Keep Sensors in Check: Disentangling Country-Level Generalization Issues in Mobile Sensor-Based Models with Diversity Scores

Alexandre Nanchen
Idiap Research Institute
Switzerland

Lakmal Meegahapola
Idiap Research Institute &
EPFL
Switzerland

William Droz
Idiap Research Institute
Switzerland

Daniel Gatica-Perez
Idiap Research Institute &
EPFL
Switzerland

ABSTRACT

[1]Machine learning models trained with passive sensor data from mobile devices can be used to perform various inferences pertaining to activity recognition, context awareness, and health and well-being. Prior work has improved inference performance through the use of multimodal sensors (inertial, GPS, proximity, app usage, etc.) or improved machine learning. In this context, a few studies shed light on critical issues relating to the poor cross-country generalization of models due to distributional shifts across countries. However, these studies have largely relied on inference performance as a means of studying generalization issues, failing to investigate whether the root cause of the problem is linked to specific sensor modalities (independent variables) or the target attribute (dependent variable). In this paper, we study this issue in complex activities of daily living (ADL) inference task, involving 12 classes, by using a multimodal, multi-country dataset collected from 689 participants across eight countries. We first show that the 'country of origin' of data is captured by sensors and can be inferred from each modality separately, with an average accuracy of 65%. We then propose two *diversity scores (DS)* that measure how a country differentiates from others w.r.t. sensor modalities or activities. Using these diversity scores, we observed that both individual sensor modalities and activities have the ability to differentiate countries. However, while many activities capture country differences, only the 'App usage' and 'Location' sensors can do so. By dissecting country-level diversity across dependent and independent variables, we provide a framework to better understand model generalization issues across countries and country-level diversity of sensing modalities.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **Empirical studies in ubiquitous and mobile computing**; **Smartphones**; **Mobile phones**; **Mobile devices**; *Empirical studies in collaborative and social computing*.

KEYWORDS

country diversity, data diversity, generalization, country, smartphone sensing, mobile sensing, bias, distributional shift

ACM Reference Format:
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. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '23), August 8–10, 2023, Montréal, QC, Canada.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3600211.3604688

1 INTRODUCTION

Current literature on mobile sensing has demonstrated the utility of multimodal passive sensor data in performing various inference tasks associated with activity recognition, context awareness, and health and well-being [24]. Examples include eating and drinking behavior [22, 25–28], activities of daily living [5, 6], energy expenditure estimation [3], mood [19, 23, 35], stress [20, 33], and depression [7, 10], all of which exhibit reasonable performance when inferred from multimodal sensing data. Even though cross-country generalization is needed for models to be deployed across diverse world regions [31, 34], most prior work has focused on homogeneous populations in one or two countries, hence limiting the understanding of model generalization to other countries [38].

Recent work has emphasized the importance of training models that generalize across multiple countries and thus higher real-world utility[5, 23]. These studies demonstrated that poor generalization across countries could be attributed to distributional shifts in data across countries. However, work on cross-country generalization has largely relied on techniques for downstream inferences, such as mood inference, social context inference, and activity recognition, and compare their performance across countries to understand distributional shifts [5, 16, 23, 38]. For example, for a two-country setting, when a model trained in Country 1 performs poorly in Country 2, studies directly attribute this finding to distributional shifts in data across the two countries. Although this approach is effective, it requires building multiple models to systematically test generalization performance across countries, which can be time-consuming and resource-intensive as the number of experiments

[1]**Alexandre Nanchen, Lakmal Meegahapola, William Droz, Daniel Gatica-Perez, ACM 2023. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record will be published in https://doi.org/10.1145/3600211.3604688.**

AIES '23, August 8–10, 2023, Montréal, QC, Canada

ACM ISBN 979-8-4007-0231-0/23/08…$15.00
https://doi.org/10.1145/3600211.3604688

grows. Furthermore, comparing models is not always straightforward due to differences in performance, attributable to choice of training algorithms, non-optimal parameter tuning, and training set characteristics, such as different numbers of training samples per country. However, even though discussed in general terms (e.g., data distribution-based shift detection and classifier performance-based shift detection [37]), prior work does not examine techniques that allow an understanding of cross-country differences in sensing modalities without relying on classifier performance.

Further, evaluations of model generalization must consider the potential for diversity at the sensor level (independent variables) and target attribute level (dependent variables). For instance, in a three-country setting, accelerometer readings may exhibit similarity between Country 1 and Country 2 but dissimilarity between Country 1 and Country 3, whereas location readings may display similarity between Country 1 and Country 3 but dissimilarity between Country 1 and Country 2. Current inference performance-based techniques do not explicitly address the sensor-level diversity and target attribute diversity across countries (also known as covariate shift and label shift, respectively [37]), which may obscure the understanding of whether shifts affecting poor generalization occur in the sensors or the targets. Moreover, if such shifts occur in the sensors, investigations into which sensor modalities are more likely to be impacted by distributional shifts have yet to be investigated. In this work, we use the terms sensors and sensor modalities, interchangeably.

Studying topics around mobile sensing and generalization is important because poor cross-country generalization of machine learning models could potentially perpetuate societal biases and result in unfair or ineffective systems. For instance, models developed in economically privileged countries might not function as well in less wealthy ones due to different data distributions, which could exacerbate existing global disparities in technology benefits. In this context, despite extensive discussion of these issues in fields such as computer vision, speech, and natural language processing, the challenges of understanding dataset shifts and generalization are relatively unexplored in the domain of mobile sensing [5, 23, 38]. Therefore, this study introduces a low-cost framework to analyze country-level diversity across sensor modalities and target attributes with a large, multi-modal, multi-country dataset from 689 participants across eight countries. We investigate whether sensor modalities can reveal the data's *country of origin* and then distinguish country differences in sensor modalities and the target variable. We suggest two *diversity scores* to measure country differences and analyze country pairs to identify generalization impacting factors. We then apply these scores to study how cross-country data diversity influences inferences of complex activities of daily living (ADL). In line with prior work [5], ADL are activities that punctuate daily routines, are complex in nature, occur over a non-instantaneous time window, and have a semantic meaning around which context-aware applications could be built. In this context, we pose the following research questions:

RQ1: Can the country of origin of data be inferred from each sensing modality independently and in conjunction, to ascertain whether each sensing modality captures country-level information?

RQ2: Can country-level diversity be methodically measured in terms of the capacity to distinguish between countries, using various sensing modalities, to gain a comprehensive understanding of the sensors that influence variations across countries?

RQ3: By considering the inference of ADL as a case study, how can we consider both sensor data (independent variables) and the target attribute (ADL—dependent variable) together to understand country-level diversity across target as shown by sensor data?

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

Contribution 1: We utilized a dataset comprising sensor data collected from 689 college students over a period of four weeks across eight countries, namely, China, Denmark, India, Italy, Mexico, Mongolia, Paraguay, and UK. Our analysis found that each sensing modality can reasonably infer the country of origin of the user, with an accuracy ranging between 0.57 and 0.71 for different sensors and an average accuracy of 0.65. This observation underscores the crucial role of sensor modalities in comprehending cross-country dataset generalization. Furthermore, the collective performance of all sensors in distinguishing countries had an average accuracy of 0.73, with a minimum of 0.59 and a maximum of 0.84, across countries. This finding is intriguing as it suggests that different sensor modalities may capture various aspects of the 'country of origin' and highlights the necessity for further investigation at the sensor modality level to better understand dataset shifts and model generalization issues.

Contribution 2: We present a novel approach to assess country-level diversity by introducing a country-level diversity score (DS1) that incorporates differences in sensor modalities and countries. While this is a simple measure, it provides insight into the distributional disparities of multimodal sensor data across countries. Based on our scoring methodology, we discovered notable variations in countries for certain sensor modalities, with high diversity scores for Italy, Mongolia, and Mexico and low scores for Denmark and Paraguay. These country-level diversity discrepancies are intriguing as they could help to understand generalization, even before training any machine learning models. Specifically, do countries with high country-level diversity across sensor modalities provide better training data in terms of generalization? Are they more challenging as test countries? By examining country pairwise differences (e.g., testing if data captured by the App modality differs significantly for Italy and the UK users), we found that 'App usage' and 'Location' are the two modalities with the highest discriminatory ability between countries. These outcomes suggest that certain sensor modalities might have a more pronounced effect on generalization than others.

Contribution 3: We propose a second country-level diversity score (DS2) that takes into account the country, sensors, and the target attribute (ADL). Under this scoring scheme, we found considerable country differences across activities. When comparing the order of countries in DS1 and DS2, we observed noteworthy differences. For instance, Italy ranks highest in DS1 but falls to fifth in DS2 scoring. Only Paraguay and Denmark maintain the same rank in both orderings. This suggests that a country's distribution of target attributes may differ from others yet remain similar in terms of sensor modalities. When analyzing pairwise country differences

across activities (e.g., examining whether the sensor data of Italy and India's users differ for a given activity), we found that no single activity stands out as a definitive differentiator between pairs of countries, but many activities can serve this purpose. These results imply that a person's 'country of origin' could influence the manner in which activities are practiced (dependent variable), as demonstrated by sensor data (independent variables).

2 BACKGROUND AND RELATED WORK

There is a plethora of research on mobile sensing related to health and well-being. These studies have utilized passive sensing data to infer behavioral, contextual, and psychological aspects of smartphone users. Recent studies have highlighted the challenges of generalization, and several issues remain unresolved. To address this, Adler et al. [2] employed two longitudinal study datasets to infer mental health symptoms and investigate their generalization across publicly available data using inference performance as a measure of generalization. They found that models trained on combined data achieved better inference than models trained on single-study data.

Muller et al. [29] investigated whether patterns in people's mobility behaviors could passively measure depression. They used a U.S.-wide sample that was socio-demographic heterogeneous as well as in mobility patterns, and found that depression inference from GPS-based mobility did not generalize well to large, demographically heterogeneous samples. Meegahapola et al. [23] studied mood inference and found that country-specific approaches performed reasonably well for two or three classes of mood inferences, but country-agnostic models did not generalize well to unseen countries. Assi et al. [5] demonstrated that country-specific models outperformed multi-country models in Human Activity Recognition (HAR) task settings, even when trained on smaller data samples. Khal et al. [16] showed that it is possible to achieve state-of-the-art accuracy in a new country when building personality models (Big 5), and investigated cross-cultural differences in features by constructing multiple country-specific models and comparing the most influential features per country.

Although these studies have highlighted the challenges of generalization in diverse datasets for a given target task, they all consider inference performance as a metric for generalization. Further, most of these studies advocate finding better techniques for model generalization (in case data and labels from target domains are unavailable) and domain adaptation (when target domain labels are available). However, they also acknowledge the challenge of adapting currently available techniques from other domains to multimodal sensing data. Prior work has seldom examined domain adaptation strategies or techniques to understand distributional shifts in mobile sensing data for in multimodal settings [8, 23]. In this work, we analyze country-level diversity directly from sensor data to provide insights into how country differences are distributed between sensor modalities and the target attribute (ADL). Our goal is to contribute to a better understanding of what factors could influence cross-country generalization in multimodal sensor datasets. The findings would allow researchers working on mobile sensing to have a better understanding of distributional shifts when developing future domain adaptation or generalization techniques in multimodal settings.

3 DATASET

We used a dataset originally collected as part of the European WeNet project and described in [13, 23]. The data was gathered from both undergraduate and graduate students in eight countries, namely China, Denmark, India, Italy, Mexico, Mongolia, Paraguay, and the UK, to capture diversity in behaviors across countries. This diversity is decomposed into two dimensions: inherent attributes (observable characteristics such as country of origin, gender, and age) and acquired characteristics[2]. Both of these dimensions of diversity were captured during four weeks in November 2020, via an online questionnaire and a smartphone application called iLog. The app was designed to record software and hardware sensors, as well as some metadata, along with hourly questionnaires assessing the participant's activity and context. Information such as what the students were doing, where they were, with whom, and how they were feeling was collected in time diaries.

The original list of activities included 34 items selected using prior work in human behavior modeling and social practice [12, 39]. As the data collection took place during the Covid-19 pandemic, it significantly influenced the students' way of life. Consequently, some activities, such as traveling and walking, were underrepresented. To address this issue, activities with similar broad semantic meanings were merged, such as 'eating' and 'cooking' and 'social media' and 'internet chatting'. Activities with very disparate semantic meanings, such as 'hobbies', which include dissimilar activities such as 'painting' or 'playing the piano', were filtered out. The resulting dataset consisted of twelve activities of daily living, that modeled the life of a student (Attending class, Eating, Online comm./Social media, Reading, Resting, Shopping, Sleeping, Sport, Studying, Walking, Watching something and Working). In total, the dataset contains 252,393 ADL reports and covers eight countries. More information about the dataset, including the process of narrowing down the ADL to 12, and data collection can be found in [5, 23]. Figure 1 displays the selected activities with their country distribution.

4 FEATURE EXTRACTION PIPELINE

In this section, we explain how we extract features from a sequence of data captured from sensors.

4.1 Obtaining Raw Features

The raw data collected from the mobile app contains a sequence of data captured from hardware and software sensors and time diary metadata. In our study, we decided to keep a traditional feature extraction approach by extracting features by means of functionals applied to a window of features [18, 28]. This approach has the advantage of yielding features that are interpretable.

Practically, feature extraction for the current activity is done as follows: first, we pool software and hardware sensor data in a

[2]We point out that inherent and acquired characteristics are the terms used by ACM as part of its "Commitment to Diversity, Equity, and Inclusion in Computing": https://www.acm.org/diversity-inclusion/about.

Table 1: Summary of features extracted from raw sensor data, aggregated around self-reports using ten-minute windows before and after a time diary entry.

Sensor modality	Count	Corresponding features
Activity	8	time spent doing following activities: still, in_vehicle, on_bicycle, on_foot, running, tilting, walking and unknown (Google Activity Recognition API)
App usage	5	time spent on apps of each category: personalization, social, communication, tools and app not found
Cellular [lte]	4	mean/std/min/max signal strength
Location	3	radius of gyration, sum of distance, altitude mean/min/max, speed mean/min/max
Notifications	4	notifications posted, notifications removed (with and without duplicates)
Proximity	4	mean/std/min/max
Screen events	6	touch events, user presence time, number of episodes, time per episode, min/max/std episode time, total time
Steps	2	steps counter (since turned on), steps detected
WiFi	5	connected indicator, number of devices, mean/std/min/max rssi

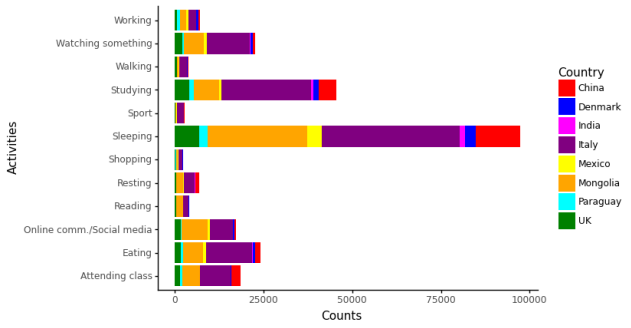

Figure 1: Distribution of countries per activity. The x-axis is the count of practiced activities.

window of 10 minutes centered around the current ADL report, in order to capture the characteristics of the activity; then the window data is discretized by applying various functionals to the continuous stream. Some functionals are specific to the sensor modalities (i.e. radius of gyration), while others are statistical functions like the mean, standard deviation, minimum, and maximum. We also decided to include some high-level features that represent the participant movement, by using the Google Activity Recognition API. This leaves us with nine modalities of sensors describing the ADL report. Table 1 shows a full description of the sensor modalities with their respective features. For more infomation about the feature processing pipeline, please refer to [5, 23]. Note that the 'Location' sensor modality here captures physical location (GPS), and we derive various features from by considering a time window and location traces that quantify the mobility.

4.2 Embedding-Based Representation

The raw extracted features, grouped by sensor modalities, differ in terms of the range of values and are sparse, i.e., they contain many zeros. Motivated by these two observations, we converted each sensor modality raw data into a continuous and dense representation using the fast.ai toolkit [14] tabular data recipe for auto-encoders[3], without the categorical input part. The number of layers for the auto-encoders is chosen empirically to maximize the evaluation set performance.

Regarding the embedding size, it is usally chosen in order to perform well on a downstream task, but in our case, we would like the embeddings to be as generic as possible, i.e., not depending on a specific task. We selected the optimal embedding size empirically, for each sensor modality, by using the elbow method [36] on the evaluation set reconstruction scores (R2 scores [1]). Optimal embedding R2 scores are close to one for all sensor modalities except for the 'Location' sensor modality, which has a score of 0.56. Then, we decided to choose the largest optimal embedding size across sensor modalities as the common embedding size for all sensor modalities (size of 22) to avoid dimensional bias in the statistical and visual analyses and to facilitate their combination. Doing so is appropriate, as R2 scores empirically increase with higher embedding dimensions.

To analyze diversity at the smartphone level, we added the 9 sensor modalities. The resulting embedding is a representation of sensor data. We will use this term in the rest of the paper to refer to this combined representation. An alternative way would have been to concatenate them, but this would have yielded a high-dimensional vector of size 198. We believe that keeping the dimension smaller is beneficial in terms of dimensionality reduction and statistical analysis. The resulting embedding can be thought of as an approximation of embedding modeling all sensor modalities. This technique is widely used in Graph Neural Network message passing [11] and has also been used in past ubicomp literature [17].

5 METHODS

In our experimental setting, for all cases where an inference is performed, the dataset was partitioned in a way that ensures similar country distributions and no overlap of users across the training, validation, and testing sets, similar to prior work that used leave-k-participants-out strategy [5, 23]. Specifically, a test user is unseen during the training phase. The use of this splitting strategy allows

[3]https://walkwithfastai.com/tab.ae

for the exploration of country diversity in the testing set with no bias toward particular users. To assess the generalization of the approach, we utilized a 10-fold cross-validation [32]. In pairwise country diversity analysis, all test embeddings from the ten folds are employed. To implement the aforementioned splitting strategy, we make ten train/validation/testing sets with respective proportions of 80%, 10%, and 10%. First, we perform a 'group stratified split' utilizing the 'StratifiedGroupKFold' class, from the scikit-learn toolkit [30], with K=10. This gives us the ten sets. Then, for each training set, we split it into training and validation using the 'GroupShuffleSplit' class. In both 'group stratified splits', the user ids are serving as the grouping variable.

5.1 Inferring Country of Origin of Sensor Data (RQ1)

The objective of this research question is to determine if the various sensing modalities, individually or collectively, contain country-level information, thereby enabling inference of *country of origin* from data. To achieve this, a one-versus-all binary classification task was set up for each country, and the performance of each sensor modality, separately and combined, was evaluated. The approach involved selecting one country for testing, and replacing labels for all other countries with an 'all' label, resulting in two labels: the country label and the 'all' label (e.g., Italy vs. All, Mongolia vs. All, etc.). To mitigate class imbalance, an equivalent number of samples as the number of samples for the selected country were randomly sampled from the 'all' sample. A binary classifier was trained using a random forest model on the sampled data, and the resulting accuracies averaged across the 10 folds. We used different models, such as multi-layer perceptron neural networks, XGboost, and Support vector machines, for the evaluation. However, we only report results for random forest models for brevity because they performed the best. Mean accuracy per sensor modality was determined by averaging all country accuracies.

5.2 Diversity Score (DS1) Considering Sensor Modalities (RQ2)

This research question aims to quantify country-level diversity based on various sensing modalities. We propose to assess a country's diversity through a country-level diversity score (DS1) that summarizes country differences across sensor modalities. To assess the significance of these differences, we rely on statistical tests. Each country pair for each modality and sensor data is tested. The experiment consisted of a two-group assessment, evaluating the country pairwise difference between the averages of the user embeddings across all activities. To accomplish this, each country pair was tested using a PERMANOVA test in conjunction with a PERMDISP test [4]. The PERMDISP test was necessary to ensure that a significant difference was not due to dispersion. It is important to note that the PERMANOVA tests the null hypothesis that "the centroids and dispersion of the groups as defined by measure space are equivalent for all groups." Failure to do so could result in type I errors, i.e., finding a difference in countries where there are none. This is especially true since our design is unbalanced (i.e., the number of users in each country differs). The scikit-bio framework[4] was utilized to conduct the tests, with 5000 permutations for the PERMANOVA test and 1000 permutations for the PERMDISP test, which tested the 'centroid'. These numbers were chosen empirically, to obtain results with high accuracy, while keeping performance considerations acceptable. A significant threshold of 5% was set for both tests, requiring the PERMANOVA p-value to be <= 0.05 and the PERMDISP test >= 0.05 for a test to be significant. Since both PERMANOVA and PERMDISP tests are permutation tests and have an element of randomness that can impact the results between different runs, especially for p-values close to 0.05, our strategy for almost reproducible results was to perform a series of combined tests (PERMANOVA and PERMDISP) incrementally. Each incremental test in the series contributed to the previous test by adding missing significant values (or nothing), with the testing procedure stopping when ten combined tests did not add new significant values.

Next, we introduce the country-level diversity score (DS1) across sensor modalities. This score is calculated for a given country by considering both country and sensor modality differences. The country count denotes the number of instances in which the given country differs from another country across all modalities, while the sensor modality count indicates the number of unique sensor modalities involved in these differences. By adding both counts we consider diversity originated from country and modality differences and obtain the country-level diversity score (DS1). For instance, according to Table 2, Denmark differs from Mexico only in terms of the 'App usage' and 'Location' sensors, resulting in a diversity score of 3 = 1 (country count) + 2 (modality count). Although this measure is simple, it allows us to gain an understanding of where distributional differences exist in multimodal sensor data across countries. Depending on the research objective, it may be worth considering a different approach to combining both counts that places greater emphasis on one aspect over the other.

5.3 Diversity Score (DS2) Considering Sensor Data and ADL (RQ3)

To assess the diversity of countries across the independent variable, we propose a country-level diversity score (DS2), which summarizes the differences between countries with respect to ADLs.

We employ pairwise statistical tests on sensor data (all modalities) for a specific activity and two countries to identify how countries differ with respect to the target variable. The test provides insights into how countries differ in terms of the target variable, and a deeper analysis at the sensor modality level is left for future work. To obtain reproducible results, we follow the same incremental procedure as in RQ2, and the same number of permutations is used for both tests. The country-level diversity score (DS2) across the target attribute is defined similarly to DS1, but this time across activities. For instance, when examining Denmark, we found that it differs from the UK and India in four unique activities, including Online comm./Social media, Shopping, Studying, and Walking (Table 4). Therefore, Denmark's diversity score is 6 = 2 (country count) + 4 (activity count).

[4] http://scikit-bio.org

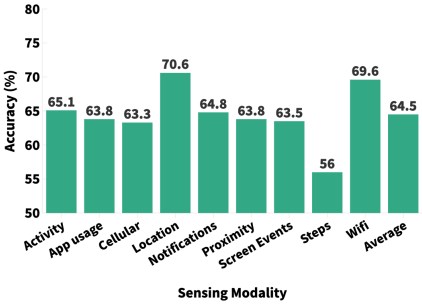

Figure 2: A comparison of the average of the 8 one-country-vs.-all binary classification accuracies, by individual sensor modalities. Random accuracy is 50%.

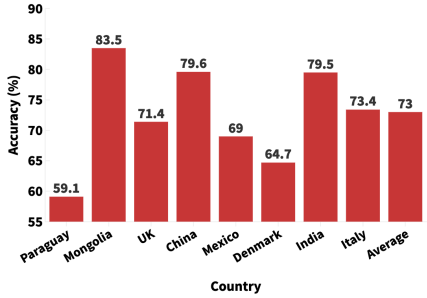

Figure 3: A comparison of the eight one-country-vs.-all binary classification accuracies with sensor data (all sensor modalities). Random accuracy is 50%.

6 RESULTS

6.1 Inferring Country of Origin of Sensor Data (RQ1)

In this section, we aim to examine whether each sensing modality contains country-level information, and to determine the degree of accuracy gained by combining all sensor modalities.

The results of the analysis are presented in Figure 2, which provides a breakdown of the accuracy levels achieved by each sensor modality. The average accuracy attained in inferring the *country of origin* from a single sensor modality is 64.5%, with two modalities performing well above the average. However, the 'Steps' sensor modality falls short of this mark. The 'Location' modality, on the other hand, exhibits the best performance with an average accuracy of 70.6%. It is worth noting that the standard deviation values are moderate (less than 10% across all sensing modalities), indicating a diversity of smartphone usage patterns across countries. These findings underscore the importance of analyzing each sensor modality separately to account for inference biases. Figure 3 presents the results obtained from combining all sensor modalities. On average, it is possible to infer a country from smartphone sensor data with an accuracy of 73.0%. The country with the highest inferred accuracy is Mongolia, with an accuracy rate of 83.5%, while Paraguay has the lowest inferred accuracy rate of 59.1%. The results suggest that sensor modalities capture complementary country-level information, thereby boosting the overall accuracy of smartphone sensor data in identifying the *country of origin*.

Hence, in summary, regarding the first research question (RQ1), our analysis has revealed that, on average, sensor modalities allow for the inference of the *country of origin* of sensor data with an accuracy of 64.5%. Furthermore, we have observed that when combining sensor modalities, there is a relative gain in performance of 13.2% compared to the average accuracy of individual modalities. Our results show that on average, a country can be inferred from smartphone sensor data with an accuracy of 73.0%. Hence, these results show that sensor data contains country-level information. This again provides us a motivation into disentangling country-level distributional shifts across different sensing modalities, rather than just relying on inference performance of a target variable.

6.2 Diversity Score (DS1) Considering Sensor Modalities (RQ2)

In this section, our objective is to quantify the diversity across countries at the sensor modality level, with the aim of gaining insights into the sensors that contribute to country differences. Specifically, our analysis seeks to achieve two goals: 1) to identify statistically significant pairwise differences between countries for each sensor modality; 2) to rank countries based on a country-level diversity score (Diversity Score 1—DS1) that combines both country and sensor modality differences. It is important to note that no distinction is made between activities in this analysis, as we will explore the influence of activities in the next research question.

Table 2 displays the significant pairwise differences between countries based on sensor modalities as part of our first goal. Please note that as a result of this choice, not all country pairs appear in the Table. The PERMANOVA F statistic is shown as an effect size indicator if the PERMANOVA p-value < 0.05 (statistically significant) and left empty if the PERMDISP p-value > 0.05 (not statistically significant). We have omitted the 'Activity' and 'Screen events' sensor modalities as no significant differences were found among countries regarding these sensors. Out of the 56 possible country pairwise comparisons (e.g., Italy vs. India, Italy vs. Mongolia, etc.), 17 showed significant differences (as shown in the first column of Table 2). Our analysis revealed that sensor modalities do not capture an equal number of country differences. Specifically, 'App usage' exhibited the highest number of differences (13), followed by 'Location' (6). On the other hand, the 'Cellular' sensor modality only captured one country difference, and 'Activity' (here we do not refer to the ADL, our dependent variable, but to the simple activity captured using the Google activity recognition API, which is an independent variable used to infer ADL) and 'Screen events' did not capture any, hence not shown on the table. We further observed that country differences can be attributed to sensor modality differences. For instance, Mongolia and Paraguay differ in their readings from the 'App usage', 'Proximity', and 'Wifi' sensors. Generally, pairwise country differences are explained by 1-3 sensor modalities (out of 9 possible), but the sensor modalities that contribute to such differences vary for specific country pairs. Therefore, we can conclude that while differences between the two countries are limited, there is a large diversity of country differences when considering

Table 2: Statistically significant differences in smartphone usage by users of different countries per sensor modality. Tests are performed on individual sensor modality embeddings. The PERMANOVA F statistic is shown if PERMANOVA p-value < 0.05 (statistically significant) and left empty if PERMDISP p-value > 0.05 (not statistically significant).

	App usage	Cellular	Location	Notifications	Proximity	Wifi	Steps
China-Mexico	3.2						
Denmark-Mexico	2.7		2.1				
Italy-China				2.5			
Italy-India		2.9					2.9
Italy-Mexico	3.4		1.9				
Italy-Mongolia	2.4					4.0	
Italy-UK	1.8		2.9				
Mexico-India	2.3						2.2
Mongolia-China	1.9						
Mongolia-India				2.0		2.7	
Mongolia-Mexico	3.5						
Mongolia-Paraguay	1.9				2.4	2.0	
Mongolia-UK					1.9		
Paraguay-Mexico	2.8		2.1				
Paraguay-UK	2.2						
UK-China	2.2		3.2				
UK-Mexico	2.4		4.4				

Table 3: Country-level diversity across sensor modalities. Country count corresponds to the number of pairwise differences for the given country. The sensor modality count is the number of unique sensor modalities involved in these pairwise differences.

Country	Diversity Score (DS1)	Country count	Sensor modality count	Involved sensor modalities
Italy	11	5	6	App usage, Cellular, Location, Notifications, Steps, Wifi
Mongolia	10	6	4	App usage, Notifications, Proximity, Wifi
Mexico	10	7	3	App usage, Human Location, Steps
India	8	3	5	App usage, Cellular, Notifications, Steps, Wifi
UK	8	5	3	App usage, Location, Proximity
China	7	4	3	App usage, Location, Notifications
Paraguay	7	3	4	App usage, Location, Proximity, Wifi
Denmark	3	1	2	App usage, Location

all countries. We also noted that most of the differences involved countries from different continents, except for 'Italy-UK', which exhibited differences in 'Location' and 'App usage'. This finding is in agreement with the previous work of Meegahapola et al. [23], which reported that on other inference tasks using the same dataset, European countries performed better for other European countries than for non-European ones.

Table 3 presents the country ordering based on the DS1. Italy holds the highest score, followed by Mongolia and Mexico. Although Italy and Mexico have almost the same DS1, Italy is distinct in terms of its sensor modalities (i.e., sensor modality count: 6) rather than its country differences (i.e., country count: 5). This observation is interesting because it implies that country-level diversity of Mexico may be caused by only a few sensor modalities (i.e., sensor modality count: 3). On the other hand, Denmark has the lowest score with 1 country and 2 sensor modalities differences. It is noteworthy that while differences often emerge between two continents, this does not hold true for country-level diversity across sensor modalities.

In summary, this research question provides insights into RQ2 by proposing a country-level diversity score that considers both country and sensor modality differences. Our findings show that country-level diversity across sensor modalities significantly varies across different countries. Moreover, we observe that the 'App usage' captures the highest country diversity, followed by the 'Location' sensor. Additionally, we note that pairwise country differences can be explained by a maximum of 1-3 sensor modalities. For example, as mentioned in Table 2, Italy-China have statistically significant differences in terms of Notifications (1 modality); Mongolia-Paraguay have statistically significant differences in terms of App usage, Proximity, and Wifi (3 modalities), etc. Therefore, our results indicate that a few specific sensor modalities play a crucial role in capturing country differences.

6.3 Diversity Score (DS2) Considering Sensor Data and ADL (RQ3)

In this section, we undertake an analysis that takes into account both sensor data and the target attribute to gain a better understanding of country-level diversity across the classes of the target variable as represented by the sensor data. For this, we specifically used ADL Recognition, where target attributes contain 12 activity classes (see Section 3 for the list of activities). Our analysis has two goals: 1) to analyze country pairwise differences across ADL statistically and 2) to rank countries using a country-level diversity

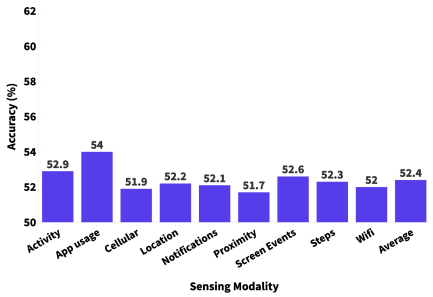

Figure 4: A comparison of the average of the 12 one-activity-vs.-all binary classification accuracies, by individual sensor modalities.

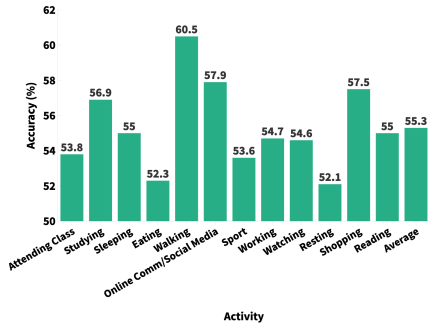

Figure 5: A comparison of one-activity-vs.-all binary classification accuracies with sensor data (all sensor modalities). Random accuracy is 50%.

score (Diversity Score 2—DS2) that considers both country and activity differences.

First, we perform an analysis to investigate the extent to which sensor modalities can be used to infer each activity. We follow the same procedure as in RQ1, but this time, we focus on activities instead of countries. Figure 4 shows that the practiced activity can be inferred from each sensing modality separately with an average activity accuracy of 52.4%. Furthermore, Figure 5 demonstrates that combining sensor modalities is beneficial when inferring ADL, indicating that sensor modalities are complementary. The average activity accuracy improves to 55.3% when all sensor modalities are combined. These results reveal that differentiating a practiced activity from the other 11 ADLs using sensor data is a challenging task, as compared to a random accuracy of 50%. Prior studies have also shown that this is a challenging task, and personalization is required to attain better performance [5].

Similar to RQ2, we conducted a test to determine significant differences across countries, this time in relation to different ADL (goal 1). The results are presented in Table 4. The 'Sport' and 'Working' activities were excluded from the tables as no significant pairwise differences were found between countries. We also discarded 4 significant differences involving one country with a sample size (number of unique students) less than 15, assuming that this case might not contain enough variability in the data that describes the activity. Out of 56 possible pairwise comparisons, 18 showed significant differences (as shown in the first column of Table 4). Similar to the findings from the analysis of sensor modalities, it was observed that different activities captured varying numbers of pairwise differences between countries. Additionally, it was noted that a pairwise difference between countries could be broken down into differences in specific activities. Consequently, two countries could exhibit differences in one activity while showing similarities in another activity, as indicated by the sensor data (see Appendix A). Unlike the analysis of sensor modalities, there was no specific activity that stood out for its diversity across countries. However, more than half of the activities had at least five pairwise differences between countries. Furthermore, a single pairwise difference between countries could be broken down into as many as seven activities out of the 12 possible, which is greater than the number observed for modalities (1-3 out of 9 possible). Taken together, these findings suggest that different countries may exhibit variations in multiple ways while engaging in a particular activity. This could pose a challenge to the generalization of ADL inference models and may explain why country-specific models perform better in prior work [5].

In response to RQ3, we have shown the relevance of taking into account both sensor data and target attribute (ADL) when assessing country diversity in mobile sensing datasets. Our proposed DS2 for countries revealed that significant differences exist between countries in terms of activity diversity, and that these differences are distinct from those observed in DS1, which captures sensor modality diversity. We also noted that no activity particularly captured country diversity, but many exhibited a substantial number of country pairwise differences (>= 5). Lastly, we found that up to seven activities could account for significant country pairwise differences. In summary, our results highlight the importance of considering both target ADL and sensor data in evaluating country diversity, as they provide complementary perspectives on the issue.

7 DISCUSSION

7.1 Summary of Results

In this study, we examined the country-level diversity of a multimodal, multi-country dataset collected from 689 participants across eight countries in the context of a 12-class ADL inference task. Our investigation aimed to disentangle the influence of sensor modalities and the target attribute on cross-country generalization.

7.1.1 RQ1. We demonstrated that individual sensor modalities could somewhat infer the *country of origin* of users and are complementary, indicating that sensors can capture significant country-level information, enabling country-level comparisons. However, the ADL inference from sensor data proved to be more challenging. Overall, we provided motivation as to why sensor-level analysis is needed for to understand cross-country model generalization issues.

7.1.2 RQ2. Further analysis was conducted to assess the effectiveness of different sensor modalities in capturing country differences. Our findings indicate that the 'App usage' and 'Location' modalities were particularly effective in this regard. This highlights the importance of understanding country differences in these modalities

Table 4: Statistically significant differences in sensing features by ADL by users of different countries. PERMANOVA F statistic is shown if PERMANOVA p-value < 0.05 (statistically significant) and left empty if PERMDISP p-value > 0.05 (not statistically significant).

	Attend class	Studying	Sleeping	Eating	Online com./ Social media	Watching some-thing	Resting	Shopping	Reading	Walking
China-India			2.7		2.4					
China-Mexico							2.0			
Italy-China	5.8	2.0	2.7	7.3	2.3	4.7				
Italy-India		3.9	5.9							
Italy-UK		2.0			2.4					2.0
Mexico-India			2.0		3.1					
Mongolia-China				2.6						
Mongolia-India			3.0		4.2					
Mongolia-Paraguay			1.9				2.8			
Mongolia-UK			3.2	4.5			1.9	2.1	2.0	
Mongolia-Mexico		2.6								
Paraguay-China	2.4			2.8			2.3			
Paraguay-India		2.8	4.0		4.2					
UK-China	5.7	3.5	2.7	6.6	5.2	5.9	2.1			
UK-India	1.9	4.5	4.9	3.5						
UK-Mexico						2.3				
Denmark-India					3.2					
Denmark-UK		2.0						1.9		

Table 5: Country-level diversity across activities. The country count corresponds to the number of pairwise differences for the given country. The activities count is the number of unique activities involved in these pairwise differences.

Country	Diversity Score (DS2)	Country count	Activities count	Involved activities
UK	15	6	9	Eating, Online comm./Social media, Reading, Resting, Shopping, Sleeping, Studying, Walking, Watching something
China	13	6	7	Attending class, Eating, Online comm./Social media, Resting, Sleeping, Studying, Watching some-thing
India	12	7	5	Attending class, Eating, Online comm./Social media, Sleeping, Studying
Mongolia	12	5	7	Eating, Online comm./Social media, Reading, Resting, Shopping, Sleeping, Studying
Italy	10	3	7	Attending class, Eating, Online comm./Social media, Sleeping, Studying, Walking, Watching some-thing
Mexico	9	4	5	Online comm./Social media, Resting, Sleeping, Studying, Watching something
Paraguay	9	3	6	Attending class, Eating, Online comm./Social media, Resting, Sleeping, Studying
Denmark	5	2	3	Online comm./Social media, Shopping, Studying

for achieving better cross-country generalization. Interestingly, we found that the two countries differ, at most, by only three sensor modalities, but the specific sensors varied across different country pairs. This suggests that country differences are captured by only a few sensors and that investigating the content of these sensors could provide a better understanding of the factors that make countries distinct. Additionally, the country-level diversity scores for sensor modalities (DS1) revealed that countries such as Italy and Denmark differ greatly in terms of diversity. Specifically, Italy exhibits a high degree of diversity with respect to both sensor modalities and country differences, while Denmark does not. Further analysis of the impact of these differences could aid in understanding the challenges of cross-country generalization.

7.1.3 RQ3. Furthermore, our analysis revealed that a large number of activities exhibited numerous pairwise country differences, suggesting that there might be important variations in how users in different countries carry out daily activities, as shown by all sensors. Specifically, we found that two countries could differ in as many as seven activities, further highlighting the challenges in cross-country ADL inference. Moreover, the country-level diversity score for activities highlighted the existence of significant diversity among countries, with highly diverse countries such as the UK exhibiting a diversity score (DS2) of 16, while less diverse countries such as Denmark had a diversity score of 6. This gap in diversity scores is an important factor to consider when developing cross-country models and merits further investigation.

In summary, our study highlights the importance of considering both sensor modalities and target attributes when assessing country diversity in mobile sensing datasets. We have provided evidence of differences in country diversity across sensor modalities and activities, which have implications for the cross-country generalization of models.

7.2 Implications, Limitations, and Future Work

Our findings suggest potential implications for future research to deepen the understanding of the relationship between country-level diversity and performance/generalization. Can the proposed Diversity Scores be used as a proxy for generalization in cross-country datasets? Firstly, it would be valuable to investigate whether there is a correlation between the ability of sensor modalities to capture country differences and the performance of models trained on them. For instance, one could study if a model, when trained on modalities that capture a high number of country differences, generalizes better. Secondly, our observation that the difference between two countries can be explained by a limited number of sensor modalities raises questions about whether accuracy differences between test countries are primarily due to the modalities where the countries differ or to other factors. In terms of practical implications, utilizing the proposed diversity scores to design experiments could facilitate a better understanding of how different countries generalize. For example, one could investigate whether training with countries that exhibit high country-level diversity scores (DS1) outperforms training with countries that exhibit low country-level diversity scores in terms of performance and generalization. Additionally, examining the impact of the country-level diversity of a test set on performance and generalization across sensor modalities and activities could provide insights into how to design more robust mobile sensing models. Finally, it may be worthwhile to explore the potential benefits of adding a diverse country to an existing dataset to improve performance and generalization.

This study has several limitations that should be taken into account. The first point to consider pertains to the dataset. The data was collected in the Fall of 2020 during the COVID-19 pandemic, a time when participants, who were university students from eight different countries, spent a significant amount of time at home. Therefore, we should not assume that this cohort represents the entire university student population of these countries. It is important to consider these aspects when interpreting the results. Additionally, regarding sample sizes, the number of unique students per country varied from 20 (Mexico) to 240 (Italy), with a median of 41 unique students. Although these numbers are statistically sufficient for testing, larger sample sizes are necessary to draw more robust conclusions at scale. Secondly, the proposed country-level diversity score (DS2) across ADL relies on tests that evaluate country differences in sensor data for combined sensors. Although this provides insights of the relationship between country differences and ADL at the smartphone level, it would be interesting to explore further how country differences relate to activities for each individual modality. However, this analysis was not provided because the applicability of the method on all sensor modalities needed to be tested first. These analyses could help disentangle the relationship between countries, sensor modalities, and target attributes. Thirdly, statistical tests were used on the user embeddings of country pairs to assess country differences. By examining the embeddings (see Appendix A), it was observed that assessing two-country differences is not always straightforward. To facilitate the tests, it may be worth exploring techniques that increase data separation prior to applying statistical tests, such as applying Linear Discriminant Analysis (LDA) [15] on embeddings prior to testing. Fourthly, this study focused solely on a specific target attribute (ADL). It would be interesting to investigate whether other target attributes, such as social context and mood, produce the same country-level diversity scores (DS2) as ADL. Exploring different target attributes could provide additional insights into country distributional shifts understanding. As a fifth point, this work focused on investigating the differences captured by sensor modalities and sensor data for the country of origin. It could be worthwhile to investigate how different states or regions within a country differ. Additionally, the proposed methodology could be extended to inherent diversity attributes like gender and age to investigate their impact on sensor modalities, the target attribute, and generalization. Understanding how differences exist across users, how they are captured by sensor modalities, and how they can potentially influence generalization is particularly important for the health and well-being related mobile sensing-related applications. Finally, for the country-level diversity scores, the choice was made to add the count of pairwise country differences and individual sensor modalities/target attribute differences. Future work could explore other ways of computing these scores (e.g., a weighted average) that are more appropriate for understanding generalization issues, depending on the requirement.

Our work adds to the important topic of generalization across countries, which has been studied in images [9] and text [21], but less on mobile datasets. Our work also contributes analysis of a multi-country dataset that includes both Global North and South countries with the goal of designing for all of them while taking into account their specificities.

8 CONCLUSION

Our study, which focuses on ADL inference, utilized a large-scale, multimodal, multi-country dataset to investigate country-level diversity across sensor modalities and activities, with the aim of disentangling both in order to gain insights on how to achieve better generalization in cross-country datasets. By proposing two country-level diversity scores for sensor modalities and activities, we identified statistically significant differences between countries that can be explained by specific sensor modalities and ADL. Our results indicate that Italy has the highest country-level diversity across sensor modalities, the UK has the highest across activities, and Denmark has the lowest for both country-level diversities. However, we observe that these diversity scores do not seem to correlate, except for Paraguay and Denmark, which have the same score. In terms of country pairwise differences, our analysis shows that the 'App usage' and 'Location' sensors have the highest ability to distinguish between countries. On the other hand, we found that no single activity stands out in terms of the ability to distinguish between countries, but many activities have a high ability to do so. Finally, we discovered that country pairwise differences could be explained by only 1-3 sensor modalities and 1-7 activities, which indicates that cross-country differences between two countries may be captured by only a few sensors but many activities. As discussed, our work opens several research directions towards diversity-aware mobile sensing systems.

ACKNOWLEDGMENTS

This work was funded by the European Union's Horizon 2020 WeNet project, under grant agreement 823783. We thank all the WeNet volunteers and local research teams, who collectively produced the datasets we used.

REFERENCES

[1] 2008. *Coefficient of Determination.* Springer New York, New York, NY, 88–91. https://doi.org/10.1007/978-0-387-32833-1_62

[2] Daniel A. Adler, Fei Wang, David C. Mohr, and Tanzeem Choudhury. 2022. Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies. *PLOS ONE* 17, 4 (04 2022), 1–20. https://doi.org/10.1371/journal.pone.0266516

[3] 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).

[4] Marti J. Anderson and Daniel C. I. Walsh. 2013. PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? *Ecological Monographs* 83, 4 (2013), 557–574. https://doi.org/10.1890/12-2010.1 arXiv:https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/12-2010.1

[5] Karim Assi, Lakmal Meegahapola, William Droz, Peter Kun, Amalia de Gotzen, Miriam Bidoglia, Sally Stares, George Gaskell, Altangerel Chagnaa, Amarsanaa Ganbold, et al. 2023. Complex Daily Activities, Country-Level Diversity, and Smartphone Sensing: A Study in Denmark, Italy, Mongolia, Paraguay, and UK. *arXiv preprint arXiv:2302.08591* (2023).

[6] Emma Bouton-Bessac, Lakmal Meegahapola, and Daniel Gatica-Perez. 2022. Your Day in Your Pocket: Complex Activity Recognition from Smartphone Accelerometers. In *Proceedings of the 16th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth).*

[7] Luca Canzian and Mirco Musolesi. 2015. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 1293–1304.

[8] Youngjae Chang, Akhil Mathur, Anton Isopoussu, Junehwa Song, and Fahim Kawsar. 2020. A Systematic Study of Unsupervised Domain Adaptation for Robust Human-Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 39 (mar 2020), 30 pages. https://doi.org/10.1145/3380985

[9] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. 2019. Does object recognition work for everyone?. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 52–59.

[10] Asma Ahmad Farhan, Chaoqun Yue, Reynaldo Morillo, Shweta Ware, Jin Lu, Jinbo Bi, Jayesh Kamath, Alexander Russell, Athanasios Bamis, and Bing Wang. 2016. Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data. In *2016 IEEE Wireless Health (WH)*. IEEE, 1–8.

[11] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *International conference on machine learning*. PMLR, 1263–1272.

[12] Fausto Giunchiglia, Enrico Bignotti, and Mattia Zeni. 2017. Personal context modelling and annotation. In *2017 IEEE international conference on pervasive computing and communications workshops (PerCom workshops)*. IEEE, 117–122.

[13] Fausto Giunchiglia, Ivano Bison, Matteo Busso, Ronald Chenu, Marcelo Rodas, Mattia Zeni, Can Gunel, Giuseppe Veltri, Amalia de Götzen, Peter Kun, Amarsanaa Ganbold, George Gaskell, Sally Stares, Miriam Bidoglia, Alethia Hume, and Jose Luis Zarza. 2020. A worldwide diversity pilot on daily routines and social practices (2020). (2020), 26. https://iris.unitn.it/retrieve/handle/11572/303769/446832/2021-Datascientia-LivePeople-WeNet2020.pdf

[14] Jeremy Howard and Sylvain Gugger. 2020. Fastai: a layered API for deep learning. *Information* 11, 2 (2020), 108.

[15] Freda Kemp. 2003. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. *Journal of the Royal Statistical Society: Series D (The Statistician)* 52, 4 (2003), 691–691. https://doi.org/10.1046/j.1467-9884.2003.t01-2-00383_4.x arXiv:https://rss.onlinelibrary.wiley.com/doi/pdf/10.1046/j.1467-9884.2003.t01-2-00383_4.x

[16] 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.

[17] Boning Li and Akane Sano. 2020. Extraction and interpretation of deep autoencoder-based temporal features from wearables for forecasting personalized mood, health, and stress. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–26.

[18] Frédéric Li, Kimiaki Shirahama, Muhammad Adeel Nisar, Lukas Köping, and Marcin Grzegorzek. 2018. Comparison of feature learning methods for human activity recognition using wearable sensors. *Sensors* 18, 2 (2018), 679.

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

[20] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T Chittaranjan, Andrew T Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. 2012. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM conference on ubiquitous computing*. 351–360.

[21] Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264* (2020).

[22] Lakmal Meegahapola, Wageesha Bangamuarachchi, Anju Chamantha, Salvador Ruiz-Correa, Indika Perera, and Daniel Gatica-Perez. 2022. Sensing eating events in context: A smartphone-only approach. *IEEE Access* 10, ARTICLE (2022).

[23] 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

[24] Lakmal Meegahapola and Daniel Gatica-Perez. 2021. Smartphone Sensing for the Well-Being of Young Adults: A Review. *IEEE Access* 9 (2021), 3374–3399. https://doi.org/10.1109/ACCESS.2020.3045935

[25] Lakmal Meegahapola, Florian Labhart, Thanh-Trung Phan, and Daniel Gatica-Perez. 2021. Examining the social context of alcohol drinking in young adults with smartphone sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–26.

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

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

[28] Lakmal Meegahapola, Salvador Ruiz-Correa, Viridiana del Carmen Robledo-Valero, Emilio Ernesto Hernandez-Huerfano, Leonardo Alvarez-Rivera, Ronald Chenu-Abente, and Daniel Gatica-Perez. 2021. One More Bite? Inferring Food Consumption Level of College Students Using Smartphone Sensing and Self-Reports. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (March 2021), 1–28. https://doi.org/10.1145/3448120

[29] Sandrine R. Müller, Xi (Leslie) Chen, Heinrich Peters, Augustin Chaintreau, and Sandra C. Matz. 2021. Depression predictions from GPS-based mobility do not generalize well to large demographically heterogeneous samples. *Scientific Reports* 11, 1 (July 2021), 14007. https://doi.org/10.1038/s41598-021-93087-x

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[31] Le Vy Phan, Nick Modersitzki, Kim K Gloystein, and Sandrine Müller. 2022. Mobile Sensing Around the Globe: Considerations for Cross-Cultural Research. psyarxiv.com/q8c7y

[32] Payam Refaeilzadeh, Lei Tang, and Huan Liu. 2009. *Cross-Validation.* Springer US, Boston, MA, 532–538. https://doi.org/10.1007/978-0-387-39940-9_565

[33] Akane Sano and Rosalind W Picard. 2013. Stress recognition using wearable sensors and mobile phones. In *2013 Humaine association conference on affective computing and intelligent interaction*. IEEE, 671–676.

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

[35] 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. In *Proceedings of the 26th International Conference on World Wide Web*. 103–112.

[36] Robert L. Thorndike. 1953. Who belongs in the family? *Psychometrika* 18, 4 (1953), 267–276. https://doi.org/10.1007/BF02289263

[37] Kush R Varshney. 2021. Trustworthy Machine Learning. *Chappaqua, NY* (2021). http://trustworthymachinelearning.com/trustworthymachinelearning.pdf

[38] 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

[39] W Zhang et al. [n. d.]. Putting human behavior predictability in context. EPJ Data Sci. 10 (1), 1–22 (2021).

A APPENDIX

In this Appendix, we present two visualization plots that depict country differences across all activities (Figure 6) and between 'Eating' and 'Shopping' (Figure 7), as revealed by sensor data from all modalities. When the p-value approaches the significance threshold, a closer examination of the embeddings can aid in a better understanding of country differences. For example, when we inspect the differences between Mongolia and the UK, which had a low PERMANOVA p-value and a low PERMDISP p-value (left plot in Figure 6), we observe that the embeddings of both countries are mixed in one of the clusters, indicating that the significant PERMANOVA test might be due to dispersion. Conversely, when we examine the UMAP plot for the UK and India (right plot in Figure 6), the separation between the two countries' embeddings is more distinct. Similarly, ADL differences can be visualized. For instance, the UK and China display differences in 'Eating' but not in 'Shopping' (see Figure 7).

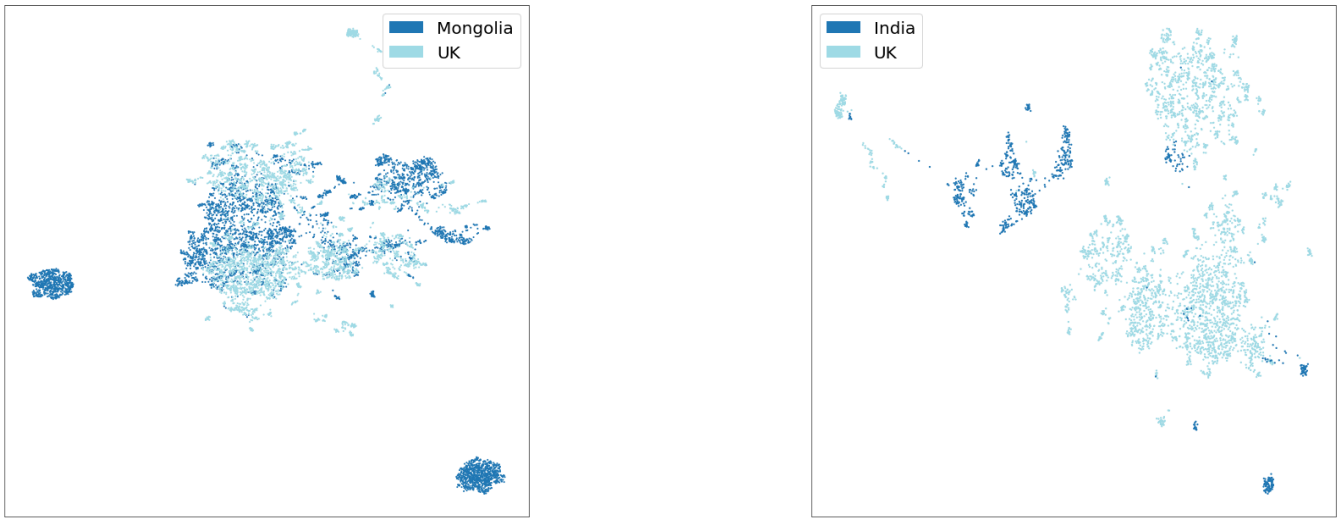

Figure 6: On the left, UMAP plot comparing Mongolia and the UK users. On the right, UMAP plot comparing the UK and India users. Each dot on a plot is the 2D projection of a user embedding capturing sensor data (all modalities) across all activities

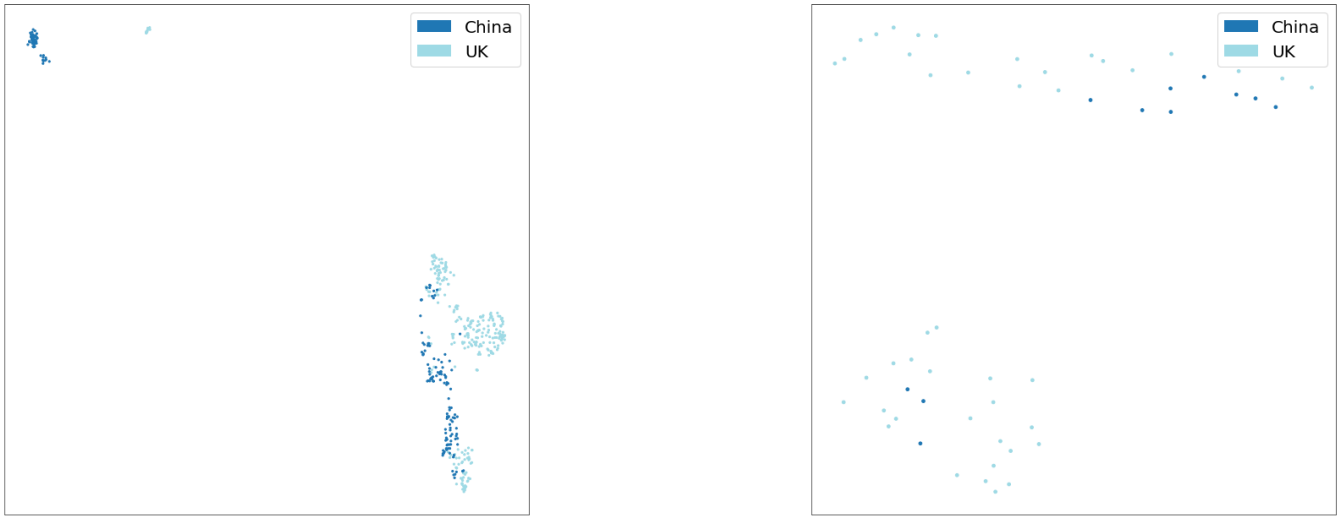

Figure 7: On the left, UMAP plot comparing China and the UK users while 'Eating'. On the right, UMAP plot comparing China and the UK users while 'Shopping'. Each dot on a plot is the 2D projection of a user embedding capturing sensor data (all modalities) for a given activity