Outliers in Smartphone Sensor Data Reveal Outliers in Daily Happiness

TEODORA SANDRA BUDA, Koa Health, Spain
MOHAMMED KHWAJA, Imperial College London, UK and Koa Health, Spain
ALEKSANDAR MATIC, Koa Health, Spain

Enabling smartphones to understand our emotional well-being provides the potential to create personalised applications and highly responsive interfaces. However, this is by no means a trivial task – subjectivity in reporting emotions impacts the reliability of ground-truth information whereas smartphones, unlike specialised wearables, have limited sensing capabilities. In this paper, we propose a new approach that advances emotional state prediction by extracting outlier-based features and by mitigating the subjectivity in capturing ground-truth information. We utilised this approach in a distinctive and challenging use case – happiness detection – and we demonstrated prediction performance improvements of up to 13% in AUC and 27% in F-score compared to the traditional modelling approaches. The results indicate that extreme values (i.e. outliers) of sensor readings mirror extreme values in the reported happiness levels. Furthermore, we showed that this approach is more robust in replicating the prediction model in completely new experimental settings.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**.

Additional Key Words and Phrases: Smartphone Sensing, Mental Health, Mobile Health, Machine Learning

ACM Reference Format:
Teodora Sandra Buda, Mohammed Khwaja, and Aleksandar Matic. 2021. Outliers in Smartphone Sensor Data Reveal Outliers in Daily Happiness. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 5 (March 2021), 19 pages. https://doi.org/10.1145/3448095

1 INTRODUCTION

Enabling smartphones to be aware of the user's emotional states has attracted a significant research attention over the past decade [12, 32]. The enthusiasm around the automatic detection of emotional states using only smartphone sensing has come from the opportunity to build responsive and personalised interfaces, to suggest more relevant content at the right time, and to power applications for mental well-being management and intervention. A plethora of smartphone studies have already explored prediction of various well-being states including stress [6, 24, 25, 34], anxiety [22], mood [4, 30] and happiness [7, 24, 25]. These studies have provided a proof of concept demonstrating that smartphone data carries predictive power for predicting user's states, however only with a moderate accuracy.

The automatic detection of emotional states primarily relies on using phone logs and sensor data to capture a set of users behaviours, which are mapped to the reported emotional states by training a machine learning model. Therefore, the two critical ingredients for developing an accurate emotional state prediction model include: 1) features that carry predictive power for detecting specific emotional states, and 2) reliable ground-truth information. The predominant approach in the literature relies on computing features that represent descriptive

Authors' addresses: Teodora Sandra Buda, sandra.buda@koahealth.com, Koa Health, Spain; Mohammed Khwaja, mohammed.khwaja16@imperial.ac.uk, Imperial College London, UK, Koa Health, Spain; Aleksandar Matic, a.matic@koahealth.com, Koa Health, Spain.

This work is licensed under a Creative Commons Attribution-NonCommercial International 4.0 License.

© 2021 Copyright held by the owner/author(s).
2474-9567/2021/3-ART5
https://doi.org/10.1145/3448095

statistics computed from sensor readings, which are used to predict scores that correspond to reported emotional states [4, 7, 30] (either an absolute score - *regression* or grouped scores into two or more classes - *classification*). However, when it comes to the prediction of dynamic user states (as opposed to predicting more stable constructs such as health symptoms [11] or traits [10, 13, 28]), this approach can limit the full potential of automatic emotional state prediction models. Firstly, using raw sensor values (without regard to the individual's typical sensor readings) to compute features can introduce noise in the models designed to distinguish dynamic (e.g. daily) emotional states. This happens due to differences in personal habits, surroundings and phone models - e.g., for one person, walking 10k steps in a day can be a part of their routine whereas for another, it can represent an extreme value; low light sensor values can be normal in northern Europe whereas an exception in typically sunny southern European cities; or the same noise/light conditions can be captured differently due to sensor sensitivity differences in dissimilar phone models. Using features that are not put in the context of a user's typical routines leaves the machine learning model ill-equipped to infer meaningful variance in sensor readings and associate it to an emotional state variation. Secondly, splitting classes based on an arbitrarily chosen value (typically the median value to create high vs low binary classes) creates a "gray" area around the border between the two neighboring classes that, due to the subjective nature of emotional reports, can result in unreliable ground-truth labels. For instance, reporting a score of 3 on a 1-7 Likert scale can have a completely different meaning for two dissimilar individuals.

In this paper, we propose a novel methodology to build models for the automatic recognition of dynamic user states by extracting more robust features as well as less subjective ground-truth information. Our approach is based on computing features as outliers [21] in a user's phone sensors data as well as using outliers in the subjective reports of emotional states. This allows development of emotional state models that are more agnostic to differences across the sample of users - for instance, by using our approach the same daily value of 7k steps will be represented in the feature space differently for Person-A who typically makes 2k steps (e.g. an elderly person) and for Person-B who typically makes 15k daily steps (e.g. a sportsperson). Similarly, we mitigate the impact of subjectivity in reporting user emotional states [16] by focusing on ground-truth information that represent individual outliers i.e. exceptionally high or exceptionally low scores. In this way, ambiguous scores that can introduce noise in the model are discarded from the learning process. We demonstrate our approach in the task of daily happiness detection. Happiness is an enormously complex construct to define, measure [1, 26] and more so to detect automatically [23]. We hypothesised that days that are reported as exceptionally happy or miserable will be mirrored in extreme values of the smartphone sensors. Using this approach, we achieved a prediction improvement of up to 13% in AUC and 27% in F-score in comparison to a detection model that used state-of-the art methodology of feature extraction as evidenced in our study.

The main contributions of this work are:

(1) A new methodological approach (summarised in Figure 4) in building predictive models of emotional states based on:
   (a) feature extraction that is agnostic to individual differences;
   (b) mitigation of subjectivity in self-reports of emotional states;
(2) Evidence that outliers in smartphone data are more predictive in detecting high and low daily happiness scores than typically used features in the literature;
(3) Evaluation of the model replicability by performing cross-validation in two completely distinct experimental cohorts.

2 RELATED WORK

Literature in mobile computing reports a handful of studies that evaluated machine learning models based on smartphone data to predict emotional states including mood, stress, and happiness. The overarching conclusion

that can be drawn from previous studies is that smartphone data carries the predictive power in predicting emotional states, although the prediction accuracy is limited. For this reason, automatic emotion recognition efforts have focused more on using body signals that are more sensitive to human emotions such as Electro-Dermal Activity, Heart-Rate Variability, voice characteristics, facial expression analysis, EEG, and so on. However, such approaches require special equipment or additional wearable devices, whereas relying solely on smartphone data allows for more scalable solutions.

Predicting daily happiness has been addressed by a few studies [7, 23, 25]. Bogomolov et al. [7] proposed machine learning based prediction of happiness using smartphone data, weather and personality traits gathered from 117 participants from a US university. The authors obtained 25 communication (calls & sms) features and 9 proximity features for each data point obtained from mobile phones, and subsequently calculated descriptive statistics of these features to capture daily user behaviour. Happiness scores were grouped into three classes based on daily reported scores as 'not happy' (for scores lower than 4), neutral (for scores equal to 4) and happy (for scores higher than 4). Using Random Forest classification, the authors reported a prediction accuracy of approx. 80%, which, however, was boosted by the fact that the dataset was imbalanced (76% happy days, 14% neutral days and 10% unhappy days). With a similar goal of detecting happiness and stress, Jaques et al. [23–25] explored predicting daily happiness harnessing students' smartphone sensing data as well as data collected from a smartwatch. The authors used binary labels for classification: 30% of the days that had the highest happiness value were assigned positive and 30% of the days that had the lowest happiness were assigned negative. Unlike in our study, they constructed the two classes based on population dependent thresholds rather than at an individual level. They achieved prediction accuracy of 56 - 69% in classifying daily happiness, with significant improvements introduced by adding data collected through a smartwatch.

In a similar line, a few studies focused on predicting mood. Ma et al. [30] built a system called MoodMiner that extracted daily aggregated features from mobile sensing (accelerometer, microphone, location, light) and communication (calls, sms) data to predict daily mood information. Self-reported mood scores were grouped into one of five levels - which represented a target variable for the developed machine learning model. The authors reported the accuracy between 48 and 52% and RMSE between 1.44 and 1.56 for the sample of 15 users. Similarly, Asselbergs et al. [4] aimed to predict daily mood that was computed from multiple momentary mood evaluations sampled through the Ecological Momentary Assessment (EMA) method. The smartphone data and the mood reports were collected from 27 Dutch student participants Calls/sms, accelerometer, screen usage, app usage and image description data were aggregated on a daily basis. The models predicted daily mood with 55 - 76% accuracy, yet the authors reported that the model performance was inferior to the baseline models. The authors concluded that passive mood detection is technically feasible with the data collected by smartphones, however that the methods used by researchers need to be further explored and improved to become practical implementations.

Applying a similar methodology, Pratap et al. [33] used the smartphone data from 271 Android phone users to predict daily mood (PHQ-2 score [29]). They obtained daily aggregated features for GPS-mobility (distance travelled, travel radius) and communication (unreturned calls, number of SMSs sent etc.) and observed that mobile sensing features were not highly predictive, whereas adding demographics improved the performance of prediction models to a median AUC of 0.5.

The focus of our review of the related work was placed on identifying studies that relied on the smartphone sensing data and on those that aimed to predict emotional states of healthy people including mood and happiness. For brevity, we did not delve into literature that relied on different sensing modalities or that aimed to predict other health variables - the respective reviews can be found in [12, 32].

To the best of our knowledge, there is no study that explored the predictive power carried by the outliers in sensor data in the context of predicting emotional states. On the other hand, the only work that processed ground-truth information in a similar manner to our approach was reported in [24] - that used population-based thresholds. Moreover, to our surprise there is no study on smartphone-based emotion detection that has attempted

to replicate their models, built with one specific user sample. Reproducibility crises has been recently emphasised in multiple domains - specifically, a very recent paper, Beam et al [5] discussed challenges in replicating machine learning models in healthcare. In our study, we also demonstrate the benefits of our approach in reproducing the algorithm in a completely new cohort with substantially different demographics.

3 BACKGROUND

3.1 Modelling Happiness

Happiness — or subjective well-being (SWB) as commonly referred to in the literature — is defined as the cognitive and emotional evaluation of life [14]. Measuring happiness is of a paramount importance in assessing the intervention impact in clinical settings or with digital health applications [37], in evaluating policy changes [18], in understanding epidemiological impact on well-being, in tracking societal progress on a national scale [3] and many more.

The measures of SWB refer to measuring experiences or evaluations. The former involves an accumulation of emotional state reports about how people feel in the moment (typically referred to as happiness). The latter relies on asking people how they feel about their lives overall (typically referred to as life satisfaction). Experiential measures are considered to be the most accurate way to approximate SWB as they capture everyday experiences, emotional states and associated happiness fluctuations [17]. The two most commonly applied methods to measure experiential SWB include Ecological Momentary Assessment (EMA) and the Day Reconstruction Method (DRM). EMA prompts people to report their well-being at specific moments (usually selected randomly throughout a day), whereas DRM requires people to report well-being relating to the past periods usually at one point during a day (e.g. in the evening). Both methods require users' cognitive load, time and may have an impact on well-being.

To replace (or complement) self-report based methods, passive smartphone sensing created a new way of tracking users' behavioural routines without requiring users to self-report how they feel or what they do. Quantifying mobility, social behaviour, and other activity- or context- related patterns have been expanded in the recent literature towards modelling complex psychological characteristics. In this paper we built a smartphone based measurement model of happiness, which we refer to as a user's state (related to experiential SWB) that can vary from one day to the other.

3.2 Smartphone-based Modelling of Emotional States

Existing literature on smartphone-based modelling of a user's psychological characteristics can be broadly categorised based on the modelling target – *states* or *traits*. The methodological difference between these two approaches stems from the conceptual definition of these two constructs. Psychology broadly refers to *states* as temporary indicators of the status of an individual – states are brief and typically caused by external circumstances, whereas *traits* represent stable statements about an individual – traits are long-lasting and internally caused [20]. In this regard, to detect a user's traits from smartphone data, researchers developed features that quantify longer-term patterns of individual's behaviours. Machine learning models were designed to map the quantified behavioural differences among people to the reported personal traits [10, 13, 28]. As opposed to the trait models, features designed for detecting emotional states are extracted for shorter periods of time (e.g. daily) and the models are trained to infer the self-reported temporary states [7, 30, 33].

To model emotional states, features are extracted to represent the periods for which the models will be queried. These features are typically computed as descriptive statistics of absolute sensor readings without taking into account the individual's behavioural routines, and therefore the model is powered to contextualise sensor readings with respect to the population behavioural patterns rather than to infer emotional response with respect to the individual routines. This may leave the model ill-equipped to capture important signals from the extracted features and to infer emotional response with respect to the individual's characteristics and behaviours as well as

with respect to the sensors characteristics which can vary across phone manufacturers. Furthermore, capturing a meaningful ground-truth information is another critical aspect when it comes to designing state detection learning models. Whereas the psychological scales that evaluate personal traits are designed to directly compare people and therefore absolute values are meaningful also for the automatic models to infer traits, self-reports of the emotional states are very subjective and the same score may not mean the same emotional state for each person.

The main goal of our approach is to improve smartphone based models of user's states by extracting features that are more agnostic to: (a) the sensor specifications, and (b) the subjectivity in ground-truth information, by contextualising each feature value to the individual's behavioural and emotional patterns by leveraging the concept of outliers.

3.3 Outliers: Definition and Traditional Resolve

To address the problem of individual specificity (coming either from sensor characteristics, behavioural differences or subjectivity in happiness reports), we embedded the concept of outliers in the feature extraction as well as in ground-truth pre-processing. We aimed to capture when the sensor readings and happiness reports were deviating from typical values for each individual independently of the distributions in the whole sample. Hawkins [21] introduced the definition of an outlier as "an observation which deviates so much from the other observation as to arouse suspicions that it was generated by a different mechanism". These points are often characterised as *discordants*, *deviants*, or *anomalies* in the data mining and statistics literature. Outliers can be indicative of unusual activity or patterns within the collected data, and their detection has been the focus in a number of application areas, such as: intrusion detection, credit-card fraud, interesting sensor events, medical diagnosis, and others [9]. In all of these applications, the data has fit into a normal model (these points are considered inliers) and outliers are considered deviations from the normal model. As Aggarwal [2] highlights, it is often a subjective judgement as to what constitutes a *sufficient* deviation for a point to be considered an outlier, and frequently the labelling has been performed through either visual inspection by a data expert, or through explorations and optimizations of different threshold mechanisms [8]. Inspired by this field, we proposed the outlier-based methodology for detecting deviations in user states based on deviations in measurements captured from their phone sensors.

4 OUTLIER-BASED APPROACH

In this section, we will first highlight the motivation behind our outlier-based approach for computing features and for preprocessing the ground-truth information. Subsequently, we will delve into the description of our approach.

4.1 Main Challenges

We performed an exploratory analysis of the features extracted in line with the existing literature, which led us to develop the outlier based approach in order to tackle the identified challenges. In the following, we overview the main challenges that we identified in the preliminary exploratory analysis.

4.1.1 Sensors Sensitivity. Witnessing different distributions and range of values captured from different phones prompted us to compare the sensor data from the phones placed in the same conditions. We collected sensor data from four different devices left stationary in the same location over a few days. Figure 1 shows the light levels and the accelerometer readings (magnitude computer using three dimensional values). Although the light levels were expected to be highly comparable across the four devices, the actual sensor readings differed significantly. Interestingly, the data captured from the accelerometer exhibited noisy readings despite the fact that the mobile phones were stationary.

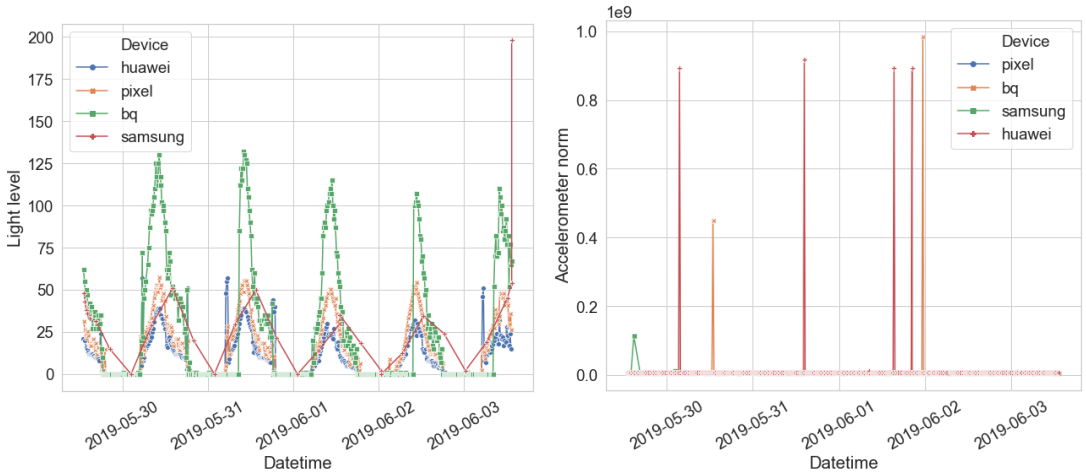


Fig. 1. Sensors from different smartphones - sampled at the same time and under the same conditions while stationary (example of light sensor and accelerometer)

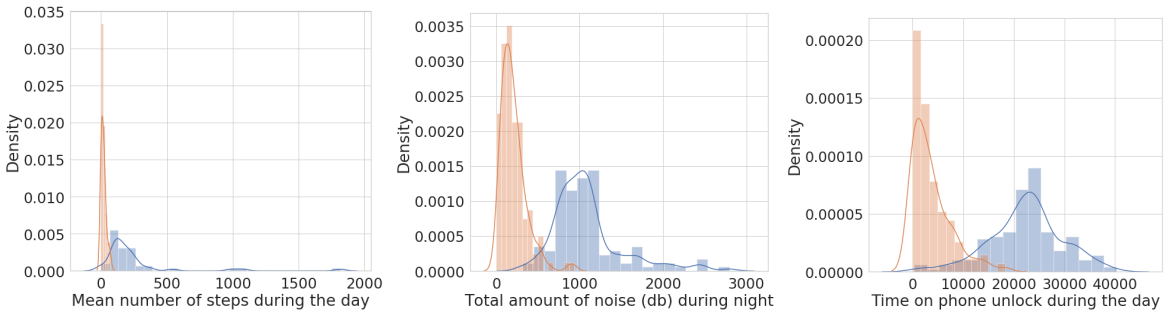


Fig. 2. Distribution of mean daily steps count, total noise level at night and the amount time the phone was unlocked for individuals with median belonging to the first quartile (left), and forth quartile (right) of the corresponding distributions.

4.1.2 Individual Routine Differences. Figure 2 shows the distribution of some daily user behaviours including, number of steps, noise level exposure and time that the phone was unlocked for individuals with median belonging to the first and forth quartile of the corresponding distributions from one of our considered cohorts. This illustrates the variation in the sensor readings across our cohort. Different distributions of the features for different groups of individuals exemplifies the fact that extracting features from shorter time-frames using the raw sensors without regard to an individual's distributions may limit the potential of user state predictive models.

4.1.3 Subjectivity in Ground-truth. Figure 3 shows the distribution of happiness scores (rated from 1 to 10) over two weeks, for individuals with median happiness belonging to the first and forth quartile of the happiness scores distribution from one of our considered cohorts. Figure 3 illustrates individual differences in the range of values used to rate subjective happiness levels, which may be mislabeled by applying population-devised reference scores.

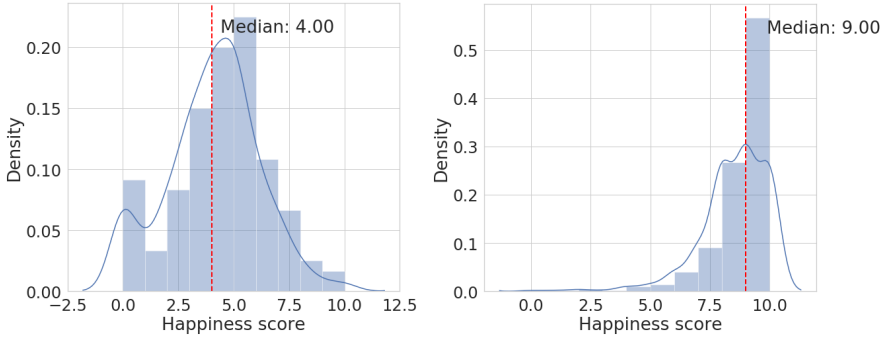


Fig. 3. Density of happiness scores for individuals with median happiness score belonging to the first quartile (left), and forth quartile (right) of the happiness distribution.

4.1.4 Traditional Ground-truth Processing. The right sub-figure of Figure 3 shows the distribution of happiness scores of individuals with a median happiness value of 9. For these individuals, pursuing the traditional approach for ground truth categorisation would label all instances under 9 as 'low' and all above as 'high'. As it can be observed, the 'low' would capture a high variation of scores, between 4 and 8, thus introducing noise in the target labelling.

4.2 Outlier-based Modelling

This section describes our approach for advancing the methodology for automatic detection of emotional states of smartphone users. The overview of the approach is presented in Figure 4 that illustrates the following process: (1) The raw sensor features are transformed into *synthetic* features based on the individual's distribution for each feature. Outliers and regular values are then converted into their corresponding mapping through an encoder. As an example, the value of the sensor 1 reading in *Day X* is mapped to a regular category based on the individual's distribution, further encoded as 2; (2) similarly, mapping the target feature to the corresponding category based on the individual's distribution. The happiness score for *Day X* is mapped in this example to a lower outlier, further encoded as 1; (3) finally, training a machine learning model based on encoded values from the sensor and target for all the individuals in the dataset with the goal of predicting outliers in the user states (in our study – happiness scores).

The proposed outlier-based approach consists of the following four phases:

(1) *Feature extraction and selection*: The process starts with computing features in the traditional way – based on raw sensor values. These features typically represent aggregated sensor readings at daily level (e.g., number of steps daily), as well as chunked into specific time windows (e.g., morning, afternoon, evening, night).

(2) *Synthetic features generation*: The second phase includes generating synthetic features based on the features extracted from the raw sensor data (in the previous phase). We map the sensor features based on each individual's distribution to lower outlier, regular or upper outlier, converting these to encoded values 0, 1 and 2, respectively. We explored various encoding mechanisms, including using the numerical values due to their ordinal properties as well as *one hot encoding*, which performed comparably.
In a variation of our proposed outlier-based approach, we first explored mapping the sensor features extracted from a specific period (daily in our study) to the percentiles according to each individual's distribution. Through this approach, we circumvent using the features extracted from the raw sensor values,

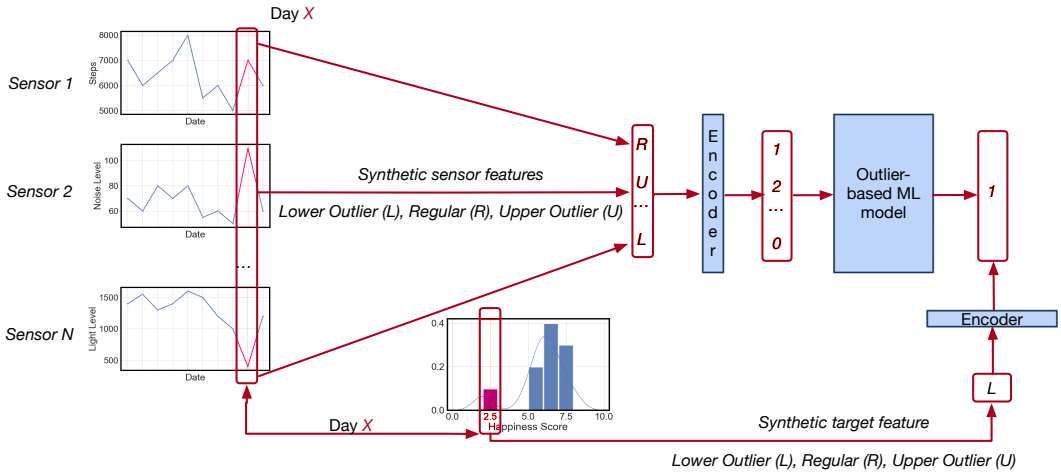


Fig. 4. Outlier-based approach for the automatic emotional state detection: First, aggregate sensor features are constructed from raw sensors. Second, based on each individual's distribution, sensor features are encoded to either lower, regular and upper outliers. Third, the target feature is also encoded to outliers based on each individual's distribution. Finally, a model is trained to map the encoded input features to the encoded target feature.

while maintaining more information than the outliers only. Interestingly, although this approach produces features that carry more information than the outliers only, the results showed a better performance of the classifier based on the outliers.

(3) *Synthetic target feature generation*: Similar to above, we map the reported target scores (happiness) to lower outliers, regular and upper outliers, converting these to 1, 0 and 2, respectively. It is important to mention that for training, we only utilise the outlier values, lower and upper, corresponding to 1 and 0, respectively.

(4) *Modelling*: In this phase, the machine learning algorithm is trained by using the synthetic input sensor features and the ground-truth (target) labels.

5 EXPERIMENTAL METHODOLOGY

5.1 Data Collection

As part of a larger study focused on life-sensing and well-being, participants were provided with our custom data collection App (built for both iOS and Android). For the period of the study (2 to 3 weeks), the App was collecting phone usage logs and sensor data. The app was also prompting users to score their momentary happiness levels 5 times per day (through Ecological Momentary Assessments – EMA), and one daily happiness question – "how happy did you feel today?", both on 1-10 scale (happiness is frequently referred to as subjective well-being [16]). Interestingly, averaging EMAs to compute one daily score yielded scores that were highly correlated with the daily happiness question (the correlation was above 0.95) and for this study we used the latter as the ground-truth information about daily happiness. Additionally, we collected demographic questions from an onboarding questionnaire. No personal identifying data, such as name, phone number or email address, was collected.

Upon the App installation, participants were first presented with a consent form detailing the objective of the study and data collected (in compliance with the GDPR regulations). Sensing data included accelerometer, microphone (only noise levels in dB), pedometer, GPS location and ambient light sensor. The phone usage

Table 1. Number of data points and users per cohort that have daily happiness recordings before and after filtering. The definition of filtering criteria 1 and 2 is provided in Section 5.2

	Diverse Cohort		University Cohort	
Filtering Criteria	Number of data points	Number of users	Number of data points	Number of users
Before filtering	2892	221	4148	481
Android users	1479	218	2869	170
Filtering criteria 1	1075	189	360	114
Filtering criteria 2	736	59	153	15

logs consisted of battery level/charging status, phone lock/unlock events and calls made/received. Data from accelerometer, GPS location and microphone was collected every 15 minutes, while data collection from the other sources was event-triggered. The battery consumption attributed to our App did not exceed 8% and we have received no major complaints about disruption of the normal phone usage. In line with the typical approaches from the existing literature [19, 28, 38], we extracted the features as outlined in Table 2, which were used to extract new synthetic outliers based features as explained in the previous section.

In accordance with the GDPR regulations, participants could decide to opt out from the collection of data such as call/text logs, location and noise data. We allowed all the users to participate in the study and to be entitled to the same monetary reward regardless of the data categories that they consented to share. The rationale behind this inclusion criteria was to attain conditions of typical real services in which users have the option to not provide specific data categories. This is due both because of (a) GDPR, which strictly defines differential data collection (b) OS policies (note that, e.g., Android applications provide an opt-out pop-up screen soon upon their installation in case location, microphone or call logs are captured). Moreover, we believe that obliging participants to accept the collection of all data categories would make our sample biased (e.g. participants from a lower socio-economical status would perhaps accept more rigorous conditions for the monetary compensation).

5.2 Participants

In total, the data from 702 users was used for this study. Participants were recruited from two different cohorts. The first one (referred to as 'Diverse Cohort') included 221 users from five different countries (Chile, Colombia, Peru, Spain and the United Kingdom) who were recruited by an external recruitment agency from February to August 2018. The second cohort (referred to as the 'University Cohort') included 481 participants who were recruited from January to March 2019 from a major UK university using an internal email sent to staff and students. Participants were asked to use the application for a period of 2-3 weeks, and upon a successful completion of the study they were rewarded with a monetary compensation.

Due to a significantly higher number of Android as opposed to iOS users in the "Diverse Cohort" and to inconsistencies in the collected sensor between the two mobile platforms, we opted to only include Android users in our analysis. We also reported the drops in the number of users that this decision caused in Table 1. Furthermore, we observed important inconsistencies in the data due to the users' opt-outs from collecting location, noise, and call/txt logs (or the combination of the three). To mitigate the modelling problems, we discarded the data from (1) days with less than 70% of all features computed for our study, (2) users that provided less than 7 days of daily well-being and smartphone data. The first filtering criteria was set based on the typical practices in the recent work [28, 31]. The second filtering criteria was used to ensure that there were enough data points per user to compute outliers.

Applying the first filtering criteria, the number of users dropped to 303 users (189 users in the Diverse Cohort and 114 users in the University Cohort) that provided 1, 435 entries for daily well-being. Importantly, despite

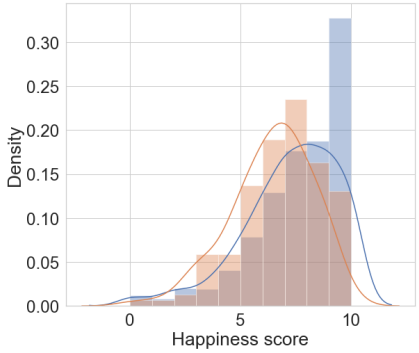


Fig. 5. Self-reported happiness density distributions in the Diverse (blue) and University Cohort (orange).

the drop from the initial number of users, the size of our data set was considerably higher from the majority of the related studies that had between 15 [30] and 271 [33] participants. The country distribution for the Diverse Cohort is as follows: most participants were from Spain (N=66), and other participants were from Peru (N=41), Colombia (N=36), Chile (N=31) and UK (N=9); 6 users didn't provide their country information. The gender ratio in the Diverse Cohort was roughly 1:1.6 (73 women : 116 men), while in the University Cohort it was roughly 1.3:1 (64 women : 49 men). In the Diverse Cohort, most of the participants were aged between 26-34 (N=117) and other participants were aged 18-25 (N=18), 35-44 (N=19), 55-64 (N=1) and 65+ (N=1).[1] Whereas, for the University Cohort, most of the participants were aged between 20-26 (64 users), with other users belonging to other ranges including 18-19 (N=13), 27-29 (N=10), 30-34 (N=13), 35-39 (N=5), 40-44 (N=5), 45-49 (N=2), 55+ (N=2)[2]. Additionally, 99 users in the Diverse Cohort and 46 users in the University Cohort were employed, while the rest were unemployed or economically inactive.

Figure 5 shows the distribution of daily happiness scores across the two cohorts, after preprocessing. The mean self-reported happiness was 6.4 and 7.3 in the university and in the Diverse Cohort, respectively.

5.3 Features Extraction and Selection

We computed daily aggregate features for the entire day (12 AM - 11:59 PM), using guidelines and methods established for various data modalities in past literature [24, 28]. The features extracted from each data modality are summarised in Table 2. For some modalities, including light, noise and pedometer, we also obtained features during the early morning (4 AM - 5:59 AM), morning (6 AM - 12:59 PM), noon (1:00 PM - 2:59 PM), afternoon (3:00 PM - 6:59 PM), evening (7:00 PM - 9:59 PM) and night (10:00 PM - 3:59 AM). Thus a total of 207 features were obtained per day per user with the dataset. Granular features were obtained during different parts of the day for the light, noise and pedometer data modalities as data was sampled and collected more frequently compared to the other modalities. Recursive Feature Elimination was performed an XGBoost model to reduce and determine the most important 25 features. This process was always performed using the data solely from the training set.

5.4 Modelling

We explored probabilistic, tree-based and neural networks, including Logistic Regression, Random Forest, Multi-layer Perceptron and XGBoost. The latter outperformed the other models and thus we present the results of the XGBoost models further. Moreover, as XGBoost is robust to missing data categories, we could use the data of

[1]We asked the age range of participants, instead of exact age values.
[2]The age ranges were slightly different for the University Cohort to accommodate the diversity in student populations

Table 2. Data Modalities and Features Extracted

Modality (Total Features)	Features Extracted
Accelerometer (15)	- Correlations of acc. in 2 directions - Energy in each direction - Std. dev. of acc. in each direction - {Min, max, mean, std. dev.} of acc. - Sampling rate
Battery (7)	- {Mean, std. dev.} of no. of charges - {Min, max, mean, std. dev.} of level - Binary flag: is phone charged? (y/n)
Light (70)	{Min, max, mean, median, total} light level (lux) and scaled light level (0-1) during {early morning, morning, noon, afternoon, evening, night, entire day}
Location (6)	- Time spent in top 5 clusters - Radius of gyration
Noise (70)	{Min, max, mean, median, total} noise level (raw) and noise light (in dB) during {early morning, morning, noon, afternoon, evening, night, entire day}
Pedometer (35)	{Min, max, mean, median, total} number of steps during {early morning, morning, noon, afternoon, evening, night, entire day}
Type of Day (2)	- Binary flag: is weekday? (y/n) - Binary flag: is weekend? (y/n)
Unlock (2)	Time spent on phone lock and unlock

heterogeneous participants who provided consent for varying data categories (e.g. provided location but did not provide noise data and vice versa). We utilised early stopping based on log loss and a validation set extracted from the training set. We evaluate our model in two different experimental settings:

(1) *Exploratory setting*: for comparing our to the traditional approach in automatic prediction of user states as well as for exploring optimal thresholds for setting outliers in the features; here we use Leave One User Out cross-validation and filtering criteria 1.
(2) *Replicability setting*: for exploring the transferability of our model to an unseen dataset; here we train our model using the data from one cohort, and test it on the other cohort, after applying filtering criteria 1 and 2.

As often criticised in many scientific domains, findings in different contexts do not replicate and are tuned to specific study demographics (e.g. students). This is particularly related to smartphone behavioural modelling where both technical and human factors endanger the replicability and usefulness of the reported models. Exploring the outlier-based modelling approach in two different experimental settings as described above is an attempt to validate our approach and its transferability to an unseen dataset.

6 RESULTS

This section presents the results of our outlier-based proposed approach compared to three other techniques:

(1) *Percentiles-based approach*: In this variation of the outlier-based approach, we replace the raw values for each feature with the corresponding percentile, based on the individual's own distribution.
(2) *Traditional sensor-based approach*: This baseline represents the state-of-the-art approaches where features built on raw sensor values are used to predict the daily happiness score. The features are detailed in Table 2.
(3) *Majority class*: This baseline represents a model that predicts the majority class at all times.

Table 3. Results from comparison between outlier-based approach and baselines in experimental environment.

Cross-validation Methodology	Model	Confusion Matrix	AUC	F-score	Kappa	Accuracy
Leave one out (Diverse cohort)	Outlier-based approach	[87% , 13% 59% , 41%]	**0.64**	**0.49**	**0.3**	**0.71**
	Percentiles-based approach	[87% , 13% 87% , 13%]	0.50	0.19	0.003	0.62
	Traditional sensor-based approach	[87%, 13% 85%, 15%]	0.51	0.22	0.02	0.62
	Majority Class	[100% , 0% 100% , 0%]	0.50	0.00	0.00	0.66
Leave one out (University cohort)	Outlier-based approach	[36% , 64% 16% , 84%]	**0.60**	**0.72**	**0.21**	**0.63**
	Percentiles-based approach	[27% , 73% 27% , 73%]	0.50	0.63	-0.002	0.53
	Traditional sensor-based approach	[30% , 70% 27%, 73%]	0.52	0.65	0.03	0.55
	Majority Class	[0% , 100% 0% , 100%]	0.50	**0.72**	0.00	0.57

Table 4. Number of days across datasets in first setting.

	# Days		
Cohort	Upper outliers (Label 0)	Lower outliers (Label 1)	Total
Diverse	178	91	**269**
University	118	157	**275**

6.1 Exploratory Setting

The main goal of this analysis was to compare the accuracy of the user state classifier built using our outlier-based approach and the typical approach. We also explored which outliers in sensors corresponded to which outliers in well-being. First, we mapped the sensor features values to their corresponding label as defined in Section 4.2. Second, we labelled the days with outstandingly high vs outstandingly low happiness scores based on each individual's distribution and encoded them with 0 and 1, respectively. Further we utilised these outlier days only for training and testing, similarly to [24]. We utilised *leave one user out* as cross-validation methodology. Moreover, we explore the following thresholds for labelling outliers: 5, 10, 15, 20, 25 percentiles and 1, 2 and 3 standard deviations from the mean. Rows that had more than 30% of the values missing were removed to increase the quality of the data, similarly to the filtering criteria applied in [31]. Table 4 shows the number of days that were left after filtering. The upper outliers correspond to days with high happiness scores, and lower outliers correspond to days with low happiness scores. The size of the dataset after filtering (i.e. the number of participants as well as daily instances) corresponds to the state-of-the-art studies in this domain (e.g. [22, 23, 34]).

Table 3 presents the results from the comparison between the outlier-based approach and the selected baseline methods in the exploratory setting. We optimised for the Area Under the Curve (AUC) score and this resulted in thresholds up to 15 percentile as optimal for mapping raw sensor values and happiness scores to outliers and regular values. However, there was no single threshold that was considerably better than the others across both datasets and testing configurations, which indicates that further studies are needed to explore the existence of

Table 5. Results from comparison between outlier-based approach and baselines in close to real-world setting.

Cross-validation Methodology	Model	Confusion Matrix	AUC	F-score	Kappa	Accuracy
Train on Diverse cohort, Test on University cohort	Outlier-based approach	[68% , 32% 46% , 54%]	**0.61**	**0.44**	**0.19**	0.64
	Percentiles-based approach	[93% , 7% 93% , 7%]	0.50	0.12	0.002	0.69
	Traditional sensor-based approach	[85%, 15% 73%, 27%]	0.56	0.32	0.13	0.69
	Majority Class	[100% , 0% 100% , 0%]	0.50	0.00	0.00	**0.73**

Table 6. Number of days across datasets in second setting.

	# Days		
Cohort	Upper outliers (Label 0)	Lower outliers (Label 1)	Total
Diverse	271	184	**455**
University	112	41	**153**

the optimal threshold. The table shows that the outlier-based approach reaches superior results across all metrics considered when compared to using traditional features.

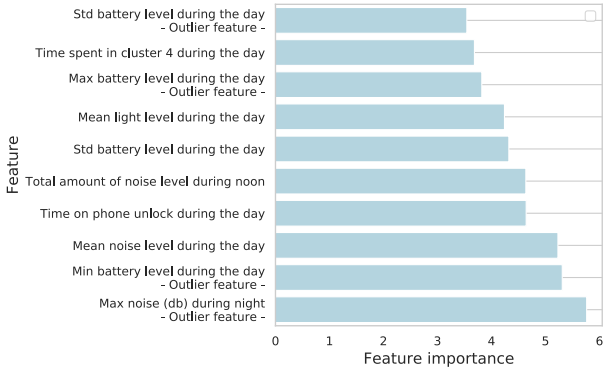
6.2 Replicability Setting

Having two cohorts with considerably different user samples across a set of demographics (nationality and country of residence, occupation, recruitment technique which may introduce different sample biases, etc.) allowed us to evaluate the replicability of the daily happiness recognition model. In this section, we explore how the outlier-based approach performs compared to the baselines when trained in one and tested in the other cohort. We applied the same procedure as in the exploratory settings – removing rows that were missing more than 30% of the data and the users that had less than 7 days worth of data. To attain the testing conditions ever closer to the real-world settings, the test data set included the days that did not correspond to outliers in happiness scores i.e. regular days. To preserve the binary classification settings, we labeled these regular days with "high" happiness label. The rationale was that distinguishing outstandingly low happiness days (in the well-being literature typically referred to as "miserable" days) can be more important for personalising phone interaction with the user as well as for mental well-being applications. Note that, as part of our main approach, we still kept only the outlier days in the training dataset. The number of days across datasets are presented in Table 6.

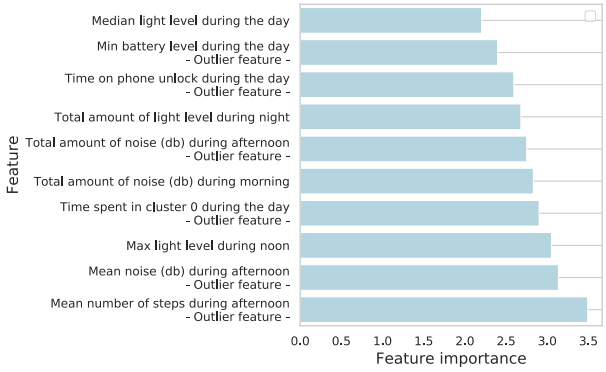
Table 5 presents the results from the comparison between the outlier-based approach and the two baseline approaches in this setting. Similarly to the exploratory settings, we optimised for the AUC metric. Due to the low number of days left in the University Cohort, we decided to train the model by using the Diverse Cohort and test it using the data from the University Cohort. Although with a lower absolute accuracy, we observed that the outlier-based approach again outperforms the baseline models including the traditional method. The confusion matrices show that the outlier-based approach is the one that is able to detect most of the days that were at the individual level labeled as exceptionally low.

Table 7. Results from a model trained on both traditional features combined with outlier-based features.

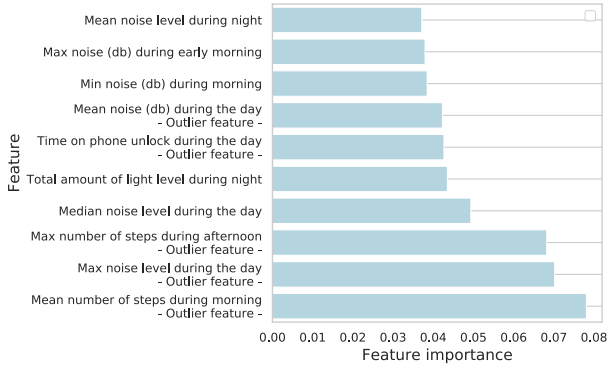
Cross-validation Methodology	Model	Confusion Matrix	AUC	F-score	Kappa	Accuracy
Leave one out (Diverse cohort)	Traditional sensor-based approach coupled with outlier-based features	[87%, 13% 77%, 23%]	0.54	0.31	0.11	0.65
Leave one out (University cohort)	Traditional sensor-based approach coupled with outlier-based features	[33% , 67% 29%, 71%]	0.52	0.65	0.04	0.55
Train on Diverse cohort, Test on University cohort	Traditional sensor-based approach coupled with outlier-based features	[89%, 11% 80%, 20%]	0.54	0.26	0.10	0.70



(a) Exploratory setting - University cohort.



(b) Exploratory setting - Diverse cohort.



(c) Replicability setting.

Fig. 6. Feature importance for a model trained on both traditional features combined with outlier-based features.

6.3 Feature Importance

In order to illustrate the predictive power of the outlier-based features in tandem with the traditional features built on top of raw sensing data, we train a new model that combines traditional features with outlier-based features in the Exploratory and Replicability settings. This model achieved moderate improvements across the evaluation metrics that can be observed in Table 7. This can be due to the large number of features compared to

data instances available for training and testing. Figure 6 illustrates the top 10 predictive features for the model. It can be observed that generally a mixture of traditional features and outlier-based features are amongst the top 10 predictive features, with the outlier-based features dominating the first three places. All figures show that the top predictive feature is indeed an outlier-based feature. We observe that outliers in *steps*, *noise*, and *battery level* are typically most indicative of outliers in daily happiness.

7 DISCUSSION

In this paper, we proposed a new methodological approach in developing smartphone based models that predict dynamic user states. Detecting emotional states has been seen as an enabler of more responsive technologies. In particular, using solely smartphone sensors to understand emotional changes of their users would allow for the scalable applications that do not require additional hardware with the sensors that are more fit-for-purpose for detecting emotions. However, the scalability comes with considerable challenges in developing user emotional state models with a set of sensors typically embedded in smartphones. In this regard, our approach aimed to improve the user state classifiers by 1) feeding the models with relevant information about the individual behavioural changes quantified with features contextualised in individual's routines, 2) focusing only on learning from extreme emotional states to avoid ambiguity in score reports that stem from subjectivity in self-reports.

7.1 Outlier-based Approach in Recognition of User States

Detecting emotions is a complex endeavour, and it comes by no surprise that the literature reported near random accuracy when relied only on smartphone sensing [4, 30]. Though the study presented in this paper did not push the accuracy towards those of emotion recognition based on specialised hardware (such as A/V, Electrodermal activity, heart-rate variability, and similar), our results demonstrated that the outlier-based approach can yield improvements in capturing relevant signals about emotional states. When tested in different configurations the newly proposed method outperformed the traditional user state modelling methods. Interestingly, extracting percentiles that carry more information that outliers only (as it precisely quantifies how (un)usual the current observation is for a specific individual) did not bring accuracy improvements as compared to outliers. In our future work, we will focus further on investigating this phenomenon and if percentiles can bring additional benefits. We believe that these results will encourage researchers to further improve the user state smartphone models. Moreover, our approach is not limited to the smartphone sensor data only – similarly, quantifying individual differences and interpreting individual sensor readings is relevant also for physiological signals (e.g. for interpreting high or low Electrodermal activity in the context of what these mean for different people).

7.2 Happiness Use Case

We demonstrated the advantages of our novel outlier-based modelling methodology in a specific use-case – happiness. The subjective nature of happiness (i.e. subjective well-being) makes the comparison among individuals difficult. Besides, even for the same individual quantifying small differences between days is difficult and error prone. Through mapping raw happiness scores to outlier labels based on each individual's distribution, out method aimed at capturing non-ordinary days while addressing subjective differences in scoring across individuals (e.g. if a person A scores very happy days as 10 and a person B scores very happy days with 6, our method will label these days consistently as upper outliers). Similarly, when it comes to the smartphone sensor based behavioural analysis, the outlier-approach for the features extraction addresses challenges in comparing raw sensor readings, due to specificity in individual behaviours, technical differences across phone models, different contexts, etc. (please see section 4.1). Our paper shows that a narrow aperture of daily behaviours captured by smartphone sensing is still sufficient to pick up important fluctuations in daily happiness.

7.3 Happiness Prediction

Our target variable, happiness, is an enormously complex construct. Its definition has been subject of debates since ancient Greeks [15, 27]. Due to its highly subjective nature, capturing reliable ground-truth information is extremely challenging. Moreover finding relevant happiness fingerprints in a narrow scope of behaviours covered by the smartphone sensors is even more so [36]. Importantly, the goal of our study was not to increase the accuracy in happiness prediction but to demonstrate the added value of our new methodology by comparing it directly, consistently and under fair conditions with other existing modelling methodologies. More interesting is the finding that unusual behaviours captured through smartphone sensors correspond to unusually high/low happiness of the users. Nevertheless, our results are fully comparable with the state of the art [4, 23, 25, 33]. The most relevant studies [23–25] constructed two classes for daily happiness based on population dependent thresholds rather than at an individual level as in our study, and did not include the neutral days neither in train nor test. They achieved a prediction accuracy of 60 - 69% in classifying sad days, leveraging also data collected from a smartwatch. In our future work, we will delve further into the most predictive features and attempt to interpret the results from the psychological perspective.

7.4 Replicability across Heterogeneous Devices and Cohorts

Replicability crisis in multiple domains and in particular in machine learning based user models have been recurrently highlighted [5]. Evaluating the same model in different experiments and under different conditions is essential to advance models towards their real-world applications. The influence of specificity of experimental conditions was also observed in our study, however we also witnessed that our approach is more robust to completely different experimental conditions and provides a higher accuracy when replicated in comparison to the traditional approaches.

An additional challenge to behavioural modelling is that it relies on smartphone data coming from sensor readings that are frequently not consistent across manufacturers and operating systems (please refer to Section 4.1) Inconsistencies in sensor readings may be one of the reasons behind frequent issues with replicability of behavioural models that rely on passive monitoring through smartphones. Although we used only Android phones in our analysis, our cohorts included in total 17 and 11 different phone manufacturers in the Diverse and the University cohort respectively (including the most popular phones such as Huawei, Samsung, Xiaomi, Motorola, LG) with 109 and 79 different phone models in the Diverse and the University cohort, respectively. Despite considerable variations in the sensor readings across Android phone models, we did not observe any influence of a specific phone model on the final model accuracy.

We encourage the use of our approach, independently or in combination with the typical models, in further exploration of replicability of the user state models.

8 IMPLICATIONS

The main contribution of this work is a novel approach that has the potential to improve automatic smartphone-based inference of user dynamic states, e.g. stress, anxiety, happiness. The relevance of having smartphones that understand users emotional states ranges from enabling personalised smartphone interfaces to designing applications that help users improve or manage their mental well-being.

(1) *Customised engagement*: Understanding better when to (or not to) engage with users. For instance, during more miserable days users may feel overwhelmed with a high number of notifications and the phone can provide a more restrictive filtering than during happier days.
(2) *Intervention*: Inferring when a user is having a worse or a better day than usual could help delivering the right intervention at the right time.

(3) *Passive assessment of the intervention effectiveness*: Asking users to report emotional states often represents an intervention per se, which may impact user's mental health negatively (showing negative trends in emotions may trigger depressive states [35]). Having a passive inference of emotional states can provide an indication of whether the intervention has been successful or not (e.g., assessing daily stress levels after the relevant interventions).
(4) *Self-reflection*: Inferring exceptionally good or bad days and raising the user's awareness represent an important support for self-reflection aiming to help a user to adapt accordingly his/her daily routines.

9 LIMITATIONS

One limitation of our study is its focus on Android phone users. Given that this represented the majority of our users in our first cohort we chose this as a filtering criteria. Moreover, this study focuses on happiness, reflecting positive and negative well-being. Other measures such as anxiety, and stress should be examined in future extension of this work.

Another limitation of our study represents the limited amount of days used to determine outliers, both in sensor readings and in the target features, as we have a maximum of 14 recordings per user in the Diverse Cohort, and 21 measurements per user in the University Cohort. This led to limited implications on what would be the optimal threshold for labelling the outliers in sensor readings and target. Thus, for our study we could not conclude on what is the optimal threshold for the outliers labelling, which is why the different thresholds generate comparable results against each other. Further studies are needed to explore what would be the minimum amount of days required in order to have more solid takeaways about the thresholds for computing outliers as this may impact the practical application of the proposed approach. Finally, the cold start problem poses a challenge for the proposed approach, given that it requires a few recordings to be able to label outliers. A potential solution to this could be to utilise the population distribution while the minimum amount of days recordings are gathered.

10 CONCLUSION

A vast number of studies have explored predicting various user states, including subjective well-being and emotions from mobile sensing data. The studies demonstrated the proof of concept yet with a moderate accuracy in predictive user states from smartphone sensing data. In this paper, we propose a novel methodology for predicting outlier user states and evaluate our approach in a challenging use-case of predicting daily happiness. Shifting away from raw sensors, the method builds synthetic features as outliers with respect to the user's individual routines and learns to predict outliers in users' reported happiness. We showed that during an irregular day, the users' smartphone sensors patterns will be irregular as well. Results demonstrated that the proposed approach outperforms the other techniques, including traditional sensor-based predictive models. Finally, due to its robustness to individual characteristics across the user sample, this work brings benefits also when it comes to replicating user state models in different settings and with different demographics.

REFERENCES

[1] Ahmed M Abdel-Khalek. 2006. Measuring happiness with a single-item scale. *Social Behavior and Personality: an international journal* 34, 2 (2006), 139–150.
[2] Charu C Aggarwal. 2015. Outlier analysis. In *Data mining*. Springer, 237–263.
[3] P Allin. 2007. Measuring societal wellbeing'. Economic & Labour Market Review, Vol. 1 No. 10.
[4] Joost Asselbergs, Jeroen Ruwaard, Michal Ejdys, Niels Schrader, Marit Sijbrandij, and Heleen Riper. 2016. Mobile phone-based unobtrusive ecological momentary assessment of day-to-day mood: an explorative study. *Journal of medical Internet research* 18, 3 (2016), e72.
[5] Andrew L Beam, Arjun K Manrai, and Marzyeh Ghassemi. 2020. Challenges to the Reproducibility of Machine Learning Models in Health Care. *JAMA* (2020).
[6] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex Pentland. 2014. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *Proceedings of the 22nd ACM international conference on Multimedia*. 477–486.

[7] Andrey Bogomolov, Bruno Lepri, and Fabio Pianesi. 2013. Happiness recognition from mobile phone data. In *2013 International Conference on Social Computing*. IEEE, 790–795.

[8] Teodora Sandra Buda, Bora Caglayan, and Haytham Assem. 2018. Deepad: A generic framework based on deep learning for time series anomaly detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 577–588.

[9] V Chandala, A Banerjee, and V Kumar. 2009. Anomaly Detection: A Survey, ACM Computing Surveys. *University of Minnesota* (2009).

[10] Gokul Chittaranjan, Jan Blom, and Daniel Gatica-Perez. 2011. Who's who with big-five: Analyzing and classifying personality traits with smartphones. In *Wearable Computers (ISWC), 2011 15th Annual International Symposium on*. IEEE, 29–36.

[11] Marios Constantinides, Jonas Busk, Aleksandar Matic, Maria Faurholt-Jepsen, Lars Vedel Kessing, and Jakob E Bardram. 2018. Personalized versus generic mood prediction models in bipolar disorder. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. 1700–1707.

[12] Victor P Cornet and Richard J Holden. 2018. Systematic review of smartphone-based passive sensing for health and wellbeing. *Journal of biomedical informatics* 77 (2018), 120–132.

[13] Yves-Alexandre de Montjoye, Jordi Quoidbach, Florent Robic, and Alex Sandy Pentland. 2013. Predicting personality using novel mobile phone-based metrics. In *International conference on social computing, behavioral-cultural modeling, and prediction*. Springer, 48–55.

[14] E Diener. 1984. Subjective well-being. *Psychological Bulletin* 95, 3 (1984), 542–575.

[15] Ed Diener, Pelin Kesebir, and William Tov. 2009. Happiness. (2009).

[16] Edward Diener, Richard E Lucas, Shigehiro Oishi, et al. 2002. Subjective well-being: The science of happiness and life satisfaction. *Handbook of positive psychology* 2 (2002), 63–73.

[17] Paul Dolan. 2014. *Happiness by design: Finding pleasure and purpose in everyday life*. Penguin UK.

[18] Paul Dolan, Richard Layard, and Robert Metcalfe. 2011. Measuring subjective well-being for public policy. (2011).

[19] Afsaneh Doryab, Prerna Chikarsel, Xinwen Liu, and Anind K Dey. 2018. Extraction of Behavioral Features from Smartphone and Wearable Data. *arXiv preprint arXiv:1812.10394* (2018).

[20] William Fleeson. 2001. Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of personality and social psychology* 80, 6 (2001), 1011.

[21] Douglas M Hawkins. 1980. *Identification of outliers*. Vol. 11. Springer.

[22] Galen Chin-Lun Hung, Pei-Ching Yang, Chia-Chi Chang, Jung-Hsien Chiang, and Ying-Yeh Chen. 2016. Predicting negative emotions based on mobile phone usage patterns: an exploratory study. *JMIR research protocols* 5, 3 (2016), e160.

[23] Natasha Jaques, Sara Taylor, Asaph Azaria, Asma Ghandeharioun, Akane Sano, and Rosalind Picard. 2015. Predicting students' happiness from physiology, phone, mobility, and behavioral data. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 222–228.

[24] Natasha Jaques, Sara Taylor, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. 2016. Multi-task learning for predicting health, stress, and happiness. In *NIPS Workshop on Machine Learning for Healthcare*.

[25] Natasha Jaques, Sara Taylor, Akane Sano, and Rosalind Picard. 2015. Multi-task, multi-kernel learning for estimating individual wellbeing. In *Proc. NIPS Workshop on Multimodal Machine Learning, Montreal, Quebec*, Vol. 898.

[26] Daniel Kahneman and Alan B Krueger. 2006. Developments in the measurement of subjective well-being. *Journal of Economic perspectives* 20, 1 (2006), 3–24.

[27] Pelin Kesebir and Ed Diener. 2009. In pursuit of happiness: Empirical answers to philosophical questions. In *The science of well-being*. Springer, 59–74.

[28] 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), 88.

[29] Bernd Löwe, Kurt Kroenke, and Kerstin Gräfe. 2005. Detecting and monitoring depression with a two-item questionnaire (PHQ-2). *Journal of psychosomatic research* 58, 2 (2005), 163–171.

[30] Yuanchao Ma, Bin Xu, Yin Bai, Guodong Sun, and Run Zhu. 2012. Daily mood assessment based on mobile phone sensing. In *2012 ninth international conference on wearable and implantable body sensor networks*. IEEE, 142–147.

[31] Souneil Park, Aleksandar Matic, Kamini Garg, and Nuria Oliver. 2018. When Simpler Data Does Not Imply Less Information: A Study of User Profiling Scenarios With Constrained View of Mobile HTTP (S) Traffic. *ACM Transactions on the Web (TWEB)* 12, 2 (2018), 9.

[32] Eugenia Politou, Efthimios Alepis, and Constantinos Patsakis. 2017. A survey on mobile affective computing. *Computer Science Review* 25 (2017), 79–100.

[33] Abhishek Pratap, David C Atkins, Brenna N Renn, Michael J Tanana, Sean D Mooney, Joaquin A Anguera, and Patricia A Areán. 2019. The accuracy of passive phone sensors in predicting daily mood. *Depression and anxiety* 36, 1 (2019), 72–81.

[34] 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.

[35] John D Teasdale. 1983. Negative thinking in depression: Cause, effect, or reciprocal relationship? *Advances in Behaviour Research and Therapy* 5, 1 (1983), 3–25.

[36] Alina Trifan, Maryse Oliveira, and José Luís Oliveira. 2019. Passive sensing of health outcomes through smartphones: systematic review of current solutions and possible limitations. *JMIR mHealth and uHealth* 7, 8 (2019), e12649.

[37] Timothy J Trull and Ulrich W Ebner-Priemer. 2009. Using experience sampling methods/ecological momentary assessment (ESM/EMA) in clinical assessment and clinical research: introduction to the special section. (2009).

[38] Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Transactions on Affective Computing* 5, 3 (2014), 273–291.