposed, but the impact of geographical diversity on sensor-based tasks remains underexplored [65], underscoring the need for thorough investigation in this area.

To enhance diversity in mobile sensing datasets, collecting data from diverse geographic regions is a key [75]. This approach, however, has seen limited exploration in mobile sensing research. For example, Khwaja et al. [44] analyzed data from five countries (the UK, Spain, Colombia, Peru, and Chile) for personality trait inference, finding that country-specific models were more effective than multi-country models, with performance improvements ranging from 3% to 7%. Similar benefits of country-specific models were observed in mood inference and complex daily activity recognition [5, 55]. However, the applicability of these findings to social context inference remains to be further examined.

In summary, this study introduces several novel aspects to the field of social context inference using smartphone sensing data. Firstly, it expands on previous research by identifying key smartphone sensing features and behavioral and contextual factors from self-reports that are instrumental in inferring social context, uniquely extending the analysis to multiple countries. This is conducted using an extensive, multimodal smartphone sensing dataset, which sets our study apart in terms of model generalization to everyday life situations across different countries. Secondly, we explore the feasibility of social context inference using a generic multi-country approach, employing both population-level and hybrid models. This represents one of the earliest efforts in social context inference with smartphone data from a considerable number of participants. Thirdly, our study includes a comprehensive comparative analysis using frameworks from previous research, examining country-specific, country-agnostic, and multi-country approaches. This analysis shows the effectiveness of multi-country models in social context inference. However, it also reveals the challenges in generalizing these models to unobserved countries, underscoring the complexity of social context inference in diverse settings.

## 3 DATASET

For this study, we used a novel smartphone sensing dataset from our previous work, that was collected in a study conducted simultaneously in five countries [5, 36]. To the best of our knowledge, this is the only dataset that allows us to perform the analysis around social context during daily life moments, across multiple countries, with smartphone data. The data were collected for four weeks from college students of the following five universities: Aalborg University (Denmark), London School of Economics and Political Science (the United Kingdom), the National University of Mongolia (Mongolia), the Universidad Católica "Nuestra Señora de la Asunción" (Paraguay), and the University of Trento (Italy). The study participants contributed three different kinds of data: *(i)* closed-ended questionnaires, *(ii)* hourly self-reports throughout the day, and *(iii)* sensor data.

*(i)* The closed-ended questionnaires consisted of three separate questionnaires, which were administered to the study participants before the start of the study, before the start of the sensor data collection, and after two weeks of sensor data collection. This design allowed the collection of a large amount of information from participants without overburdening them. The questionnaires were designed to capture surface diversity information (age, sex, country, etc.) and deep diversity information (personality, values, intelligence, etc.— with validated scales).

*(ii)* During the data collection, a mobile app was deployed. Using the app sensor data were captured passively, and participants also self-reported details about their behavior and context. These hourly self-reports were meant to capture how people spend their time. A notification was sent to participants once every hour, and they were asked to report their current activities (studying, cooking, etc.), semantic location (home, library, etc.), mood (valence— in a five-point scale from very negative to very position), and social context (alone, friends, relatives, classmates, roommates, colleagues, partner, other). Hence, the dependent variable for the discussed

**Table 2: Dataset Summary.**

| Country | # of Participants | # of Participants With Sufficient Data | # of Reports | Alone Percentage |
|---|---|---|---|---|
| UK | 72 | 53 | 26,687 | 69.05% |
| Denmark | 25 | 17 | 10,058 | 49.63% |
| Italy | 238 | 221 | 151,335 | 64.78% |
| Paraguay | 29 | 24 | 9,745 | 54.52% |
| Mongolia | 217 | 138 | 94,249 | 24.52% |
| All | 581 | 453 | 292,074 | 51.31% |

alone-or-not inference is based on the answers from the social context self-report.

*(iii)* The sensor data originally consisted of 34 different sensors, which are divided into continuous and interaction sensing. Continuous sensing modalities (captures sensor data regardless of user activities) included activity type, step count, Bluetooth, WiFi, location, cellular, and proximity; and interaction sensing modalities (measures the interactions participants have with the phone) included app usage, touch events, the screen on/off episodes, notification, etc.

In total, four weeks of sensor and self-report data were collected from study participants. However, our analysis revealed that the completeness and the number of responses to time diaries and questionnaires vary greatly between different countries (Table 2) and different participants (Figure 2). Furthermore, the quality and quantity of sensor data also vary between different participants. This is also observable in the amount of missing sensor data (Nan) for certain sensors or study participants (Figure 1). A more detailed discussion about the dataset and feature extraction can be found in papers by original authors that presented the dataset [5, 55].

The final analysis of this study consists of inferring the social context of a participant. Given the large size of the original dataset, which amounts to approximately 30GB of data in different data formats, the sensor data were processed and aggregated to sensible features at the self-report level. To do this, we used an approach similar to prior mobile sensing studies where sensor data are aggregated around the time of an in-situ self-report [6, 13, 57, 58, 76]. More specifically, all sensor data corresponding to ten minutes around the social context self-report was aggregated to features as summarized in Table 6. Thus, the aggregation time frame corresponds to a ten-minute window around the answering time of a self-report, which captures the individual's current activity, semantic location, mood, and, finally, the social context. It is worth noting that various time windows ranging from 1 to 20 minutes were explored. Extending beyond a 20-minute window was not feasible due to the risk of data leakage from overlapping sensor data segments. Among the tested durations, the ten-minute window emerged as reasonable, offering satisfactory performance. Additionally, this specific time frame aligns with previous studies that used the same dataset [5, 55], thereby facilitating meaningful comparisons. Moreover, in total, 117 distinct sensor features were created in line with the work of the original authors [5], including 48 features that correspond to app usage time period based on app categories in Google Playstore (e.g., action, dating, music, puzzle, social, etc.) similar to [48, 73]. Additionally, time and day features were also

included in the analysis, where the day of the week corresponds to a categorical variable, weekday or weekend, while the time was noted with the hour of the corresponding time dairy event (a numeric value between 0 and 23). Furthermore, missing data markers were introduced, which indicate when data is missing for certain sensor groups [49]. In summary, the smartphone sensor features of the final dataset used for inference are detailed in Table 6 in the Appendix.

## 4 EXPERIMENTAL SETUP

### 4.1 RQ1: Situational and Contextual Aspects Around Social Contexts

The purpose of this analysis was to investigate various situational and behavioral cues related to social context (alone or not) that can be extracted from self-reports and sensor data. This examination provides us with a comprehensive overview of the dataset and positions our study in relation to prior research. Initially, the original dataset featured detailed social context categories, including labels such as alone, with classmates, with colleagues, with a partner, with relatives, with roommates, and with others. Utilizing these labels, we examined how patterns of fine-grained social contexts differed across countries (Figure 3). Subsequently, by aggregating data from all countries, we explored how social context varied across these fine-grained categories throughout the day (Figure 4). Next, as previous studies have linked social context to mood (discussed in Section 2), we conducted an analysis focusing on binary social context (i.e., alone or not) across different levels of valence, assessed using a five-point Likert scale ranging from very positive to very negative. Lastly, we delved into the variations in social context across different locations and activities (Figure 5).

To identify which sensor features, including the hour of the day and the day of the week, are most indicative of whether someone is alone or not, we used mixed effects models [78]. The results are presented in Table 3 sorted by the descending order of statistical significance (p-values). This is done, ensuring that the most robust and least likely chance-based findings are presented first. This approach also aids in swiftly identifying the most relevant features impacting the dependent variable, making it easier to focus on key variables for further investigation or discussion. Furthermore, these experiments were conducted separately for each of the five countries, comparing the data distributions of sensing features when a participant is alone and when they are not alone. Participants were chosen as the random effect, to make sure that in-person variations are taken into account when calculating p-values. To
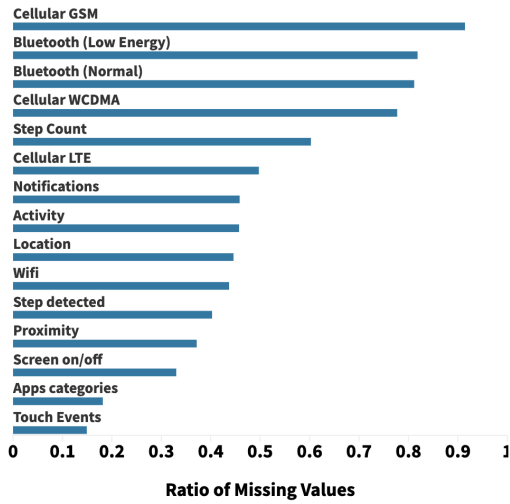
**Figure 1: Ratio of missing data (Nan values) in percentage aggregated to sensor group.**
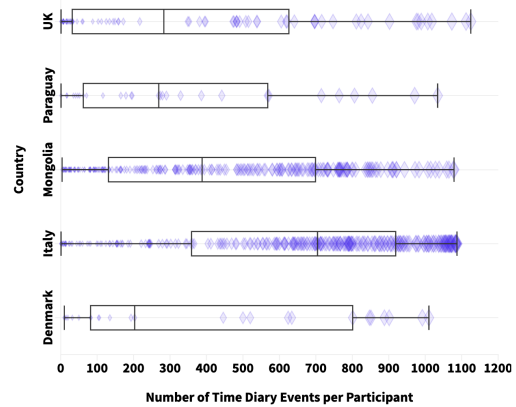


**Figure 2: Boxplot of the number of self-reports per participant, grouped by country. The boxes extend till the upper and lower quartiles, the median is indicated by an orange line, and whiskers extend till 1.5x IQR.**

account for multiple comparisons, we adjusted the p-values using the Bonferroni correction [85].

## 4.2 RQ2, RQ3 & RQ4: Social Context Inference

To gain a deeper insight into the importance of country diversity in social context inference, we adopted the strategy of dividing our dataset into several subsets [1], each corresponding to the origin of the data [44, 58]. Given that countries can exhibit distinct cultural and socio-economic norms, we presume that these factors influence people's behavior, smartphone usage, and social interactions [44]. Thus, our fundamental assumption posits that data originating from a specific country is better suited for training and constructing a model specific to that country. To test this assumption rigorously, we needed to compare the results of models trained on different data splits. Therefore, drawing inspiration from prior research that has explored multi-country data [5, 44, 55], we used several experimental approaches: *i)* Multi-Country: training and testing with all available countries; *ii)* Country-Specific: training and testing within the same country; and *iii)* Country-Agnostic: training on one or more countries and testing on an unseen country. Indeed, these scenarios closely mirror real-world practices, where companies often endeavor to gather smartphone and wearable sensor data from diverse countries to train their models. However, there are instances where collecting data from multiple countries is impractical. In such cases, models are trained within one country or using data from a select few countries and subsequently deployed to other countries that were not included in the model's training dataset.

Additionally, we trained various machine learning models, including logistic regression, random forest classifiers, XGBoost, and AdaBoost. Initially, all models were trained using all available sensor features. Subsequently, we employed sequential forward feature selection (FS) [69] to identify the most predictive features for training in both country-specific and multi-country approaches. Furthermore, drawing inspiration from previous research [55, 60],

we developed two types of models with varying levels of personalization:

- Population-level models (non-personalized): Similar to leave-k-users-out strategy [2]—that is, data in training and testing splits are not from the same users, hence ≈ 80:20 training and testing split.
- Hybrid models (partially personalized): First, data in training and testing splits are not from the same users, and we did an ≈ 60:40 split. Then, add 50% of data from testing users into the training set to achieve partial personalization for users in the testing set, leading to ≈ 80:20 training and testing split. Moreover, temporally, only the first half of data produced by testing users are removed from the testing set and included in the training set to make sure there is no temporal leakage [43]. That is, no data points from the future of the testing users are used in model training.

In summary, to address **RQ2**, encompassing *(i)* the feasibility of employing a generic, non-personalized model for social context inference using multimodal smartphone sensor data, and *(ii)* the potential performance improvements through data separation by countries, we implemented both multi-country and country-specific approaches with population-level models. Subsequently, to explore **RQ3**, which delves into the impact of partial personalization on model performance, we leveraged multi-country and country-specific approaches, this time with hybrid models. Lastly, to investigate **RQ4**, pertaining to the generalizability of models

---

[2]In comparing leave-one-out and leave-k-out (k-fold, but without data from the same user across folds) evaluation methods [60], we find that while leave-one-out is thorough in using almost all available data for training except for the testing user, it is computationally intensive, especially for large datasets. Leave-k-out, on the other hand, offers more flexibility and efficiency. This approach is particularly useful when dealing with larger datasets, as it balances the need for thorough evaluation with practical constraints on computational resources. Considering these factors, we chose to use the leave-k-out method for our study. This choice was driven by the need to efficiently handle our sizable dataset while still ensuring a robust evaluation of model performance.

trained in one or more countries to unseen country data, we utilized the country-agnostic approach with population-level models in two distinct setups: Setup 1, where models were trained with data from a single country and tested on all other countries separately; and Setup 2, where models were trained with data from four countries and tested on the remaining country.

In all cases, data were randomly sampled, and models were randomly initiated ten times with different seeds to obtain results for each setup, and results were averaged to obtain the mean and standard deviation of the area under the receiver operating characteristic curve (AUC) and F1-score metrics. To deal with the class imbalance and, for some countries, the small amount of data, Synthetic Minority Over-sampling (SMOTE) [22] was used in training while evaluation was performed on the testing set. All results can be evaluated against a majority baseline AUC of 50%. Given run time restrictions, GPU implementations of the models were used. All models were run in Python with commonly used libraries such as scikit-learn [64] for logistic regression and Ada Boost, cuML [72] as it implements a random forest model on the GPU, and lastly, a python XGBoost implementation [24]. Finally, in all cases, the hyper-parameter search was done using grid search (e.g., random forest: number of trees, max depth, min sample split; XGboost: learning rate, min split loss, max depth, reg lambda, etc.).

Furthermore, behavioral sensor data often exhibit sparsity and tend to contain high volumes of missing data [92]. Within this dataset, missing data can be attributed to various distinct scenarios: *(i)* sensor malfunction, *(ii)* sensor damage or absence, and *(iii)* intentional sensor deactivation by the user (e.g., disabling Bluetooth/WiFi, activating flight mode, etc.). However, distinguishing between these cases based solely on the available sensor data is challenging, and as in previous studies [66, 80], all these scenarios are treated equivalently. The extent of missing data also varies depending on the sensor (Figure 1) and the country where the experiment was conducted (e.g., 60% of Mongolian sensor data is missing). To address the issue of missing data, the first step involved discarding sensor groups with over 90% missing data, aligning with prior work [73]. This exclusion specifically applied to the 'cellular_gsm' features (refer to Figure 1). Subsequently, the remaining missing observations were estimated using the k-nearest-neighbor (kNN) algorithm [92, 96], with an experimentally determined optimal value of two for k. Following the procedure outlined in prior research [5, 92, 96], estimation was performed on the training set, and the learned model was then employed to impute missing data in the testing sets. To retain information that might be encoded in the missing entries of a given sensor group, binary markers were used to denote that the data had been interpolated [49].

## 5 RESULTS

### 5.1 RQ1: Situational and Contextual Aspects Around Social Contexts

As illustrated in Table 2, the number of study participants and the number of self-reports varied depending on the location of the pilot study. In total, 581 study participants provided self-reports regarding their social context and also supplied good-quality sensor data. The number of self-reports submitted by each participant displayed substantial variation, contingent upon individual participants and

their respective countries. Figure 2 visually depicts this diversity, indicating that the median count of self-reports per participant ranged from 200 (Denmark) to 800 (Italy). On average, participants contributed approximately 300 to 400 self-reports. Nevertheless, there were some users who provided well in excess of 1,000 self-reports, while others contributed close to zero self-reports despite providing sensor data over numerous days. To ensure an adequate volume of data for model personalization, individuals with fewer than 100 self-reports or fewer than six instances of the minority class were excluded from the analysis, resulting in a total of 453 participants for the subsequent analysis (Table 2).

The distribution of social contexts exhibits differences depending on the participant's respective country, as evident in Figure 3. In most countries, with the exception of Mongolia, individuals predominantly self-reported to have been alone. However, in Mongolia, participants reported to have been in the company of their families in over 60% of cases. This could suggest varying cultural norms related to living arrangements or family interactions. Although it is not possible to generalize the culture and practices of the entire Asian region to Mongolia, prior work has suggested that Asian households often have stronger family ties compared to Western societies [38]. Consequently, young adults in these societies may tend to live with their parents for longer periods [25, 45, 95]. It is indisputable that the participant's country influenced the prevalence of different social context types (e.g., alone, with family/friends/partner, etc.) within the dataset. For example, in Denmark and the UK, participants reported spending more time with their partners, sometimes even more than with their families. Country-specific differences in the frequency and nature of social contexts are corroborated by findings in life sciences literature [82] and recent reports [63].

In addition to the country-specific impact on the frequency and nature of social contexts among participants, a temporal effect was also evident. Figure 4 illustrates how the distribution of social context reports varies by the hour of the day. Participants reported to be predominantly alone or with their romantic partners during the morning hours. As the day progressed to lunch and dinner times, participants reported to have spent their time (particularly eating) with their families.

Figure 5 depicts the co-occurrence of specific moods (positive and negative valence), activities, and semantic locations with the social context (alone or not). Regarding concurrent activities and whether someone is alone or not, activities such as studying, listening to lectures, sleeping, and personal care were predominantly carried out alone. In contrast, activities involving social interaction, such as family care, eating, and nurturing one's social life, were typically done with others. Furthermore, regarding semantic location, participants reported having been alone at home on 125,270 instances, which could be influenced by the COVID-19 pandemic and the prevalence of remote work and study environments [33]. However, given the remote work culture that is prevalent even after the pandemic, we could expect these patterns to hold to an extent [17, 51, 52]. In terms of gender-specific differences, there were no notable disparities between men and women. Men reported being alone in 54.6% of all self-reports, while women reported to be slightly less alone, accounting for 49.1% of all self-reports. Additionally, women reported to have been with their families, averaging
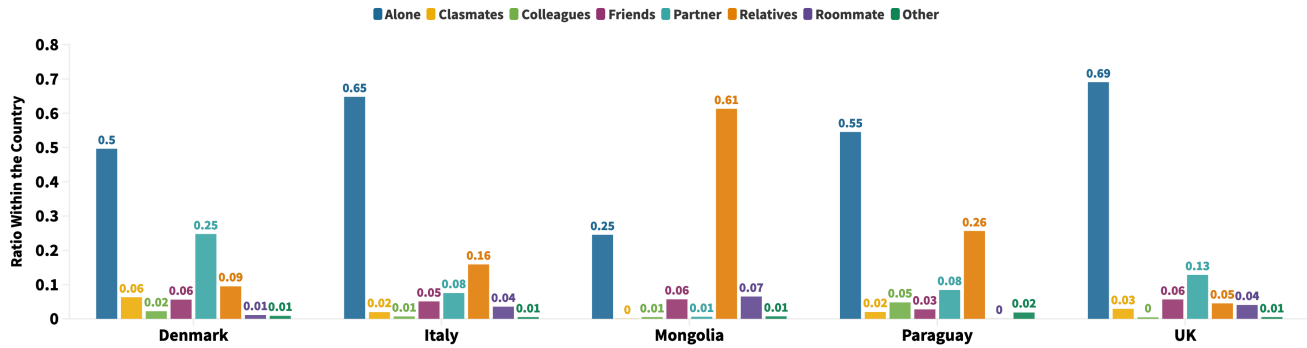
**Figure 3: Distribution of social context across countries, as a ratio of the total number of self-reports.**
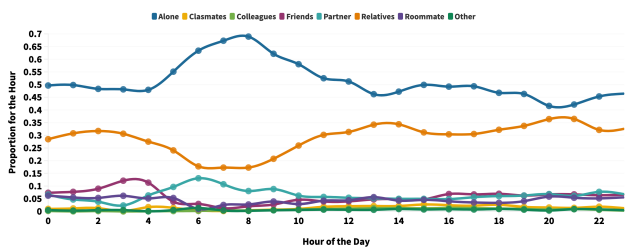


**Figure 4: Social context distribution according to the hour of the day, across all participants.**

31.6%, which is approximately 5% more than men. All other distinctions between women and men were minor in the broader context. These findings align with existing research indicating that young men tend to experience more loneliness [8] and tend to be alone more in certain contexts [34]. Prior work also discussed similar findings that women tend to seek more social ties compared to men [81]. However, it is important to note that being alone could be associated with a wide range of factors, including culture, geography, age, etc., in addition to gender [8].

The statistical analysis presented in Table 3 highlights the top ten sensing features effective in discerning whether a person is alone or not, based on the p-value after Bonferroni correction. This analysis accounts for both fixed (features) and random effects (study participants), offering a nuanced understanding of the data. In Denmark, app usage features such as news and magazines, simulation, and books and references show moderate to high negative coefficients, indicating that using these apps is inversely related to being with others. Conversely, in Paraguay, app usage features like finance and productivity, along with the number of notifications received, have smaller negative coefficients. In the UK, features related to user presence and proximity are statistically significant but show almost no relationship with social context. Similarly, in Mongolia, features like trivia, comics, weather, and arts and design have notable negative relationships with being alone. Italy displays a unique pattern where physical activities such as being in a vehicle, on foot, and walking, though statistically significant, have very small coefficients, showing no clear association. This pattern, at

least in terms of feature set, is akin to the UK and Paraguay but contrasts with Denmark and Mongolia, where app usage features have the top statistical significance values. It is also worth noting that, in the UK, except for two features (user_presense_time and proximity_minimum), other features had $p \geq 0.05$, suggesting no statistical significance. The 'All' category, aggregating data from all countries, indicates that app features like news and magazines, books and references, maps, sports, and simulation are generally negatively associated with being alone. These findings highlight country-specific differences in the relationship between sensor data features and social context, emphasizing the need to consider user behavior diversity in different countries for understanding and inferring everyday life behaviors with multimodal sensor data [5, 44].

## 5.2 RQ2: Social Context Inference Without Personalization

This section addresses the question of whether social context can be inferred using smartphone sensing data (**RQ2**). A comparison of the results obtained is presented in Table 4. Overall, in the generic multi-country setup, random forest models demonstrated the best performance. This performance was approximately 2% higher than that of XGBoost. Consequently, all subsequent results are only shown using random forest models for the sake of brevity. The test AUC of the random forest models ranged from 63% to 73%, depending on the country. However, due to the relatively small dataset sizes in some countries (e.g., only 10K self-reports in Denmark and 9.7K self-reports in Paraguay) and the high number of features, we employed feature selection to explore whether performance could be improved.

In Figure 6 and Figure 7, we compare the AUC and F1-Score, respectively, of different country-specific models with and without feature selection. The improvements in AUC vary depending on the country and, consequently, the number of participants, ranging from 0.3% (Italy) to 3.5% (UK). Considering the F1 metric slightly bigger improvements can be observed, which range from around 1% (Italy) up to 5.5% (UK). This relatively modest performance increase can be attributed to the fact that random forest classifiers already incorporate embedded feature selection during the training phase, making explicit feature selection prior to training less impactful [55]. Even when examining other model types, we did not

(a) Valence



(b) Semantic Location



(c) Concurrent Activity

Figure 5: Alone or not alone depending on a) Valence, b) Semantic Location, and c) Concurrent Activity of the participants.

observe notable improvements in feature selection. Additionally, the optimal number of features varied by country but tended to hover slightly above 30. Specifically, the final feature counts used in training were 31 in the UK, 32 in Denmark, 36 in Italy, 30 in Paraguay, 36 in Mongolia, and 90 in the Multi-Country approach. Hence, compared to country-specific models with lower feature counts, feature selection in the multi-country approach retained over 90 features, suggesting that when data from diverse countries are used to train a single model, more features are needed for better representation. It is also important to note that this outcome might not just be due to the number of data points or users in the training data. Countries with a higher number of users (Italy, Mongolia) still resulted in models with fewer features after feature selection, similar to countries with a lower number of data points (UK, Denmark, Paraguay). Therefore, this result is logical and should be expected when using larger behavioral datasets from diverse countries. This underscores the notion that different cultural and social practices in various countries impact the performance of social context inference models, as revealed in the analysis presented in Table 3.

## 5.3 RQ3: Effect of Partial Personalization on Model Performance

Next, our exploration delves into the impact of model personalization on performance. In Figure 8 and Figure 9, we draw comparisons between population-level and hybrid models across individual countries and the multi-country approach. These results are derived after feature selection. First and foremost, our findings underscore the importance of personalization, as it consistently shows improved performance across all countries and the multi-country approach, with performance gains of the range 20-30% in AUC. This aligns with prior research in domains like mood inference and eating behavior prediction, where smartphone sensing-based models have demonstrated enhanced performance following personalization [7, 16, 48, 55]. Thus, it comes as no surprise that we observe a similar trend in the context of social context inference. Hence, the notable performance boost in personalized models highlights the individualistic and variable nature of social context, prompting further exploration into conceptualizing and modeling behavioral diversity within sensing models.

**Table 3: Top ten sensing features to discern alone or not alone social contexts, depending on the ascending value of p-value after Bonferroni correction, from mixed effects models.**

| | feature | coefficient | std. error | z-score | p-value | | feature | coefficient | std. error | z-score | p-value |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Denmark** | app_news_and_magazines | -0.105 | 0.041 | -2.537 | 0.011 | **Paraguay** | app_finance | -0.043 | 0.012 | -3.468 | 0.000 |
| | app_casual | -0.041 | 0.016 | -2.492 | 0.012 | | notifications_w/o_duplicates | -0.018 | 0.005 | -3.248 | 0.001 |
| | app_simulation | -0.180 | 0.074 | -2.416 | 0.015 | | app_productivity | -0.000 | 0.000 | -3.195 | 0.001 |
| | app_action | -0.078 | 0.032 | -0.241 | 0.015 | | location_altitude | -0.000 | 0.000 | -3.175 | 0.001 |
| | app_books_and_reference | -2.651 | 1.099 | -2.412 | 0.015 | | activity_in_vehicle | -0.004 | 0.001 | -2.807 | 0.004 |
| | app_sports | -0.582 | 0.241 | -2.412 | 0.015 | | activity_running | -0.025 | 0.010 | -2.588 | 0.009 |
| | app_adventure | -0.109 | 0.045 | -2.409 | 0.015 | | notifications_removed | -0.003 | 0.001 | -2.577 | 0.009 |
| | proximity_std | -0.027 | 0.011 | -2.375 | 0.017 | | wifi_number_of_devices | -0.007 | 0.003 | -2.466 | 0.013 |
| | app_business | -0.017 | 0.007 | -2.364 | 0.018 | | activity_on_bicycle | -0.016 | 0.006 | -2.405 | 0.016 |
| | app_maps_and_navigation | -0.352 | 0.149 | -2.354 | 0.018 | | wifi_minimum_rssi | -0.006 | 0.002 | -2.172 | 0.029 |
| **UK** | user_presense_time | -0.000 | 0.000 | 1.987 | 0.046 | **Mongolia** | app_video_players_editors | 0.000 | 0.000 | 2.613 | 0.008 |
| | proximity_minimum | -0.008 | 0.004 | -1.975 | 0.048 | | app_shopping | -0.056 | 0.021 | -2.609 | 0.009 |
| | proximity_mean | -0.018 | 0.009 | -1.909 | 0.056 | | app_travel_and_local | -0.056 | 0.022 | -2.500 | 0.012 |
| | screen_number_of_episodes | 0.011 | 0.006 | -1.844 | 0.065 | | app_trivia | -0.674 | 0.284 | -2.373 | 0.017 |
| | app_photography | 0.006 | 0.003 | 1.802 | 0.701 | | app_books_and_reference | -0.016 | 0.007 | -2.312 | 0.020 |
| | steps_counter | 0.000 | 0.000 | 1.775 | 0.075 | | app_educational | -0.109 | 0.047 | -2.300 | 0.021 |
| | proximity_maximum | -0.015 | 0.009 | -1.608 | 0.107 | | app_weather | -3.481 | 1.516 | -2.295 | 0.021 |
| | app_entertainment | 0.020 | 0.012 | 1.588 | 0.112 | | app_comics | -0.379 | 0.165 | -2.291 | 0.021 |
| | activity_on_bicycle | 0.007 | 0.004 | 1.559 | 0.118 | | app_art_and_design | -0.705 | 0.315 | -2.238 | 0.025 |
| | app_business | 0.003 | 0.002 | 1.440 | 0.149 | | app_news_and_magazines | -0.134 | 0.060 | -2.227 | 0.025 |
| **Italy** | activity_in_vehicle | 0.004 | 0.000 | 5.970 | 0.000 | **All** | app_news_and_magazines | -0.105 | 0.041 | -2.537 | 0.011 |
| | activity_onfoot | 0.003 | 0.000 | 4.092 | 0.000 | | app_casual | -0.041 | 0.016 | -2.492 | 0.012 |
| | activity_walking | 0.003 | 0.000 | 4.010 | 0.000 | | app_simulation | -0.180 | 0.074 | -2.416 | 0.015 |
| | steps_counter | 0.000 | 0.000 | 3.243 | 0.001 | | app_action | -0.078 | 0.032 | -2.414 | 0.015 |
| | activity_tilting | 0.000 | 0.000 | 3.225 | 0.001 | | app_books_and_reference | -2.651 | 1.099 | -2.412 | 0.015 |
| | location_altitude | 0.000 | 0.000 | 3.164 | 0.001 | | app_sports | -0.582 | 0.241 | -2.412 | 0.015 |
| | proximity_mean | -0.013 | 0.004 | -3.067 | 0.002 | | app_adeventure | -0.109 | 0.045 | -2.409 | 0.015 |
| | notifications_w/o_duplicates | -0.006 | 0.002 | -2.409 | 0.015 | | proximity_std | -0.027 | 0.011 | -2.375 | 0.017 |
| | wifi_std_rssi | 0.004 | 0.001 | 2.394 | 0.016 | | app_business | -0.017 | 0.007 | -2.364 | 0.018 |
| | screen_number_of_episodes | 0.003 | 0.001 | 2.111 | 0.034 | | app_maps_and_navigation | -0.352 | 0.149 | -2.354 | 0.018 |

**Table 4: Different machine learning models and their inference AUC and F1 metrics with standard deviations in brackets according to different countries.**

| | Logistic (L2) | | Random Forest | | XG Boost | | Ada Boost | |
|---|---|---|---|---|---|---|---|---|
| | $\overline{AUC}$ ($AUC_\sigma$) | $\overline{F1}$ ($F1_\sigma$) | $\overline{AUC}$ ($AUC_\sigma$) | $\overline{F1}$ ($F1_\sigma$) | $\overline{AUC}$ ($AUC_\sigma$) | $\overline{F1}$ ($F1_\sigma$) | $\overline{AUC}$ ($AUC_\sigma$) | $\overline{F1}$ ($F1_\sigma$) |
| Majority Baseline | 50.0 (0.0) | 33.7 (0.1) | 50.0 (0.0) | 33.7 (0.1) | 50.0 (0.0) | 33.7 (0.1) | 50.0 (0.0) | 33.7 (0.1) |
| Random Baseline | 50.1 (0.1) | 50.0 (0.2) | 50.1 (0.1) | 50.0 (0.2) | 50.1 (0.1) | 50.0 (0.2) | 50.1 (0.1) | 50.0 (0.2) |
| Multi-Country | 50.5 (2.8) | 36.6 (9.5) | **67.6** (4.1) | **61.3** (2.8) | 65.8 (2.3) | 60.5 (1.7) | 61.0 (2.8) | 56.1 (3.2) |
| UK | 49.5 (2.7) | 41.6 (7.1) | **63.5** (4.5) | **55.7** (4.6) | 60.8 (5.2) | 55.0 (3.3) | 59.3 (6.5) | 54.9 (5.1) |
| Denmark | 51.1 (1.0) | 32.7 (10.0) | **66.0** (8.7) | 55.0 (9.8) | 60.3 (9.5) | 53.3 (8.8) | 61.2 (8.8) | **56.1** (7.5) |
| Italy | 49.9 (0.5) | 29.3 (2.4) | **66.6** (1.9) | **60.9** (1.2) | 66.1 (2.6) | 60.8 (2.3) | 63.8 (2.8) | 59.2 (2.0) |
| Paraguay | 49.7 (1.4) | 37.1 (5.9) | **72.0** (12.8) | 63.1 (9.4) | 70.9 (11.8) | **64.1** (9.3) | 65.7 (13.4) | 60.7 (10.5) |
| Mongolia | 49.5 (2.0) | 43.5 (1.8) | **73.8** (6.2) | 65.3 (6.1) | 71.9 (5.6) | **66.4** (4.9) | 63.9 (5.3) | 58.4 (3.9) |

In fact, it is worth noting that the weighted average AUC (weighted by the number of data points) of the country-specific models slightly outperforms the population-level multi-country model, achieving 68.7% compared to 67.6%, albeit by a narrow margin. This trend is consistent across hybrid models as well. Nevertheless, it is important to consider that the standard deviations, calculated based on ten iterations of experiments as explained in Section 4, reveal a larger variability in the performance of country-specific models (7.1%) compared to the multi-country model (4.0%). This discrepancy can be partly attributed to the relatively small number of participants in certain countries, such as Denmark and Paraguay. Furthermore, the hybrid models exhibit good performance, surpassing 90% AUC
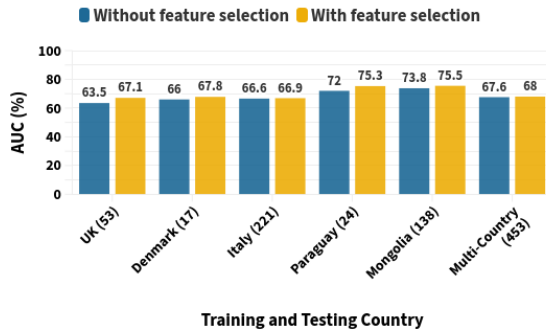
Figure 6: Comparison of random forest model performance for feature selection and without feature selection.The performance metric shown is AUC.
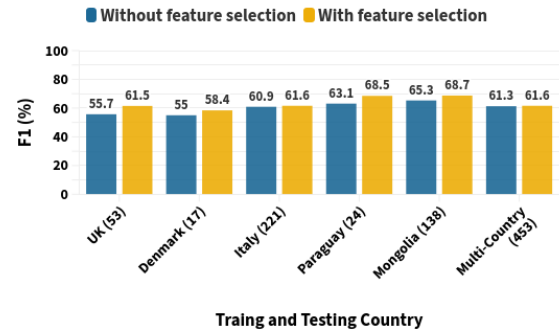


Figure 7: Comparison of random forest model performance for feature selection and without feature selection. The performance metric shown is F1 score.
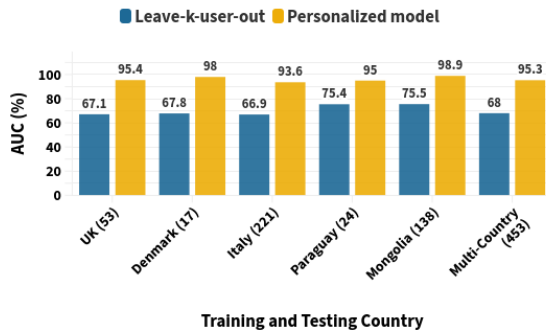


Figure 8: Comparison of random forest model performance for population-level (leave-k-out) and hybrid (personalized) models. The performance metric shown is AUC.
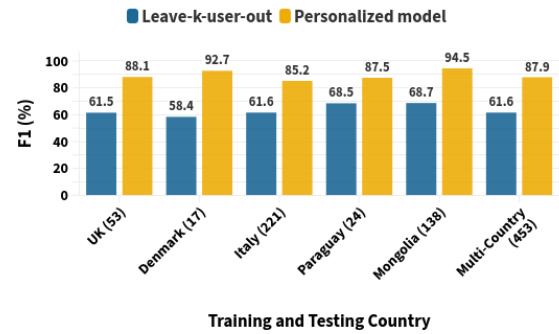


Figure 9: Comparison of random forest model performance for population-level (leave-k-out) and hybrid (personalized) models. The performance metric shown is the F1 score.

across all individual countries and in the multi-country setup. This outcome underscores the effectiveness of model personalization in enhancing the performance of social context inference models. In conclusion, our findings indicate that training a multi-country model yields reasonably good results for social context inference. This result diverges from prior research that explored the performance of country-specific and multi-country models across mood inference, personality inference, and activity recognition [5, 44, 55], where multi-country models tended to under-perform by larger margins. Hence, it is important to highlight that this distinction does not extend to social context inference, where the multi-country approach proves to be effective. On a broader scheme of things, these results also suggest that there is feasibility in building models for social context inference that capture the complex everyday life behavior of individuals well.

The multi-country approach, with its access to a larger data volume, includes diverse country-specific characteristics. This diversity makes it challenging to definitively attribute the differences in inference performance between multi-country and country-specific models to either data quantity or country-specific data heterogeneity. Prior research in personality modeling and mood inference has grappled with the complexity of providing a clear answer in such

scenarios involving multi-country data. Employing under-sampled models, which use an equal amount of data from each country to train country-specific models (often matched to the data volume of the country with the least data), could inadvertently result in reduced representational expressiveness for individual countries [44, 55].

## 5.4 RQ4: Generalization of Models to Unseen Countries

To gain a deeper understanding of how data heterogeneity tied to a particular country aids in detecting the behavior of its participants and the impact of additional data on inference performance, we conducted another experiment. Here, we utilized data from four countries to test it in a fifth, unseen country. This setup emulates a scenario where an application is developed and trained for a specific set of countries and then enters a new market in a different country. The results of this experiment are illustrated in Figure 10. Notably, the AUC of these country-agnostic models consistently falls below those of the country-specific models, although the differences are minor. The magnitude of the AUC variance between country-agnostic and country-specific approaches is minimal in Denmark and the UK ($\Delta$ 3%), the country with the least available

**Table 5: Country-agnostic: Using the countries in the left column, to predict data from the countries on the top row, with AUC, F1 and their tenfold standard deviation. Values along the diagonal are grayed out as they came from the performance obtained in Figure 6, with feature selection under the country-specific approach.**

| Testing / Training | UK | Denmark | Italy | Paraguay | Mongolia |
|---|---|---|---|---|---|
| UK | 67.1 (4.0) / 61.5 (3.2) | 54.2 (10.0) / 50.1 (8.1) | 60.3 (3.1) / 56.7 (1.8) | 61.8 (8.9) / 55.6 (6.9) | 53.8 (3.9) / 42.5 (5.9) |
| Denmark | 51.9 (4.2) / 47.0 (3.8) | 67.8 (9.2) / 58.4 (9.6) | 55.4 (3.3) / 49.0 (2.7) | 56.0 (7.8) / 51.7 (5.8) | 52.2 (6.8) / 50.7 (3.9) |
| Italy | 61.9 (4.3) / 55.8 (1.8) | 60.4 (10.8) / 54.5 (11.8) | 66.9 (2.6) / 61.5 (2.0) | 54.2 (6.6) / 50.8 (7.0) | 57.6 (5.1) / 42.0 (8.5) |
| Paraguay | 58.0 (7.1) / 51.8 (7.5) | 50.1 (11.2) / 43.8 (9.4) | 56.1 (3.3) / 53.1 (2.1) | 75.4 (14.4) / 68.5 (13.0) | 57.4 (5.1) / 48.8 (4.5) |
| Mongolia | 55.5 (6.8) / 31.0 (11.1) | 54.4 (13.9) / 45.4 (8.4) | 52.8 (3.7) / 34.8 (4.5) | 62.8 (9.8) / 47.5 (13.7) | 75.5 (5.4) / 68.7 (6.1) |

data, and most pronounced in Mongolia (Δ 12.7%), a country with high data volume.

Lastly, we explored another experimental setting in which data from one country were used to train a model, which was then tested separately in all the other countries. The results are presented in Table 5. This setup allowed us to investigate the geographical proximity between different countries and how it translates to social context inference models. In the resulting matrix, the rows represent the training countries, while the columns represent the testing country (e.g., using UK data to test in Denmark yields a 54.2% AUC). Interestingly, culturally similar countries appeared to extrapolate better to each other in certain cases. For instance, it can be observed that Italian model worked comparatively well for the UK and Denmark, while it did not perform as well for Paraguay or Mongolia. However, this pattern did not hold to models trained in Denmark and the UK when tested in other European countries. This finding contrasts with prior research on mood inference and complex daily activity recognition in a multi-country setting [5, 55], where the authors found models performing reasonably well when tested in geographically similar European countries. However, similar to those studies and another study [91], we observed a performance degradation, which could be attributed to distributional shifts between countries. Furthermore, Mongolia appeared to be a challenging country to predict without Mongolian data in the training set. If we tested a model in Mongolia using Mongolian data, we could achieve a mean AUC of 75.5%. However, using training data from any other country, the highest AUC achieved was 57.6%, with a model trained on Italian data, marking a difference of approximately 18%. For other countries, the differences ranged between 5% and 15%. This disparity could be attributed to either distributional differences in sensor data across countries or variations in the target attribute, i.e., the social context. In fact, as seen in Figure 3, the distribution of the social context in Mongolia differed from that of all other countries, featuring fewer alone self-reports. Additionally, as shown in Table 3, features with the highest statistical significance for Mongolia were predominantly app usage-based, a pattern that was distinct from other countries except Denmark. Therefore, this could potentially explain the disparity in inference performance in this context.

## 6 DISCUSSION

### 6.1 Summary of Results

In summary, our results shed light on the research questions as follows:
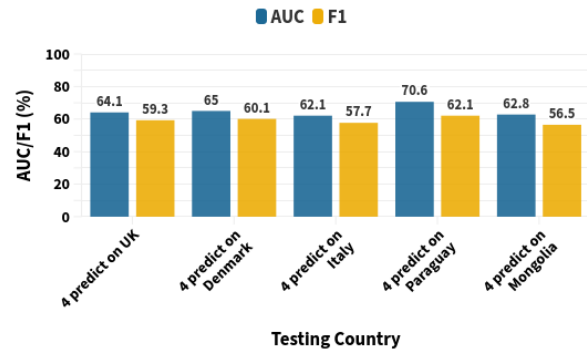


**Figure 10: Country-agnostic model performance with random forest classifiers, for different testing countries where models were trained on the other four countries. AUC and F1 score is shown.**

- **RQ1**: We conducted a comprehensive analysis of a smartphone sensing dataset collected from 483 young adults across five countries. Our analysis identified specific individual features, such as types and quantities of app usage (including tools, communication, productivity, and social), activity types (like being in a vehicle, on foot, walking), location, Bluetooth and WiFi usage, proximity, and screen episodes, as top indicators for distinguishing between being alone and not alone. Moreover, we found that the features significant in differentiating social contexts varied across countries, emphasizing the diversity in behaviors and contextual features associated with various social settings.
- **RQ2**: We operationalized three approaches to assessing the inference of social context (as described in Table 1)—these approaches were country-specific, country-agnostic, and multi-country. Without accounting for the diversity of data sources at the country level, we found that the generic multi-country approach exhibits a moderate AUC of 67.6% without feature selection and 68% with feature selection when using population-level models. The country-specific approach resulted in similar performance to the multi-country approach, with most countries displaying margins of less than 5% for both with and without feature selection. This outcome does not follow prior studies that utilized similar multi-country datasets for health-related inferences, which suggested that country-specific models were generally superior by substantial margins. As a result,

we concluded that this may not always hold true for social context inference.

• **RQ3**: The utilization of hybrid models resulted in the multi-country approach yielding an AUC of 95.3%. Country-specific models performed similarly, with their differences in AUC being in the range of 3%. The results indicate that a generic multi-country model may generally suffice for social context inference, regardless of personalization level. Country-specific models might offer slight improvements in certain countries (i.e., UK, Denmark, and Mongolia). Notably, this deviates from previous research that use smartphone sensor data for health-related inferences, where country-specific models were recommended to achieve better performance, even when personalized. This contrast highlights the differing effectiveness of model types across various domains and objectives.

• **RQ4**: With regards to mood inference and activity recognition, despite initial prior research [5, 55] indicating that models may exhibit reasonable generalization to geographically proximate countries in Europe, our analysis of social context inference performance did not reveal such associations. In fact, in all cases, the models performed poorly when applied to unseen countries. This finding is consistent with prior studies, which concluded that models lack generalization capabilities when deployed to previously unseen countries compared to the performance of country-specific or multi-country approaches. In practice, this means that social context inference models could either be developed in deployment countries or developed with data collected from multiple countries by capturing a broader behavioral diversity, leading to robust models.

## 6.2 Implications of Social Context Inference for Future Applications

This paper discusses the operationalization and achieved performance of a social context inference task in a multi-country setting. The presented work offers theoretical and practical implications for researchers and practitioners in human-computer interaction and ubiquitous computing.

The theoretical implications of the present study are manifold. Firstly, we propose an experimental framework for implementing and comparing social context inference experiments in a multi-country context. The conducted experiments include country-specific, country-agnostic, and multi-country approaches, offering insights into the extent to which country-specific social context information is encoded in multimodal sensing data and whether this information can enhance model performance. In future human sensing research exploring multi-country settings, this experimental framework may be adopted to ensure comparability across studies, facilitating cross-dataset comparisons. Secondly, our findings indicate that country-specific models outperform multi-country models by only a marginal degree. This does not follow previous research on mood inference and activity recognition, which suggested that country-specific models are superior to multi-country models. However, given the complexity of the data collected during the COVID-19 pandemic and other data limitations, the results presented here are not definitive. Consequently, future studies should examine the impact of country diversity on model performance in context-aware sensing tasks, paying attention to cultural practices that may drive differences in social behavior and that may be apparent in the multimodal data. Thirdly, our results demonstrate the importance of personalization for the social context inference task. Hybrid models outperform population-level models by a large margin, suggesting that users exhibit significant differences in social behavior. Consequently, future studies should take note of the significance of collecting sufficient data from individual users to enable personalized models to be implemented.

The practical implications of our study in social context inference extend to various real-world applications. Firstly, mobile health applications that include social context models need to address the initial low inference performance which might lead to inappropriate notifications for new users. One strategy, similar to some mobile health apps that delay providing insights due to lack of personalization [59], is to initially disable the social context feature and enable it only after the model has been adequately personalized for an individual user. Secondly, applications should be designed to monitor social context over an extended period, enabling behavior and health monitoring apps to provide users with long-term insights into their social patterns, such as the amount of time spent alone, empowering users to make informed decisions about their social behavior. This could have implications on their mental well-being. Thirdly, employing social context models that are pre-trained on large datasets from various countries and subsequently personalized with user-provided data can ensure high accuracy in inferring individual social contexts. As a fourth implication, these models have notable potential in smart environments and wearable technology. By understanding the social context, smart devices and environments can adapt functionalities to enhance user experience. For instance, a smart home system could modify its ambiance, such as lighting and music, based on whether the user is alone or hosting guests. Similarly, wearable devices and phones can adjust their notification settings—vibrating for notifications when the user is in a meeting and switching to visual or audible alerts when they are alone. This adaptation not only improves the user's interaction with technology but also respects social norms and privacy preferences. In educational and workplace settings, this technology can be used to tailor communication strategies and content delivery. For instance, learning platforms can adjust content presentation based on the user's social environment, and workplace tools can optimize task allocation and collaboration features to align with the user's current social context. This wide range of implications underscores the extensive potential of social context inference across daily life situations in improving interactions, user experience, health, and well-being.

## 6.3 Cross-Cultural Considerations for Human-Computer Interaction Researchers

Our study also has implications for HCI researchers, particularly in the realm of global applications. It underscores the critical importance of incorporating cross-cultural considerations into the design, development, and deployment of context-aware systems. Understanding that social behaviors and contextual features can significantly differ between countries and cultures emphasizes the need for adaptable and culturally sensitive models. HCI researchers should recognize the wide-ranging variations in individual behaviors, preferences, and social norms across diverse cultural contexts.

Consequently, when developing context-aware applications, conducting extensive cultural research and adapting algorithms and models to local practices and expectations is essential. This not only acknowledges the diversity of behaviors but also respects the ethical norms inherent to each culture. Moreover, achieving a balance between personalization and universality is a central challenge. While personalization is critical for accurate social context inference, there is an increasing need for models and systems that can function across different cultural settings. Creating models that can adapt to varying cultural contexts without compromising user privacy or model performance remains a significant goal. In summary, recognizing the profound impact of cross-cultural diversity on context-aware systems is an imperative consideration for HCI researchers. By embracing this perspective, researchers can craft more culturally sensitive applications that enhance user experiences while respecting country-specific cultural norms.

## 6.4 Ethical and Privacy Considerations of Social Context Inference

This study sheds light on the important ethical and privacy concerns that arise with the implementation of context-aware systems, especially with social context inference. Researchers and practitioners are faced with a complex challenge: how to provide valuable services while also respecting user privacy and autonomy. These findings underscore the importance of informed consent and user control in the development of applications that perform social context inference. Transparency should be a top priority when creating such applications. Users must have a clear understanding of how their data is collected, processed, and used to infer social context. Additionally, they should be granted precise control over when and how this inference takes place. Empowering users to make informed decisions regarding context-aware features, especially those involving social context, is crucial, given that this information not only concerns the individual, but also those in their proximity. Furthermore, privacy-preserving techniques and data anonymization should be at the core of the design of context-aware systems. Even though the social context may not constitute health-related information, it holds great relevance to individuals and is closely linked to various aspects of health. Moreover, researchers should be attuned to the variations in privacy expectations and requirements in different countries, adhering to international privacy standards such as the General Data Protection Regulation (GDPR) and local regulations to ensure ethical and legal compliance. Another critical aspect is the potential for bias and discrimination within context-aware applications. Researchers must be vigilant in identifying and mitigating biases that could arise during social context inference. Regular audits and assessments should be conducted to detect and rectify any discriminatory outcomes associated with the inferred social behavior. Finally, researchers have a pivotal role in advocating for robust ethical guidelines and regulations governing the development and deployment of context-aware systems, including social context. Collaboration with policymakers, privacy advocates, and industry stakeholders is essential to establish ethical frameworks that safeguard user privacy and uphold ethical standards within the field.

## 7 LIMITATIONS AND FUTURE WORK

We conclude by discussing the limitations of and future work resulting from our study. These concern the data generation process, experimental setup, and the computational approaches used. First, a primary limitation of our study is the use of the term "country" to refer to the student populations in universities within those countries, which may not represent the entire country. Although we have been transparent about this from the beginning of the paper, it is important to interpret the results cautiously, keeping this aspect in mind. All data used for this study was generated in a time frame from November to December 2020. This coincides with a major surge of the COVID-19 pandemic in Europe but also in Paraguay and Mongolia. In all countries, social distancing measurements were recommended. Obviously, the COVID pandemic deeply influenced and possibly altered the behaviors, social practices, and moral norms of the pilot study participants. It is, however, highly difficult to assess how those changes influenced the analyzed data. It is quite likely that the pandemic changed the frequency of social contexts (fewer reports being done with others), but it is impossible to say if and how it precisely changed social interactions if they occurred and how this, on the other hand, influenced the mobile sensing data produced. On the other hand, given that many universities and companies are already adhering to remote work/study settings, we expect this trend to continue for the years to come. Hence, despite the study being done during the pandemic, we expect these results to hold in the future to an extent. However, future studies could explore this further.

This study considers only the two-class social context inference task of predicting if someone is alone or not. In principle, however, the used dataset differentiates between 8 different social contexts. The decision to limit this analysis to a two-class inference is mostly caused by class imbalance and page limits. First of all, the alone or not alone situation is well populated in all country data subsets, which is not necessarily given for any other two-class social contexts (e.g., with family, without family would not be well represented in Denmark). Thus, to obtain comparable results across all countries, only this most naive social context was exhaustively investigated. Furthermore, feasibility reasons also tempered further experiments with other social context inference tasks. The performance of the most simple social inference task lies at minimally 66.9% AUC, in the case of, for example, Italy, already quite low; thus, it can be assumed that models would perform even worse on more difficult social inference tasks. Future work could explore other constructs of the social context. Additionally, while we experimented with various time windows, we have only reported results using a ten-minute window for aggregating sensor data. This was done to facilitate easier comparison with other studies that used the same dataset [5, 55]. However, it is important to note that the choice of time window depends on the nature of the dataset and the specific inference task. While the ten-minute window was suitable for our tasks, including social context inference, different time windows could be explored in future work.

The results presented in the personalization section show that social behavior differs greatly from user to user. This indicates that even leave-k-users-out models could be greatly improved if user profiles or specific user groups could be more precisely modeled.

This study tries to model different subgroups of users according to their country of residence and assumes similar cultural and social behaviors of people living in the same country. However, with additional information provided by users regarding other diversity attributes, models could potentially be further improved [75]. This would open the possibility of constructing partially personalized models on a set of diverse information provided by a new user. Future work should try to leverage the information on user diversity to improve mobile sensing models. Finally, it is important to note that although our study aimed to raise awareness about the diversity and generalization issues in modeling human behavior, our sample was primarily composed of young adults from various countries. While our research contributes positively to the broader conversation about generalization, we acknowledge that it also faces its own limitations, particularly regarding the lack of different age groups, occupations, and racial profiles.

## 8 CONCLUSION

In our study, we conducted an in-depth analysis to understand the social context of young adults across five countries using smartphone data. We compared the performance of models trained on data specific to each country against a model trained on data from all countries combined. Interestingly, we found that the country-specific models slightly outperformed the multi-country model, despite having less data. Additionally, our results showed that hybrid models were effective, indicating that a general model using data from various countries can successfully infer social context, with country-specific models offering minor performance improvements in certain cases. These findings have several practical applications. Mobile health apps incorporating social context models can provide valuable information, feedback, and interventions, allowing users to track their social context over time and gain insights into their well-being. In smart homes and wearable devices, social context inference models can adapt functionalities based on whether the user is alone or with others, enhancing the user experience. Our study highlights the potential of using smartphone data to passively infer social context, opening up various application opportunities. Furthermore, we discussed the importance of training these models with country diversity in mind to create robust systems that can adapt to diverse daily life situations across multiple countries.

## ACKNOWLEDGMENTS

## REFERENCES

[1] Saeed Abdullah, Nicholas Lane, and Tanzeem Choudhury. 2012. Towards population scale activity recognition: A framework for handling data diversity. *Proceedings of the AAAI Conference on Artificial Intelligence* 26, 1, 851–857.

[2] Brett Adams, Dinh Phung, and Svetha Venkatesh. 2008. Sensing and using social context. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 5, 2 (2008), 1–27.

[3] Raian Ali, Carlos Solis, Mazeiar Salehie, Inah Omoronyia, Bashar Nuseibeh, and Walid Maalej. 2011. Social sensing: when users become monitors. *Proceedings of the 19th ACM SIGSOFT symposium and the 13th European conference on Foundations of software engineering*, 476–479.

[4] Yasith Amarasinghe, Darshana Sandaruwan, Thilina Madusanka, Indika Perera, and Lakmal Meegahapola. 2023. Multimodal Earable Sensing for Human Energy Expenditure Estimation. *arXiv preprint arXiv:2305.00517* (2023).

[5] Karim Assi, Lakmal Meegahapola, William Droz, Peter Kun, Amalia de Götzen, Chaitanya Nutakki, Shyam Diwakar, Salvador Ruiz Correa, Donglei Song, Hao Xu, Miriam Bidoglia, George Gaskell, Altangerel Chagnaa, Amarsanaa Ganbold, Tsolmon Zundui, Carlo Caprini, Daniele Miorandi, Alethia Hume, Jose Luis Zarza, Luca Cernuzzi, Ivano Bison, Marcelo Rodas Britez, Matteo Busso, Ronald Chenu-Abente, Can Günel, Fausto Giunchiglia, Laura Schelenz, and Daniel Gatica-Perez. 2023. Complex Daily Activities, Country-Level Diversity, and Smartphone Sensing: A Study in Denmark, Italy, Mongolia, Paraguay, and UK. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023). https://doi.org/10.1145/3569483

[6] Sangwon Bae, Denzil Ferreira, Brian Suffoletto, Juan C. Puyana, Ryan Kurtz, Tammy Chung, and Anind K. Dey. 2017. Detecting Drinking Episodes in Young Adults Using Smartphone-Based Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 2, Article 5 (jun 2017), 36 pages. https://doi.org/10.1145/3090051

[7] 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. https://doi.org/10.1109/ACCESS.2022.3179702

[8] Manuela Barreto, Christina Victor, Claudia Hammond, Alice Eccles, Matt T Richins, and Pamela Qualter. 2021. Loneliness around the world: Age, gender, and cultural differences in loneliness. *Personality and Individual Differences* 169 (2021), 110066.

[9] Roy F Baumeister and Mark R Leary. 1995. The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychological bulletin* 117, 3 (1995), 497.

[10] Dror Ben-Zeev, Emily A Scherer, Rui Wang, Haiyi Xie, and Andrew T Campbell. 2015. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric rehabilitation journal* 38, 3 (2015), 218.

[11] Ethan M Berke, Tanzeem Choudhury, Shahid Ali, and Mashfiqui Rabbi. 2011. Objective measurement of sociability and activity: mobile sensing in the community. *The Annals of Family Medicine* 9, 4 (2011), 344–350.

[12] Ananya Bhattacharjee, Dana Kulzhabayeva, Mohi Reza, Harsh Kumar, Eunchae Seong, Xuening Wu, Mohammad Rashidujjaman Rifat, Robert Bowman, Rachel Kornfield, Alex Mariakakis, et al. 2023. Integrating Individual and Social Contexts into Self-Reflection Technologies. *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–6.

[13] Joan-Isaac Biel, Nathalie Martin, David Labbe, and Daniel Gatica-Perez. 2018. Bites 'n'bits: Inferring eating behavior from contextual mobile data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–33.

[14] Sarah-Jayne Blakemore. 2012. Development of the social brain in adolescence. *Journal of the Royal Society of Medicine* 105, 3 (2012), 111–116.

[15] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).

[16] Emma Bouton-Bessac, Lakmal Meegahapola, and Daniel Gatica-Perez. 2022. Your Day in Your Pocket: Complex Activity Recognition from Smartphone Accelerometers. In *International Conference on Pervasive Computing Technologies for Healthcare*. Springer, 247–258.

[17] Douglas Broom. 2021. Home or office? Survey shows opinions about work after COVID-19. *World Economic Forum. July* 21.

[18] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on fairness, accountability and transparency*, 77–91.

[19] Michelle Nicole Burns, Mark Begale, Jennifer Duffecy, Darren Gergle, Chris J Karr, Emily Giangrande, and David C Mohr. 2011. Harnessing context sensing to develop a mobile intervention for depression. *Journal of medical Internet research* 13, 3 (2011), e55.

[20] Yekta Said Can, Bert Arnrich, and Cem Ersoy. 2019. Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. *Journal of biomedical informatics* 92 (2019), 103139.

[21] Youngjae Chang, Akhil Mathur, Anton Isopoussu, Junehwa Song, and Fahim Kawsar. 2020. A systematic study of unsupervised domain adaptation for robust human-activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–30.

[22] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[23] Guanling Chen and David Kotz. 2000. A survey of context-aware mobile computing research. (2000).

[24] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

[25] Hoi Shan Cheung and Elinor Lim. 2021. A systematic review of parenting in Singapore: Insights to the culture-specific functions of styles and practices.

(2021).

[26] Sunny Consolvo, David W McDonald, Tammy Toscos, Mike Y Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, et al. 2008. Activity sensing in the wild: a field trial of ubifit garden. *Proceedings of the SIGCHI conference on human factors in computing systems*, 1797–1806.

[27] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248–255.

[28] Google Developers. 2021. Activity Recognition API. https://developers.google.com/location-context/activity-recognition. Accessed: 2022-02-15.

[29] Alisa Devlic, Roland Reichle, Michal Wagner, Manuele Kirsch Pinheiro, Yves Vanrompay, Yolande Berbers, and Massimo Valla. 2009. Context inference of users' social relationships and distributed policy management. *2009 IEEE international conference on pervasive computing and communications*, 1–8.

[30] Anind K Dey. 2001. Understanding and using context. *Personal and ubiquitous computing* 5, 1 (2001), 4–7.

[31] Anind K Dey, Gregory D Abowd, and Daniel Salber. 2001. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human–Computer Interaction* 16, 2-4 (2001), 97–166.

[32] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

[33] Eiko I Fried, Faidra Papanikolaou, and Sacha Epskamp. 2022. Mental health and social contact during the COVID-19 pandemic: an ecological momentary assessment study. *Clinical Psychological Science* 10, 2 (2022), 340–354.

[34] Richard Fry and Kim Parker. 2021. Rising share of US adults are living without a spouse or partner. (2021).

[35] Andrew J Fuligni and Jacquelynne S Eccles. 1993. Perceived parent-child relationships and early adolescents' orientation toward peers. *Developmental psychology* 29, 4 (1993), 622.

[36] Fausto Giunchiglia, Ivano Bison, Matteo Busso, Ronald Chenu-Abente, Marcelo Rodas, Mattia Zeni, Can Gunel, Giuseppe Veltri, Amalia De Götzen, Peter Kun, et al. 2022. A worldwide diversity pilot on daily routines and social practices (2020-2021). University of Trento Technical Report-DataScientia dataset descriptors.

[37] Zhiqiang Gong, Ping Zhong, and Weidong Hu. 2019. Diversity in machine learning. *IEEE Access* 7 (2019), 64323–64350.

[38] Jack Goody. 1996. Comparing family systems in Europe and Asia: Are there different sets of rules? *Population and Development Review* (1996), 1–20.

[39] John Heron. 1970. The phenomenology of social encounter: The gaze. *Philosophy and Phenomenological Research* 31, 2 (1970), 243–264.

[40] Julianne Holt-Lunstad, Timothy B Smith, Mark Baker, Tyler Harris, and David Stephenson. 2015. Loneliness and social isolation as risk factors for mortality: a meta-analytic review. *Perspectives on psychological science* 10, 2 (2015), 227–237.

[41] Rui Hu, Hieu Pham, Philipp Buluschek, and Daniel Gatica-Perez. 2017. Elderly people living alone: Detecting home visits with ambient and wearable sensing. *Proceedings of the 2nd International Workshop on Multimedia for Personal Health and Health Care*, 85–88.

[42] Nathan Kammoun, Lakmal Meegahapola, and Daniel Gatica-Perez. 2023. Understanding the Social Context of Eating with Multimodal Smartphone Sensing: The Role of Country Diversity. *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI '23)* (2023).

[43] Sayash Kapoor and Arvind Narayanan. 2022. Leakage and the reproducibility crisis in ML-based science. *arXiv preprint arXiv:2207.07048* (2022).

[44] Mohammed Khwaja, Sumer S Vaid, Sara Zannone, Gabriella M Harari, A Aldo Faisal, and Aleksandar Matic. 2019. Modeling personality vs. modeling personalidad: In-the-wild mobile data analysis in five countries suggests cultural impact on personality models. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–24.

[45] Kyungmin Kim, Yen-Pi Cheng, Steven H Zarit, and Karen L Fingerman. 2015. Relationships between adults and parents in Asia. *Successful aging: Asian perspectives* (2015), 101–122.

[46] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. 2017. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages* 2, 3 (2017), 18.

[47] Guanqing Liang and Jiannong Cao. 2015. Social context-aware middleware: A survey. *Pervasive and mobile computing* 17 (2015), 207–219.

[48] Robert LiKamWa, Yunxin Liu, Nicholas D Lane, and Lin Zhong. 2013. Moodscope: Building a mood sensor from smartphone usage patterns. *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, 389–402.

[49] Roderick JA Little and Donald B Rubin. 2019. *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons.

[50] Jeffrey W Lockhart, Tony Pulickal, and Gary M Weiss. 2012. Applications of mobile activity recognition. *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, 1054–1058.

[51] Susan Lund, Anu Madgavkar, James Manyika, Sven Smit, Kweilin Ellingrud, Mary Meaney, and Olivia Robinson. 2021. The future of work after COVID-19. *McKinsey global institute* 18 (2021).

[52] S Mass. 2021. Work from home likely to remain elevated post pandemic. *The Digest* 6 (2021).

[53] Timothy Matthews, Andrea Danese, Jasmin Wertz, Candice L Odgers, Antony Ambler, Terrie E Moffitt, and Louise Arseneault. 2016. Social isolation, loneliness and depression in young adulthood: a behavioural genetic analysis. *Social psychiatry and psychiatric epidemiology* 51, 3 (2016), 339–348.

[54] Julia M Mayer, Starr Roxanne Hiltz, and Quentin Jones. 2015. Making social matching context-aware: Design concepts and open challenges. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 545–554.

[55] Lakmal Meegahapola, William Droz, Peter Kun, Amalia de Götzen, Chaitanya Nutakki, Shyam Diwakar, Salvador Ruiz Correa, Donglei Song, Hao Xu, Miriam Bidoglia, George Gaskell, Altangerel Chagnaa, Amarsanaa Ganbold, Tsolmon Zundui, Carlo Caprini, Daniele Miorandi, Alethia Hume, Jose Luis Zarza, Luca Cernuzzi, Ivano Bison, Marcelo Rodas Britez, Matteo Busso, Ronald Chenu-Abente, Can Günel, Fausto Giunchiglia, Laura Schelenz, and Daniel Gatica-Perez. 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. https://doi.org/10.1145/3569483

[56] Lakmal Meegahapola and Daniel Gatica-Perez. 2020. Smartphone Sensing for the Well-Being of Young Adults: A Review. *IEEE Access* (2020).

[57] 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 (2021), 1–26.

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

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

[60] 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. https://doi.org/10.1145/3448120

[61] 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. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 217–228. https://doi.org/10.1145/3600211.3604688

[62] Subigya Nepal, Weichen Wang, Vlado Vojdanovski, Jeremy F Huckins, Alex Dasilva, Meghan Meyer, and Andrew Campbell. 2022. COVID student study: A year in the life of college students during the COVID-19 pandemic through the lens of mobile phone sensing. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–19.

[63] Esteban Ortiz-Ospina, Charlie Giattino, and Max Roser. 2020. Time use. *Our World in Data* (2020).

[64] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[65] Le Vy Phan, Nick Modersitzki, Kim K Gloystein, and Sandrine Müller. 2022. Mobile Sensing Around the Globe: Considerations for Cross-Cultural Research. https://doi.org/10.31234/osf.io/q8c7y

[66] Ivan Miguel Pires, Nuno M Garcia, Nuno Pombo, and Francisco Flórez-Revuelta. 2016. Identification of activities of daily living using sensors available in off-the-shelf mobile devices: Research and hypothesis. *International Symposium on Ambient Intelligence*, 121–130.

[67] David Porfirio, Allison Sauppé, Aws Albarghouthi, and Bilge Mutlu. 2020. Transforming robot programs based on social context. *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–12.

[68] Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications* 32, 10 (2020), 6363–6381.

[69] Pavel Pudil, Jana Novovičová, and Josef Kittler. 1994. Floating search methods in feature selection. *Pattern recognition letters* 15, 11 (1994), 1119–1125.

[70] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D Lane, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. 2018. Multimodal deep learning for activity and context recognition. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1, 4 (2018), 1–27.

[71] Mika Raento, Antti Oulasvirta, Renaud Petit, and Hannu Toivonen. 2005. ContextPhone: A prototyping platform for context-aware mobile applications. *IEEE*

*pervasive computing* 4, 2 (2005), 51–59.

[72] Sebastian Raschka, Joshua Patterson, and Corey Nolet. 2020. Machine Learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *arXiv preprint arXiv:2002.04803* (2020).

[73] Darshan Santani, Florian Labhart, Sara Landolt, Emmanuel Kuntsche, Daniel Gatica-Perez, et al. 2018. DrinkSense: Characterizing youth drinking behavior using smartphones. *IEEE Transactions on Mobile Computing* 17, 10 (2018), 2279–2292.

[74] Blake Sawyer, Francis Quek, Wai Choong Wong, Mehul Motani, Sharon Lynn Chu Yew Yee, and Manuel Perez-Quinones. 2012. Using physical-social interactions to support information re-finding. 885–910.

[75] 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. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 905–915.

[76] Sandra Servia-Rodríguez, Kiran K Rachuri, Cecilia Mascolo, Peter J Rentfrow, Neal Lathia, and Gillian M Sandstrom. 2017. Mobile sensing at the service of mental well-being: a large-scale longitudinal study. *Proceedings of the 26th International Conference on World Wide Web*, 103–112.

[77] 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. *arXiv preprint arXiv:1711.08536* (2017).

[78] Lewis B Sheiner and Thaddeus H Grasela. 1991. An introduction to mixed effect modeling: concepts, definitions, and justification. *Journal of pharmacokinetics and biopharmaceutics* 19 (1991), S11–S24.

[79] Gürkan Solmaz and Fang-Jing Wu. 2017. Together or alone: Detecting group mobility with wireless fingerprints. *2017 IEEE International Conference on Communications (ICC)*, 1–7.

[80] Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. 2015. Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. *Proceedings of the 13th ACM conference on embedded networked sensor systems*, 127–140.

[81] Nina Tamminen, Tarja Kettunen, Tuija Martelin, Jaakko Reinikainen, and Pia Solin. 2019. Living alone and positive mental health: a systematic review. *Systematic reviews* 8, 1 (2019), 1–8.

[82] Hiromi Taniguchi and Gayle Kaufman. 2021. Family, Collectivism, and Loneliness from a Cross-Country Perspective. *Applied Research in Quality of Life* (2021), 1–27.

[83] Jean M Twenge, Brian H Spitzberg, and W Keith Campbell. 2019. Less in-person social interaction with peers among US adolescents in the 21st century and links to loneliness. *Journal of Social and Personal Relationships* 36, 6 (2019), 1892–1913.

[84] Janne Vanhalst, Bart Soenens, Koen Luyckx, Stijn Van Petegem, Molly S Weeks, and Steven R Asher. 2015. Why do the lonely stay lonely? Chronically lonely adolescents' attributions and emotions in situations of social inclusion and exclusion. *Journal of Personality and Social Psychology* 109, 5 (2015), 932.

[85] Victoria Vickerstaff, Rumana Z Omar, and Gareth Ambler. 2019. Methods to adjust for multiple comparisons in the analysis and sample size calculation of randomised controlled trials with multiple primary outcomes. *BMC medical research methodology* 19, 1 (2019), 1–13.

[86] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. *Proceedings of the international AAAI conference on web and social media* 9, 1, 454–463.

[87] Fabian Wahle, Tobias Kowatsch, Elgar Fleisch, Michael Rufer, Steffi Weidt, et al. 2016. Mobile sensing and support for people with depression: a pilot trial in the wild. *JMIR mHealth and uHealth* 4, 3 (2016), e5960.

[88] Weichen Wang, Shayan Mirjafari, Gabriella Harari, Dror Ben-Zeev, Rachel Brian, Tanzeem Choudhury, Marta Hauser, John Kane, Kizito Masaba, Subigya Nepal, et al. 2020. Social sensing: assessing social functioning of patients living with schizophrenia using mobile phone sensing. *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–15.

[89] Zhiyuan Wang, Maria A Larrazabal, Mark Rucker, Emma R Toner, Katharine E Daniel, Shashwat Kumar, Mehdi Boukhechba, Bethany A Teachman, and Laura E Barnes. 2023. Detecting Social Contexts from Mobile Sensing Indicators in Virtual Interactions with Socially Anxious Individuals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–26.

[90] Molly Stroud Weeks and Steven R Asher. 2012. Loneliness in childhood: Toward the next generation of assessment and research. *Advances in child development and behavior* 42 (2012), 1–39.

[91] 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. https://doi.org/10.1145/3569485

[92] Xuhai Xu, Jennifer Mankoff, and Anind K Dey. 2021. Understanding practices and needs of researchers in human state modeling by passive mobile sensing. *CCF Transactions on Pervasive Computing and Interaction* 3, 4 (2021), 344–366.

[93] Zhixian Yan, Jun Yang, and Emmanuel Munguia Tapia. 2013. Smartphone bluetooth based social sensing. *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*, 95–98.

[94] Özgür Yürür, Chi Harold Liu, Zhengguo Sheng, Victor CM Leung, Wilfrido Moreno, and Kin K Leung. 2014. Context-awareness for mobile sensing: A survey and future directions. *IEEE Communications Surveys & Tutorials* 18, 1 (2014), 68–93.

[95] Chenyue Zhao, Feng Wang, Xudong Zhou, Minmin Jiang, and Therese Hesketh. 2018. Impact of parental migration on psychosocial well-being of children left behind: a qualitative study in rural China. *International journal for equity in health* 17, 1 (2018), 1–10.

[96] Yuchao Zhou, Suparna De, Wei Wang, Ruili Wang, and Klaus Moessner. 2018. Missing data estimation in mobile sensing environments. *IEEE Access* 6 (2018), 69869–69882.

[97] James Zou and Londa Schiebinger. 2018. AI can be sexist and racist—it's time to make it fair.

**Table 6: Summary tables of sensors and extracted sensor features by authors of the paper that conducted feature processing. The sensor features are computed by aggregating ten minutes of sensor data.**

| Sensor | # of Features | Features and Description |
|---|---|---|
| Location (GPS) | 3 | Radius of gyration, the sum of distance covered, altitude |
| Bluetooth (low energy) | 5 | Number of devices connected, mean rssi (received signal strength indicator), max rssi, min rssi, std rssi |
| Bluetooth (Normal) | 5 | Number of devices connected, mean rssi (received signal strength indicator), max rssi, min rssi, std rssi |
| Wifi | 6 | Number of devices connected to the device wifi hotspot, if the device is connected to a wifi, wifi mean rssi (received signal strength indicator), max rssi, min rssi, std rssi |
| Cellular gsm (2G) | 4 | Strength of the mobile signal as defined by mean, max, min, std |
| Cellular wcdma (3G) | 4 | Strength of the mobile signal as defined by mean, max, min, std |
| Cellular LTE (4G) | 4 | Strength of the mobile signal as defined by mean, max, min, std |
| Notifications | 4 | notifications posted, notifications removed, notifications posted without duplicates, notifications removed without duplicates |
| Proximity | 4 | Measures of the proximity sensor as mean, max, min, std |
| Activity | 8 | Activities as classified from the accelerometer data by the Google Activity Recognition API [28]; activity still, activity tilting, activity invehicle, activity onbicycle, activity onfoot, activity walking, activity running, activity unknown |
| Steps | 2 | step counted, steps detected |
| Touch | 1 | Number of touch events |
| Screen | 7 | User presence time, screen number of episodes, screen time total, screen time per episode, screen time max episode, screen time min episode, screen time std episode |
| Apps Usage | 48 | Time spent using app categories. App categories as classified by the google playstore [73]. E.g.: Action apps, dating apps, music apps, puzzle apps, etc. |
| Time | 2 | Hour of the day, day of the week |