Passive mobile sensing and psychological traits for large scale mood prediction

Dimitris Spathis
University of Cambridge
ds806@cam.ac.uk

Sandra Servia-Rodriguez
University of Cambridge
ss2138@cam.ac.uk

Katayoun Farrahi
University of Southampton
k.farrahi@soton.ac.uk

Cecilia Mascolo
University of Cambridge
cm542@cam.ac.uk

Jason Rentfrow
University of Cambridge
pj39@cam.ac.uk

ABSTRACT

Experience sampling has long been the established method to sample people’s mood in order to assess their mental state. Smartphones have started to be used as experience sampling tools for mental health state as they accompany individuals during their day and can therefore gather in-the-moment data. However, the granularity of the data needs to be traded off with the level of interruption these tools introduce on users’ activities. As a consequence the data collected with this technique is often sparse. This has been obviated by the use of passive sensing in addition to mood reports, however this adds additional noise.

In this paper we show that psychological traits collected through one-off questionnaires combined with passively collected sensing data (movement from the accelerometer and noise levels from the microphone) can be used to detect individuals whose general mood deviates from the common relaxed characteristic of the general population. By using the reported mood as a classification target we show how to design models that depend only on passive sensors and one-off questionnaires, without bothering users with tedious experience sampling. We validate our approach by using a large dataset of mood reports and passive sensing data collected in the wild with tens of thousands of participants, finding that the combination of these modalities has the best classification performance, and that passive sensing yields a +5% boost in accuracy. We also show that sensor data collected for the duration of a week performs better than when only using data collected for single days for this task. We discuss feature extraction techniques and appropriate classifiers for this kind of multimodal data, as well as overfitting shortcomings of using deep learning to handle static and dynamic features. We believe these findings have significant implications for mobile health applications that can benefit from the correct modeling of passive sensing along with extra user metadata.

ACM Reference Format:

1 INTRODUCTION

Experience sampling—which involves asking participants to report on their behaviors or environment on repeated occasions over time—has long been used as a mechanism to longitudinally assess the mental health of individuals by prompting them to report their mood using questionnaires traditionally delivered through pen and paper, but also through the web. Psychologists have used different tools or scales that facilitate users to assess their mood. These include the Positive and Negative Affect Schedule (PANAS) [47], a self-report questionnaire of two 10-item scales that measures both positive and negative affect; and the Affect Grid [33] scale, a 2-dimensional grid, where the x-axis indicates the feeling in terms of its positiveness or negativeness while the y-axis indicates its intensity. Independently of the scale used, timely and accurate mood report is important to anticipate clinical outcomes such as depression [8], longevity [44] or mortality [2].

The pervasiveness of smartphones and wearable devices has enabled timely delivery of experience sampling [11], allowing a near real-time detection of clinical outcomes and relapses. This led to the development of several mobile phone applications that prompt their users to assess and report their mood one or more times per day, using one or more different scales [36, 38]. Apart from potentially inducing biases in the measurements, interrupting users during their daily lives at a high frequency and with the same purpose is seen as a high burden by many users [27], as it is evidenced by the high dropout rates reported in these applications. Indeed, according to recent statistics, more than 2/3rds of people who download a mobile health app used it only once [22].

Previous research has pointed out the link between self reported mood and some personality traits such as emotional stability [9, 12]. Exploiting this link to track mental health would mitigate users’ burden, as assessing their personality as well as other psychological traits would only require one off questionnaires. At the same time, personal mobile devices come also equipped with a growing set of built-in sensors, such as an accelerometer, microphone and gyroscope. A proper and rigorous analysis of the data passively collected with these sensors provides valuable insights for the users’ physical behaviour [3], but could also act as a proxy of their mental health [45]. However, how to use psychological traits and passive sensing data to accurately track mental health is still an open research question. Also, the use of low sampling rates for passive sensing data collection due to battery consumption issues often lead to very sparse sensing data, which adds to the challenge.

The penetration of mobile devices has also introduced scale: many more individuals can now be reached and assessed. For example, in a hospital environment, mobile experience sampling enabled the collection of 11,381 survey responses over a 12-month period from 304 physicians and nurses, completed with minimal initial training [40]. Mobile sensors enable researchers to collect not only the explicit reports of the participants, but also the context in which
Rising Stars track

Spathis D. et al.

Figure 1: Aggregate 735,778 self-reported mood scores in the Emotionsense dataset collected from 17,251 users. Most users report neutral (around 0.5,0.5) and calm-happy (down right quadrant) mood on the affect grid (a). The two multi-modal distributions (Pearson r = -0.23, p < 0.00001) of the mood(b).

Figure 2: The mood tracking application. Users can report their mood in an affect grid, complete personality and other questionnaires.

these answers were provided. Indeed, a recent survey of 110 papers concluded that a total of 70 studies (63.6%) passively or actively collected sensor data from the participants’ study device [42]. On a larger scale, Utsureko [38] and Emotionsense [36], two different smartphone applications for mood monitoring through self-reports were used by more than 24,000 and 17,000 users, respectively. However, most of the studies on investigating the use of smartphones to track and improve mental health and well-being have been conducted through controlled experiments, and limited number of participants and observations [20, 23, 34, 46]. Conducting such studies in the wild would allow reaching many more participants, broadening the significance of the findings. However, the absence of rigid control over participation and the limited mechanisms to promote engagement, make the data collected noisier and sparser than in controlled setups, and it is unclear whether previous findings and methodologies can be transferred to these large natural datasets. Robust methodologies for anticipating clinical outcomes and relapses using very sparse data are key to the widespread adoption of smartphones as tools to provide mental health support.

Mobile sensing applications often require inputs from sensors in the form of high-dimensional time-series, coming from accelerometers, gyroscopes, microphones or other user-generated data [19]. However, these sensor measurements are quite noisy and although for some purposes simple first-order features have proved to be effective, it is not straightforward how to select robust features from different noise levels of individual user behaviors, since every user introduces different levels of noise according to its device, environment etc. For example, the MoodExplorer study [48] extracted the mean, variance, and signal-to-noise ratios from the microphone sensor, while the Emotionsense study [36] calculated the standard deviation of the magnitude of acceleration from the three axes (x, y, z) of the accelerometer. Noise in mobile measurements is hard to model because it is correlated over time [30] and presents a non-linear structure [4].

In this paper we investigate whether individuals’ perceived mood can be obtained through their psychological traits collected through one-off questionnaires, as well as passively collected mobile sensing data, thus avoiding sending frequent experience sampling questionnaires. More specifically, we investigate whether these psychological traits and passive sensing data can be used to detect individuals whose general mood deviates from the common relaxed mood distinctive among mentally healthy individuals [33]. To do so, we propose a machine learning methodology to classify individuals according to their general mood, that takes as inputs sparse answers to one-off surveys covering different profile-related characteristics of the individuals, as well as features extracted from noise and sparse accelerometer and microphone sensors readings passively collected with their smartphones. We evaluate our methodology using a large-scale dataset of mobile sensing and self-reported data collected in the wild for more than 3 years and that contains data from more than 17,000 participants. We conduct extensive experimentation by training over 100 models in order to find out the
best combination of modalities. We also conduct extensive first and second-order feature extraction from the sensor time-series.

This paper makes the following contributions:

- We conducted an extensive data exploration of the self-reported moods provided by 17,251 of the users of an experience-sampling based smartphone applications, with the aim of identifying the most common reporting behaviour so as to characterize mentally healthy individuals in the context of our research. Our findings showed that the majority of the population in our dataset reported feeling, on average, relaxed (down-right side of the affect grid), which is in line with previous research [33].
- We provide a supervised learning methodology to detect individuals whose general mood deviates from the common relaxed mood distinctive among mentally healthy individuals [33]. Our methodology does not involve any kind of cumbersome experience sampling, but only uses one-off questionnaires (demographics, personality, etc.) as well as sparse and noisy passive sensing data collected with the accelerometer and microphone sensors of individuals’ smartphones.
- We performed an extensive evaluation of our methodology using a large scale dataset collected in the wild. Our results showed that the combination of one-off questionnaires and passive sensing data gives the best performance in mood prediction. Indeed, by adding passive sensing data we achieve a +5% in accuracy (75% in absolute) with respect to only using questionnaires.

These findings have the potential of informing future developers of mobile health applications as well as psychologists on how to properly use one-off questionnaires and passive sensing data for the early detection of symptoms of mental disorders at scale.

2 THE PROBLEM AND THE DATA

Mobile health applications aimed at assisting users with their mental health so as to prevent clinical outcomes should minimize the burden to the user so as to increase adherence and satisfaction with the app. Instead of the timely and continuous collection of mood self-reports, psychological traits obtained through one-off questionnaires, as well as passive sensing data, should be preferred in order to design effective and useful applications. Our aim in the rest of this paper is to investigate how psychological traits and passive sensing data can be used to detect individuals who might not feel mentally well, i.e., users who have been reported moods that deviate from the general reports of the population.

To do so, we first conduct an exploratory analysis of the mood reports provided by more than 17,000 individuals for a period of more than 3 years, in order to identify the most common set of mental states (moods) reported by any of these individuals (Section 3). Given the scale and the in the wild nature of the data collection, we believe our results are general enough to be representative of the whole population. We then use these findings as the ground truth to validate our machine learning methodology to identify individuals whose record of reported moods deviates from that of the majority, by only using one-off questionnaires and passive sensing data (Section 4). We provide further details of the data used in our analysis and experiments in the rest of this section.

2.1 The data

We use the Emotion Sense dataset [36], a dataset that contains sensor and self-reported data collected with a mobile phone application for Android (Fig. 2) designed to study subjective well-being and behavior. From February 2013 until October 2016, this application collected 735,778 self-report data from 17,251 users, through surveys presented on the phone via experience sampling, and behavioral data from physical and software sensors in the phone (accelerometer, microphone, location, text messages, phone calls, etc.). The participants signed a consent form that restricts the use of the data to the University of Cambridge researchers, according to the Institutional Review Board (IRB). For this analysis, we consider self-reported mood collected graphically using the Affect Grid [33], profile-related surveys, as well as sensed data collected with the accelerometer and microphone sensors. Twice per day, between 8AM and 10PM and with a difference of at least 120 minutes apart, participants received a notification asking them to report their mood in the affect grid (Figure 1). Meanwhile, sensed data were collected passively in the background at different moments during the day depending on the different versions of the application. At different stages of the application, participants were requested to complete profile-related questionnaires covering a broad range of topics: demographics, personality, gratitude, health, sociability, job satisfaction, life aspirations and connectedness, where the questions were answered using Likert scales. Below we describe the specific data we use in our experiments.

Experience sampling. The Emotion Sense application for mood monitoring prompted their users to report, twice per day, how they felt using an Affect Grid scale. Figure 1 shows the aggregate of mood self-reports for all the users of the application, where the down-right quadrant, corresponding to relaxed mood, is the most densely populated, a result that matches previous studies in the area [33]. Due to the in the wild nature of the data collection, users did not always report their mood even if they were prompted to do so, which might be consequence of the burden that experience sampling brings to the users. In more detail, Figure 3 shows the CCDF of moods reported per participant, included the ones they were expected to do given the time they were using the app, the ones they were prompted to do but they did not (missed) and those that
they actually did so. Thus, alternatives to experience sampling are required to design effective, long-term, mobile health applications for mental health. As we will show later, by using the reported mood as a classification target we can design systems that depend only on passive sensors and one-off surveys.

**One-off questionnaires.** Previous research has found a link between self reported mood and personality traits such as emotional stability [9, 12]. However, to the best of our knowledge, it is not clear yet how to use personality, and other psychological traits, to detect potentially mentally unhealthy individuals. In the Emotion Sense dataset, a subset of the users (12,106, 70% of total) completed some one-off surveys providing information regarding their demographics, personality, gratitude, health, sociability, job satisfaction, life aspirations, connectedness, and satisfaction with life.

**Passive sensing data.** Data collected through the built-in accelerometer sensor of our smartphones provide valuable insights into our activity level throughout the day. At the same time, previous research has demonstrated the link between activity level and happiness [21, 36]. We hypothesize that our activity level throughout the day has a high impact on how we feel on that day and therefore also use these sensing data in our experiments. In the Emotion Sense dataset, accelerometer samples consist of \([x, y, z](m/s^2)\) axes data for periods of 5, 8 or 10 seconds, collected at different intervals throughout the day depending on the version of the application. Microphone samples, on the other hand, provide insights into the noise level in the user’s environment. As with activity, we hypothesize that how we feel (our mood) influences is influenced by the kind of places or environments we visit and the level of noise in these spaces. Therefore we use this in our experiments. To preserve privacy, the Emotion Sense application only recorded the amplitude level of noise at 20Hz for periods of 5, 8, 10 seconds at different intervals throughout the day depending on the version of the application.

Varied amounts of data are available for each of the sensors and self reports, mainly due to the uncontrolled way in which users were recruited. Also, the in the wild nature of the data collection makes the available data noisy and sparse, which adds to the challenge. We present more details on how we dealt with these noisiness and sparseness, as well as on the number of participants and days of sensed and self reported data used for each analysis, later on Sections 3 and 4.

3 FINDING GROUPS OF USERS FROM SELF-REPORTED MOOD TRAJECTORIES

The main goal of our research is on investigating whether psychological traits and passive sensing data can be used to identify users whose set of mood reports deviates from those of the general population, which might be indicative of some mental condition. Fig 1a shows a visualization of the aggregation of self reports provided by the users in the Emotion Sense dataset, where the most common mood reported is in the down-right side of the affect grid, corresponding to the relaxed mental state. However, it is not clear whether to fix the boundaries on the affect grid. We propose not to hard code the thresholds and potentially inducing biases in our labels, but instead relying on clustering techniques to make labels naturally emerge from the data. The rest of this section describes in detail the methodology to label users into relaxed/non-relaxed in the Emotion Sense dataset.

3.1 Methodology

A mood self-report in the affect grid is described by means of two coordinates: the x-coordinate that indicates the feeling in terms of its positive and negative and the y-coordinate indicates the intensity of alertness. The history of mood-reports of an individual consists of time-series trajectories of \([x,y]\) tuples recorded over time in the affect grid. Also, the noise and sparseness of an in-the-wild setup result in that (i) the number of self-reports reported by different individuals might be different, and (ii) that for a given individual, the reported moods might not be consecutive (as a consequence of users missing reports). In order to cope with this variability and obtain independent features to allow clustering algorithms to learn representative clusters, we extract 8 simple features for each axis or coordinate, namely \(\text{counts, mean, std, min, max and quantiles (25%, 50%, 75%)\},\) resulting in 16 final features for every user. Missing values are replaced with zeros and \(\text{minmax [0,1]}\) normalization is applied to the final features column-wise. Due to the sparsity of the mood and the power law distribution of the counts, these two count features that measure non-missed reports are affected the most by the normalization, concentrating all their mass close to zero.

We then apply the \(k\)-means [25] clustering algorithm to produce mutually exclusive clusters of spherical shapes based on distance. In order to come up with the optimal number of clusters, we conduct the Elbow method [41] where we increase the number of clusters and observe the drop of the evaluation metric. Here, we use the silhouette metric [32] which measures how similar a sample is to its own cluster compared to other clusters.
3.2 Findings

We applied our methodology to identify non-relaxed users (or those that deviate from the most common mood feeling reported) in the Emotion Sense dataset. For each of the 17,251 users that have reported their mood at least once, we obtain 2,682 sparse mood reports completed over 3 years, for valence and arousal. This is the final sample we used for this experiment.

**Exploratory analysis.** As a first exploratory analysis, we apply hierarchical clustering to the historical mood of the users. Figure 4 shows the resulting trees. Specifically, the y-axis shows the cluster of users whereas the x-axis the cluster of features (16 features, 8 per valence and 8 per arousal). We observe that there are multiple user groups shown on the left side tree, pointing out that some mood reporting behaviors resemble other users. However, it is not easy to spot clear relationship due to the number of users. The features are also clustered with the most prominent 2 groups being the valence and arousal. However, there are some intruders in those clusters: for example, the maximum arousal (max_y) belongs to the valence cluster while the counts (counts_x) and the minimum (min_x) of valence goes into the arousal group. These feature clusters provide hints regarding the non-linear relationships of the mood components.

**k-means.** We now apply k-means to obtain the labels to use in our experiments. We repeat the experiments by varying k, the resulting number of clusters in order to visually identify them in the affect grid. Figure 5 shows the resulting clusters when increasing the number of clusters from 2 to 4. For 2 clusters (Fig. 5d), by plotting the mean valence and arousal in the affect grid, we notice a group of consistently relaxed users on the down-right quadrant and another group that consists of depressed, stressed and excited users on the rest of the grid. When we further increase the number of clusters, things get more complicated for pattern finding. For example, with 3 clusters (Fig. 5e) we spot a central neutral group which is now distinct, while the rest is similar to the previous plot.
We now describe our methodology to identify non-relaxed with seven clusters or more. These two groups will be used as a individual from their psychological traits obtained through one-off satisfaction with life. These 92 features are represented as Likert-scales file attributes such as demographics, personality, gratitude, health, Questionnaires.

4.1 Feature extraction

features are finally fed to classifiers. We detail these steps below. To obtain feature combinations with the maximum variance. These univariate Mann-Whitney U test [26], and then we transform these feature selection, where we first calculate the feature significance regarding users’ psychological traits. We then perform a two-step as one-hot encoding the answers to the one off questionnaires re-follow the workflow in Figure 8, where we begin by extracting features. Intuitively, this means that the clustering algorithm found 2 centers in the high-dimensional space and we just plot the values for every feature of these points. For instance, for 2 clusters (Fig. 5a) the largest distance seems to be between maximum valence as well as the minimum arousal. These two features could be enough to separate the two clusters. By moving up in the number of clusters, things get more complicated since we have to find features for which all the features reside equally apart. Namely, for 3 clusters (Fig. 5b) minimum valence seems to be different for every group, while for 4 clusters (Fig. 5c) there is not a single feature with different centroids.

Finally, we perform the elbow method to quantitatively find the optimal number of clusters. Figure 6 shows that the top silhouette score is 0.30 (higher is better) with two clusters while it goes down 0.23 with three clusters. We observe that it plateaus around 0.20 with seven clusters or more. These two groups will be used as a label in the machine learning pipeline to infer non-relaxed users from one-off questionnaires and passive sensing data in the next section. We are aware that these clusters are inferred information and thus could include some errors, however we incorporate the silhouette score with the lowest error. Please note that there is a class imbalance between the clusters on the user level: cluster 1 (65%), cluster 2 (45%), which we will address later in that section.

4 PSYCHOLOGICAL TRAITS AND MOBILE SENSING TO PREDICT NON-RELAXED MOOD

We now describe our methodology to identify non-relaxed individuals from their psychological traits obtained through one-off questionnaires, and passive sensing data collected using the accelerometer and microphone sensors of their smartphones. We follow the workflow in Figure 8, where we begin by extracting features from the accelerometer and microphone raw data, as well as one-hot encoding the answers to the one off questionnaires regarding users’ psychological traits. We then perform a two-step feature selection, where we first calculate the feature significance of a real-valued feature to a binary target as a p-value using the univariate Mann-Whitney U test [26], and then we transform these selected features with Principal Component Analysis (PCA) [31] to obtain feature combinations with the maximum variance. These features are finally fed to classifiers. We detail these steps below.

4.1 Feature extraction

Questionnaires. One-off surveys cover a wide range of a user profile attributes such as demographics, personality, gratitude, health, sociability, job satisfaction, life aspirations, connectedness, and satisfaction with life. These 92 features are represented as Likert-scales or categories. In order to be appropriate for machine learning models, the categorical features are transformed to individual features with one-hot encoding; so that a feature with e.g. 3 possible choices (Yes, No, missing), is transformed to 3 different features. Categorical features include the gender, age group, education level and ethnic group among others. The total list of questionnaire features is 131.

Accelerometer. We consider the 3 (x,y,z) dimensions of the accelerometer and compute the magnitude of the acceleration for 5, 8, and 10-second samples, resulting in 48 time-steps for every user-day (336 time-steps for every user-week). We aggregate the sensor in 30-min bins since this level of granularity is the best trade-off between data sparsity and modeling the sub-hourly movement of individuals. By doing this light processing, we end up with one time-series instead of three, combining the three axes into one time-series. Based on the sparsity histogram (Fig. 7b), we filter those samples that have at least 50 time-steps during the week (20 time-steps during the day). This time-series is normalized with minmax scaling to a [0.05-1] range and the missing values are replaced with zeros. We extract 721 simple and second order features that cover a wide range of attributes of a sensor such as the energy, autocorrelation, entropy, trends, wavelet and Fourier coefficients, peaks, etc. For a comprehensive list of the features we refer the reader to the documentation of the tsfresh library [10].

Microphone. Similarly with the accelerometer data, we compute the mean of the 5, 8, and 10-second window over the initial raw microphone data over the amplitude level of noise at 20 Hz, ending up with 48 time-steps for every user-day (336 time-steps for every user-week). We apply the same filtering, normalization and feature extraction as the accelerometer above, resulting in 717 features.

Seasonality. Temporal features are extracted by the end of the sensor user-week time-stamp in order to capture the inherent seasonality patterns. Namely, we compute these 5 increasingly detailed time-aware features: the number of the quarter, month, week, day of week and hour of day. We consider these features to belong to the sensor modality that we introduce later.

4.2 Classifiers

We considered three different classifiers for our inference task: Logistic Regression, Gradient Boosting Trees and a Deep Neural Network. Below we describe the details of our implementation.

Logistic Regression (LR). An sklearn implementation of a binary logistic regression, with penalty of L2 regularization along with a $C = 1$ (inverse of regularization strength), was tested.

Gradient Boosting Trees (GB). An sklearn implementation of a gradient boosting was tested. Reportedly the state-of-art in feature-based machine learning [29], this classifier forms an ensemble of weak prediction models, typically decision trees.

Deep Neural Network (NN). We use a straightforward bottleneck architecture of 4 feed forward Dense layers of dimensionality 100-50-100. The reduced dimensionality in the middle (50 units) has been shown to lead to better generalization in deep learning architectures [14, 24]. A rectified linear unit (ReLU) [13] activation is applied at the output of every layer, followed by a batch normalization layer that transforms the output to have zero mean and unit variance [15]. Dropout of 50% probability is applied to every layer to reduce overfitting [37]. The final layer performs a softmax
Rising Stars track

Figure 7: Sparsity analysis of the sensors. Missing values for the sensors on the weekly (a,c) level. Cumulative distribution functions (CDF) for the missing time-steps (b,d) show the long tail distribution of sparsity. Some weekly periodicity is also spotted. Similar conclusions are drawn with the daily level sensors.

Figure 8: Workflow of the data processing and model development for the mood prediction task.

5 EVALUATION

We now detail the evaluation of our methodology to identify non-relaxed users from one-off questionnaires and passive sensing data described in Section 4. We used the Emotion Sense dataset, for our experiments, and the clustered mood we obtained using k-means in the Section 3 as the labels for the classifiers. Below we indicate how we merged the data from the different modalities and how we partition the dataset for our experiments in Section 5.1. Findings and results are provided in Section 5.2.

5.1 Experimental setup

Modality merge. Experiments in the wild such as this one do not guarantee complete and fine-grained data, especially when they involve battery consuming tasks like sensor-tracking or input-based prompts like self-reports from users. Therefore, not all modalities appear for the same users. We start by merging the accelerometer and microphone modalities resulting in 141,261 user-weeks while we concatenate their features along with the seasonality ones. Then, we find which users from those weeks have completed at least a single questionnaire and concatenate these static features to the feature vector, resulting in 131,793 user-weeks. Finally, we merge with the clusters that we produced in the previous section, so that every user-week feature vector corresponds to one of the two user mood clusters. Please note that these clusters came up by taking into account the full mood history of the users and therefore we do not imply that mood is static. Apparently, the high class imbalance on the user level earlier is exaggerated here because only 7% of the user-weeks belong to cluster 2 (green in Fig. 5). As a result, we subsample the majority class, resulting in 18,998 balanced user-weeks from 2,812 users. The same processing is followed for the daily sensors: 112,161 user-days after sensor merge, 106,672 after questionnaire merge, and we end up with 16,470 user-days from 1,859 unique users when we merge with the labels and sub-sample.
We now present the classification results of predicting whether a user’s weekly/daily sensors and questionnaire metadata belong to the same mood cluster group (relaxed or not) based on sensors, questionnaires and other meta-data. As discussed earlier, we performed extensive experiments and trained 180 models to evaluate the impact of the different modalities and user splits. In Table 1, we present the mean classification performance of the experiment setup described in the previous section, that of predicting the mood cluster group (relaxed or not) based on each user’s weekly/daily sensors and questionnaire metadata.

<table>
<thead>
<tr>
<th>Modality</th>
<th>Weekly AUC</th>
<th>Daily AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensors (S)</td>
<td>0.575 (0.03)</td>
<td>0.550 (0.04)</td>
</tr>
<tr>
<td>Questionnaires (Q)</td>
<td>0.690 (0.05)</td>
<td>0.687 (0.09)</td>
</tr>
<tr>
<td>All (S + Q)</td>
<td><strong>0.749 (0.06)</strong></td>
<td><strong>0.725 (0.06)</strong></td>
</tr>
</tbody>
</table>

The sole use of sensors achieves less than 60% for all the models. The NN presents similar performance for the combined (with a lower max AUC in the combined representation though). By using the sensors on the week level we achieve 75% AUC by classifying users into relaxed or not. Also, we observe that by tracking the users for more time (week over day level), we achieve better performance. In hindsight, this is intuitive since movement and noise levels are more time (week over day level). Beyond the binary task, the GB model achieves an AUC of 0.740, while the LR comes second with 0.706. The NN presents similar performance for the combined and questionnaire representation, hinting that the daily sensors do not contribute much for it. However, the rest models show a rise of +1.1% (GB) and +3.5% (LR) in AUC when we add the sensors to the questionnaires.

### Discussion
These results show that by adding passive sensing to traditional personality and demographics surveys we are able to predict the mood group of individual users with a higher precision. Specifically, for our task we achieve ~ 75% AUC by classifying users into relaxed or not. Also, we observe that by tracking the users for more time (week over day level), we achieve better performance. In hindsight, this is intuitive since movement and noise levels are expected to be related with relaxation levels. Beyond the binary task, extra experiments with 3 or 4 clusters (multi-class) yielded worse results due to the significant cluster overlap and less data-points per class to learn. Last, putting our results in the context of related work we see that similar datasets yield lower accuracy (around 65%) for slightly different tasks such as predicting tomorrow’s mood or daily mood average.

### 6 RELATED WORK
As noted in one of the first seminal review papers in 2010 [19], the main obstacle to the field of mobile sensing and pervasive health is not lack of adoption, since billions already carry sensor-rich devices, but rather on how to perform privacy-aware and resource-sensitive reasoning with noisy and missing data, and to deliver effective interventions. When these issues are solved, mobile sensing will act as a catalyst for diverse domains such as social networking, health, and energy. Here, we focus on the challenges regarding learning robust and informative features from noisy signals and how they can assist with user modeling and interventions.

While the motivation for building mood prediction systems seems well-founded, the implementation thereof appears to be challenging. Numerous mobile apps for mental health monitoring have been proposed, like BeWell [20], or MoodScope [23]. Specific groups,
like undergrad students, have been studied in controlled setups, e.g. StudentLife [46] measured the impact of student workload on stress with sensors and self-reports, whereas Snapshot [34] tracked mood and sleep. Other efforts have focused on detecting depression by tracking medication, sleep patterns and actions [38], location [6], or even keypress acceleration [7]. Like in our case, static personality metadata have been combined with sensor time-series [5]. Please note that the paper that introduced this dataset [36] also predicted mood by using smartphone sensor data, but used a smaller subset of users and most importantly a different prediction target (mood at time $t$ with data sensed before and after $t$).

One of the most important limitation of the above works is the relatively small sample size, with participants often belonging to similar socioeconomic backgrounds, in order to draw robust conclusions. Besides, participants were often tracked during a short period of time and in controlled setups. For instance, the MoodScope study [23] monitored 32 people over 2 months, the StudentLife project [46] tracked 48 students over 10 weeks, whereas Snapshot [34] is probably the biggest general published study with 206 students tracked for over 1 month. In contrast, we draw robust conclusions from an initial dataset of more than 17,000 users, collected in the wild for more than 3 years.

Putting aside the limitations of the sample size, perhaps the most closely related work to ours is the Snapshot [1] study. This study investigated how daily behavior gathered through passive sensing data influence sleep, stress, mood, and other wellbeing-related factors. Multiple papers focused on different aspects of the collected dataset, such as personalization with multi-task learning to predict tomorrow’s mood, stress, and health [39], prediction of happy/sad mood based on sleep history [35], or a denoising autoencoder to fill in missing sensor data for mood prediction [16]. Similar to us, they first cluster the users before going into classification [39], although their goal here is to provide personalized predictions to these clusters. However, our models do not distinguish between healthy and depressed patients, but predict the clustered mood group which roughly correspond to relaxed or not-relaxed users. From a more practical perspective, personalized models are difficult to be deployed on a real world scenario, since they require training $N$ personalized models, with $N$ being the number of users. Even though previous research has shown that better performance can be achieved by averaging the individual model accuracies [6, 23], no results are reported on unseen disjoint users. Instead, we provide single end-to-end trainable models while in all of our experiments we report performance from a disjoint user set that the model has not seen during training.

The majority of related literature has applied some kind of supervised learning algorithms, like Logistic Regression or Support Vector Machines, without focusing on systematic first and second order feature extraction from the sensors. The only alternative seems to be using some kind of deep learning which although yields moderate results (e.g. StudentLife dataset with deep feed-forward neural networks [28]). Other neural approaches include the Deep-mood paper that uses RNNs for depression prediction [38]. We build upon this growing piece of literature of employing machine learning on mood prediction by proposing end-to-end models that exploit a thorough feature extraction of the sensors as well as well rich information about the demographic and personality data of the users.

7 CONCLUSION

The pervasiveness of smartphones have converted them into experience sampling tools to collect people’s mood so as to assess their mental state. However the granularity of the data needs to be traded off with the level of interruption these tools introduce on users’ activities, which often results into very sparse data. In this paper we propose a machine learning methodology to detect if an individual perceived mood differs from that of the general population, by solely considering their psychological traits collected through one of questionnaires and passively collected mobile sensing data, thus avoiding the use of experience sampling questionnaires.

We evaluate our methodology by using a large-scale dataset collected in the wild for more than 3 years and 17,000 participants. An exploratory analysis of the data revealed that relaxed is the most common state reported by our population. Our experiments also confirmed that our methodology is able to distinguish between generally relaxed/non-relaxed individuals with a 75% AUC when using a combination of weekly sensors (accelerometer and microphone) and one-off questionnaire data (personality, demographics, etc) as inputs. Besides, the use of passive sensing data yields a +5% boost in accuracy. In healthcare context, this accuracy states that we can group users 3 out of 4 times correctly using only short-time mobile phone sensing and sparse surveys. While that level of accuracy might not be adequate for medical deployments, our focus is mostly on the positive contribution of passive sensing.

As future work, we plan to study data imputation techniques in order to ameliorate the significant data loss while merging the modalities [16] as well as focus on feature importance analysis. Also, in our current setup we use the aggregate approach of target clusters of users which someone can argue that might change over time; we are working on continuous predictions of both the sensors and the mood predictions. We also plan to adapt models that operate on raw time-series such as WaveNet [43] and combine them with multi-modal approaches for the static features.

ACKNOWLEDGMENTS

This work was partially funded by the Embiricos Trust Scholarship of Jesus College, Cambridge and the EPSRC Doctoral Training Partnership (grant reference EP/N509620/1).

REFERENCES
