Peers know you
A feasibility study of the predictive value of peEr's observations to estimate human states
Berrocal, Allan; Wac, Katarzyna

Published in:
Procedia Computer Science

DOI:
10.1016/j.procs.2020.07.031

Publication date:
2020

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY-NC-ND

Citation for published version (APA):
The 17th International Conference on Mobile Systems and Pervasive Computing (MobiSPC)  
August 9-12, 2020, Leuven, Belgium  

Peers Know You: A Feasibility Study of the Predictive Value of Peer’s Observations to Estimate Human States  

Allan Berrocal\textsuperscript{a}, Katarzyna Wac\textsuperscript{a,b,}\textsuperscript{*}  

\textsuperscript{a}Quality of Life Technologies Lab, Institute of Service Science, University of Geneva, Switzerland  
\textsuperscript{b}Quality of Life Technologies Lab, Department of Computer Science, University of Copenhagen, Denmark  

Abstract  
This paper examines the predictive value of peer’s observations of an individual, applied to computational models of certain states of that individual. In a study of 28 days, 13 participants provided self-assessments about their level of stress, fatigue, and anxiety, while their smartphone passively recorded the sensor’s data. Simultaneously, their designated peers provided assessments about the level of stress, fatigue, and anxiety they perceived from the participant using the PeerMA method. We extracted sensor-derived features (sDFs) from the participant’s smartphone, and peer-derived features (pDFs) from the peer’s assessments. We evaluated the pDFs on a binary classification task using three machine learning algorithms (Decision Tree-DT, Random Forest-RF, and Extreme Gradient Boosting-XGB). As a result, the classification accuracy consistently increased when the algorithms were trained with the sDFs plus the pDFs, compared the tradition of using only the sDFs. More importantly, the classification accuracy was the highest when we trained the algorithms only with the pDFs (73.3\% DT, 73.7\% RF, and 71.1\% XGB), which represents a unique contribution of this paper. The findings are encouraging about the incorporation of peer’s observations in machine learning with potential benefits in the fields of personal sensing and pervasive computing, especially for mental health and well-being.  

© 2020 The Authors. Published by Elsevier B.V.  
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)  
Peer-review under responsibility of the Conference Program Chair.  

Keywords: Peer-ceived Momentary Assessment; PeerMA; Ecological Momentary Assessment; machine learning; well-being  

1. Introduction  
Thanks to the sensing capacities of mobile, wearable and ubiquitous devices, researchers are using them \cite{19} to passively collect raw data to model individual states such as mental health and educational outcomes \cite{29}, stress \cite{12, 7, 22}, depressive mood \cite{25}, or schizophrenia \cite{28} among others. In general, passively sensed data is converted into informative features to create computational models, commonly using machine learning (ML) algorithms to make predictions about individual states. The ground truth about these modeled individual states is commonly obtained  

* Corresponding author. Tel.: +41-22-379-024  
E-mail address: katarzyna.wac@unige.ch  

1877-0509 © 2020 The Authors. Published by Elsevier B.V.  
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)  
Peer-review under responsibility of the Conference Program Chairs.
using Ecological Momentary Assessment (EMA) [24], which provides quantitative scores used as labels to train supervised (or semi-supervised) computational models for classification or regression tasks [19]. Despite its value, passively sensed data does not always enable accurate modeling of highly subjective individual’s perceptions[10], and its use may pose privacy risks [9]. Besides, smartphone sensors data may vary in time due to hardware or sensing platform differences [19]. We also know that individuals tend to abandon smart devices after a brief use [16] for reasons such as poor fit to expectations, not perceiving any direct value from the collected data, or the perceived high maintenance (e.g., battery charging), especially if these devices are not their own smartphones.

Our research evaluates the Peer-ceived Momentary Assessment (PeerMA) method [5, 4], which is a form of EMA, completed by a designated peer of an individual (defined as a close, trusted friend, or family member), during the same time-observation window when the individual is prompted to complete an EMA. Using PeerMA, trusted peers report their perception of the states for an individual with momentary and ecological validity, potentially enabling peers to become a source of information to complement the individual’s self-assessments [15, 1]. However, it is unknown if the data gathered via PeerMA about an individual is informative as an input to computational models trained to estimate certain states of that individual. As PeerMAs are paired with EMAs, we evaluate if PeerMA derived features (denoted “pDFs”) could be combined with smartphone-sensors derived features (denoted “sDFs”) to be exploited by certain ML algorithms to predict the state of the given individual (self-reported via EMA).

In this paper, we conduct an empirical investigation of whether or not pDFs improve the accuracy of a ML algorithm classifying individual’s physical and mental states (stress, fatigue, anxiety, and well-being). PeerMA derived features are the novelty aspect of this research, since they are extracted from someone else’s assessments of the participant’s momentary state. More specifically, we examine the following research questions:

**RQ.1** Does the classification accuracy improve when using pDFs, compared to using only sDFs? And if so,

**RQ.2** What are the most predictive pDFs during the classification tasks?

This paper describes background and related work in Section 2, followed by the study design in Section 3. Then, Section 4 outlines the data preparation process, followed by a description of the data analysis process in Section 5, and a discussion of the main results in Section 6. Finally, we conclude the paper and outline future work in Section 7.

### 2. Related Work

In behavioural sciences, Vazire [27] provided empirical evidence of how informant-reports (a.k.a. other / proxy / observer assessment-report) improve the validity of personality assessments, and how they enabled research on questions for which self-reports alone would be insufficient. Balsis et al. [1] used the self and informant-report versions of the NEO-PI-R inventory for personality and health assessment, and showed that informant-reports had greater internal consistency than self-reports. Although, that alone does not imply that informant-reports were more valid than self-reports, their internal consistency made them better predictors of overall health measures than the self-reports. Watson et al. [30] studied the acquaintance effect in the context of self vs. other agreement in low visibility aspects such as affect traits (like attentiveness or serenity) measured with the PANAS-X and Big Five instruments. Their analysis showed that self-other agreement was high among married couples, and moderate to high in several components of the scales for dating couples and friendship dyads. Similarly, Krom et al. [15] showed that spouses and friends were able to discriminate between eight distinct sources of chronic stress of a target person.

Regarding smartphone-based assessments, Harari et al. [11] reviewed studies using Smartphone Sensing Methods to identify physical movement, social interactions, and other daily activities as objective and automated proxies of behaviour. Others have also used smartphone sensing to model individual states, such as mental health and daily stress [12, 7, 22], academic performance [29], or even depressive moods [25] or schizophrenia [28].

Because peer-based assessments have not been used in the context of personal sensing, this paper contributes with the first exploration of their predictive value towards modeling individual human states with potential implications for personal health, ubiquitous technologies, and mobile human-computer interaction.

### 3. Study Design

To answer our research questions, we conducted an in-the-wild study (i.e., in the participants’ daily life context) during the autumn of 2018. The study lasted 28 days during which 13 participants, recruited at the University of
Geneva (UNIGE) in Switzerland, self-assessed, during the day, their perceived level of stress [17], fatigue [23] and (state) anxiety [14] using EMA, while their peers assessed the level of stress, fatigue, and anxiety perceived from the participant using the PeerMA method. The study is part of the exploratory (observational) phase of our research. The Institutional Review Board of the UNIGE approved the study under protocol N. CUREG.201807.

**Participants and Tools:** We recruited 13 (6♀, 7♂) adult participants around the campus of the UNIGE using flyers, email distribution lists, and word of mouth. Six participants had one peer, and seven had two peers, totaling 20 (8♀, 12♂). Those who completed the study entered a raffle for two Amazon Gift Cards worth 50.00 CHF.

We implemented the PeerMA method with the *mQoL Peers*¹ application using the *mQoL Lab* platform [8, 4]. To begin, participants had a 15-minute meeting with a researcher to (1) explain the study, (2) sign the informed consent, (3) train the participant to use the app, (4) answer any question from the participant. We explained that peers had to be individuals with whom they were regularly in contact (at least daily, face to face, or via communication tools) like a spouse, close relatives, or friends from school or work. We had no interaction with the peers. After enrolling peers, participants began the study and received daily EMAs and PeerMAs (Section 3).

**Subjective Variables:** The *mQoL Peers* app triggered single-item questions proposed by [21] on a visual analog scale as shown in [5]. We chose stress, fatigue, and anxiety because they often occur among healthy adults [2], compromising a person’s well-being (e.g., high stress, bad sleep quality) [6] and they can be studied with introquestive methods [26]. Besides, they are not trivially observable by peers, and presumably, early detection of these conditions could inform diagnosis, or therapeutic decisions. Table 1 describes the subjective variables collected from participants and peers. Peers also indicated how confident they were with each assessment [5], which informs the data analysis.

**Objective Variables:** The *mQoL Peers* app [8, 4] collects passive data reflecting the participant’s daily usage of the smartphone (summarized in Table 2), from which we derive features explained in Section 4.2.

### Table 1: Subjective variables: daily EMA/PeerMA surveys

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>EMA</th>
<th>PeerMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>Self-perceived stress, fatigue, anxiety</td>
<td>Peer-perceived stress, fatigue, anxiety</td>
</tr>
<tr>
<td>Frequency</td>
<td>max 8x day 9am-9pm. Uniformly randomized, separated by at least 45 mins</td>
<td>immediately following each EMA</td>
</tr>
<tr>
<td>Questions</td>
<td><em>How much {stress, fatigue, anxiety} are you experiencing?</em></td>
<td><em>How much {stress, fatigue, anxiety} is your peer projecting?</em></td>
</tr>
<tr>
<td>Confidence</td>
<td>[0-4] zero, low, moderate, high</td>
<td></td>
</tr>
<tr>
<td>Channel</td>
<td>Silent push notification (no sound, no vibration), expiring in 30 minutes if not answered (to prevent questions from piling up).</td>
<td></td>
</tr>
<tr>
<td>Format</td>
<td>Single-item question with a visual analog scale with values from 0 to 1 [5]</td>
<td></td>
</tr>
</tbody>
</table>

### Table 2: Objective variables: collected by the *mQoL Peers* application

<table>
<thead>
<tr>
<th>Name</th>
<th>Description of variables recorded</th>
<th>Triggered by</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applications Used</td>
<td>Frequency of use of application’s categories defined in the Play Store</td>
<td>Changes of the application on the</td>
</tr>
<tr>
<td></td>
<td>(e.g., communication, productivity, entertainment, and others).</td>
<td>screen</td>
</tr>
<tr>
<td>Screen Touches</td>
<td>Number of times the person touches the screen while it is on.</td>
<td>New screen session</td>
</tr>
<tr>
<td>User Activity</td>
<td>Current user’s physical activity as detected by Google Services: (e.g.,</td>
<td>Changes in the user’s activity</td>
</tr>
<tr>
<td></td>
<td>still, in-vehicle, on-bicycle, on foot, running).</td>
<td></td>
</tr>
<tr>
<td>User Presence</td>
<td>Frequency of screen rotation and session duration (from ON to OFF).</td>
<td>Screen activation or rotation</td>
</tr>
<tr>
<td>Network Connectivity</td>
<td>Values of the network connection (e.g., WiFi level, WiFi interface</td>
<td>Changes in network connection state</td>
</tr>
<tr>
<td>Battery</td>
<td>Battery state (e.g., charging, full), level and temperature.</td>
<td>Changes in battery state</td>
</tr>
</tbody>
</table>

4. Data Preparation

We show the socioeconomic characteristics of all the study participants and peers in Table 3. However, for this paper, we selected only 9 participants and 10 peers whose *continuity* (defined as the percentage of days with at least

---

¹ Application name in the Google Play Store: *mQoL Peers*
one completed survey) was at least 65% (~2/3 of the period). That way, we discard dyads who did not regularly answer the surveys to avoid result’s bias. Table 4 shows information about EMA and PeerMA responses, as well as objective variables passively obtained from the participant’s smartphones during the study period. A related paper [3] presents a detailed qualitative and quantitative analysis of the complete dataset.

4.1. EMA: Defining Target Variables

The target variables are derived from the EMA assessments of each participant. Fig 1 illustrates eight daily EMA scores for stress, fatigue, and anxiety (horizontally). The values are normalized to [0-1] based on the highest and lowest assessment given by each person along the study. From these values, we calculate an additional variable called “well-being” (shown in Fig 1 as “x”) inspired by Huppert [13]. For this study, we define well-being as one minus the arithmetic mean of stress, fatigue, and anxiety. Empirically, high stress, fatigue, and anxiety reduces well-being. Conversely, low levels of stress, fatigue, and anxiety contribute to a healthy state of well-being. Therefore, there are 4 target variables for EMA: stress, fatigue, anxiety and (derived) well-being.

For each day, we sampled the EMA data using a 2h non-overlapping time window (Fig 1 under day 2 with dashed lines) resulting in six windows between 9am-9pm. We chose the time window experimentally to hold at most two EMAs. When the window has one or two EMAs, the score for stress, fatigue, and anxiety for that window is the arithmetic mean of EMAs. When the window has zero EMAs (shown in Fig 1 as “?”), the missing scores are imputed using the “nearest” method; which has the advantage of imputing the missing values with neighboring values that were effectively given by the participant (although at a nearby moment of that day), consequently the imputations are not strictly synthetic. Fig 2 illustrates the process of splitting absolute values of EMAs (and PeerMAs, see Section 4.2) into 2h time windows, after which, missing values are imputed for both, as described above. Table 5 shows the size of the datasets at each stage; after this process, the amount of EMAs and PeerMAs is equal.

Table 4: Summary of Participant and Peer’s Data Contributions.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
<th>Participants</th>
<th>Peers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>male</td>
<td>6 (46%)</td>
<td>8 (40%)</td>
</tr>
<tr>
<td></td>
<td>female</td>
<td>7 (54%)</td>
<td>12 (60%)</td>
</tr>
<tr>
<td>Age</td>
<td>18-20</td>
<td>1 (8%)</td>
<td>2 (10%)</td>
</tr>
<tr>
<td></td>
<td>21-29</td>
<td>9 (69%)</td>
<td>8 (40%)</td>
</tr>
<tr>
<td></td>
<td>30-39</td>
<td>2 (15%)</td>
<td>4 (20%)</td>
</tr>
<tr>
<td></td>
<td>40-49+</td>
<td>1 (8%)</td>
<td>6 (30%)</td>
</tr>
<tr>
<td>Marital status</td>
<td>single</td>
<td>10 (77%)</td>
<td>12 (60%)</td>
</tr>
<tr>
<td></td>
<td>married</td>
<td>2 (15%)</td>
<td>5 (25%)</td>
</tr>
<tr>
<td></td>
<td>other</td>
<td>1 (8%)</td>
<td>3 (15%)</td>
</tr>
<tr>
<td>Education level</td>
<td>undergrad</td>
<td>8 (62%)</td>
<td>11 (55%)</td>
</tr>
<tr>
<td></td>
<td>graduate</td>
<td>5 (38%)</td>
<td>9 (45%)</td>
</tr>
<tr>
<td>Currently employed</td>
<td>yes</td>
<td>3 (23%)</td>
<td>9 (45%)</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>10 (77%)</td>
<td>11 (55%)</td>
</tr>
<tr>
<td>Others notice my  #</td>
<td>yes</td>
<td>9 (69%)</td>
<td>15 (75%)</td>
</tr>
<tr>
<td>my stress?</td>
<td>no</td>
<td>4 (31%)</td>
<td>5 (25%)</td>
</tr>
</tbody>
</table>

4.2. Extracting features from PeerMA and Smartphone Sensors

PeerMA derived features (pDFs): With every PeerMA, peers answer four questions (stress, fatigue, anxiety and confidence level, Section 3) for which we record the score (value), and the time in seconds used to answer the question.

Table 5: Size of EMAs and PeerMAs datasets for each <participant, peer> dyad on each stage of data processing.

| Dataset stage | S1P1 S1P1 S1P1 S1P1 S1P1 S1P1 S1P1 S1P1 S1P1 S1P1 |
|---------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|               | E1    | E1    | E1    | E1    | E1    | E1    | E1    | E1    | E1    | E1    |
| Original      | 127   | 99    | 213   | 74    | 111   | 50    | 216   | 101   | 153   | 91    |
| Imputed       | 91    | 82    | 157   | 66    | 82    | 46    | 163   | 83    | 97    | 74    |
| 2h Window     |      |       |       |       |       |       |       |       |       |       |

EMAN PeerMA
pace), leading $4 \times 2 = 8$ features per PeerMA. Then we compute the peer-based “well-being” from the PeerMA value of stress, fatigue and anxiety (shown in Fig 1 as “x” next to PeerMA) analogous to the EMA-based “well-being” score.

Next, we focus on the 2h windows (of Fig 1) for the analysis. Because there can be more than one PeerMA in each 2h window, we take the above-defined 8 features, and compute their min, max, mean for each window, producing $8 \times 3 = 24$ numeric features. Then, we compute the min, max, mean of the well-being values, adding 3 more features. Finally, we compute the min, max, mean of the time in seconds the peer took to complete the PeerMA from beginning to end (per 2h window), adding 3 more features. As a result, pDFs account for $24 + 3 + 3 = 30$ quantitative features (per each 2h window). If there are zero PeerMAs in the window (e.g., shown in Fig 1 as “?” under PeerMA), then the values are imputed using the “nearest” method.

**Smartphone-sensors derived features (sDFs):** We used six raw data sources from the smartphone sensors, as listed in Table 2. For applications used, screen touches, user physical activity, and user presence events, we reproduced the approach of [7, Sec 5.2] with slight modifications. In their paper, they derived 88 features from these data streams, while we derived 74. Additionally, we added two more data sources, network connectivity and battery events, from which we derived 36 features. In total, we derived 110 quantitative numerical sDFs, both continuous and discrete. Due to space constraints, we describe the sDFs in this public repository^2. Similar to the pDFs, we aggregate the sDFs using the min, max, mean of the values for each 2h window (Section 4.1). If there are zero records for a sensor within a time window, we treat those columns as missing values without imputation. We decided to do that due to the heterogeneity of the sensors which requires a case-by-case exploration to choose adequate imputation techniques for each.

5. Data Analysis

5.1. Prediction Scenarios

Recapitulating, our goal is to examine how informative are the pDFs on a ML context to predict the state of an individual (in this case, stress, fatigue, anxiety, and derived well-being obtained via EMA). To accomplish that, we take the final (imputed) dataset from Table 5, and for each of the 10 participant, peer dyads (Table 4), we extract five sub-datasets/prediction scenarios (named DS.1-DS.5 in Fig 3) to examine our research questions from different angles. Conceptually, one could train two models, one with, one without pDFs, and then compare both results. However, we train several models using these sub-datasets to examine RQ.1 more thoroughly.

DS.1 consists of sDFs only (110 features, Section 4.2) and is the baseline scenario as used in the literature [29, 12, 7, 22]. We form DS.2 by adding the pDFs to DS.1, and it allows us to identify the effect of the pDFs when combined with sDFs. We form DS.3 by replacing the pDFs of DS.2 with PeerMA-like random noise to crosscheck the results obtained from DS.2, since that way, DS.3 should not yield better results than DS.2, if the PDFs are informative. DS.4 is formed by pDFs-only, representing a direct and unique contribution of our research as we examine the predictive

^2 https://gitlab.unige.ch/qol/peerma_ml
To examine the results, we focus on comparing DS.1 with DS.2, and DS.4 (which have pDFs), and DS.1 with DS.3 and DS.5 (which have PeerMA-like random noise).

5.2. Definition of Target Classes and Classification Task

The simplest approach to predict the state of the individual is to transform the four target variables (stress, fatigue, anxiety, and well-being) into two classes (class 0, class 1) to perform binary classification. The cutoff point in each case is the median of the assessments. We recall that prior to computing the median, the assessments were normalized using the min and max value provided (for each user separately). Thus, EMA values smaller than the median are assigned to class 0, and those larger than or equal to the median are assigned to class 1. For the classification task, we chose four standard classification algorithms used in related literature: Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGB), and ZeroR as a baseline which always chooses the majority class. We used the scikit-learn framework for machine learning in Python.

For each algorithm, we trained a model with the five sub-datasets DS.1-5 of Fig 3 to classify the well-being target. Then, due to the simplicity of interpretability, we picked the DT and trained another set of models for DS.1-5 to classify stress, fatigue, and anxiety as the targets to compare these models’ accuracy with the one obtained for well-being (which derives from those variables).

To train the models, we used a nested cross-validation approach consisting of two loops. The outer loop uses a cross-validation (CV) to split the dataset into stratified train/test segments using the RepeatedStratifiedKFold method with parameters n_splits=10 and n_repeats=4. The inner loop uses a GridSearchCV that further splits each training set into stratified training/validation segments using the StratifiedKFold method with n_splits=10. Thus the inner loop is repeated to search the best performing hyperparameters set, choosing from a list of predefined values for each of the algorithms DT, RF, and XGB. As a result, the outer loop yields 40 trained models from which we compute the mean accuracy as the evaluation metric. We repeat this process 4 x 5 x 10 = 200 times (n. of algorithms: DT/RF/XGB/ZR, n. of sub-datasets: DS.1-5, n. of dyads) for a total of 8000 models. In this paper, we examined only “within-subject” models, i.e., separate models for dyads. Experiments on a global model (combining data from all participants and peers into a single dataset) are left as future work.

6. Results and Discussion

Fig 4 shows the results of the binary classification task (described in Section 5.2), predicting the individual’s well-being. The top chart of Fig 4 shows that the accuracy of the classifiers across all the dyads increased from the baseline DS.1 (sDFs Only) dataset to DS.2 (sDFs+pDFs), and also from DS.1 to DS.4 (pDFs Only). The classification accuracy obtained from DS.2 and DS.4 is very similar, especially for RF and XGB. Also, DS.4 produced very similar results for the RF and DT classifiers, and slightly lower for the XGB. Not surprisingly, the bottom chart of Fig 4 shows that the accuracy of the classifiers decreases with DS.3 and DS.5, since they rely on PeerMA-like random noise.

---

Fig. 3: Graphical representation of five sub-datasets used to examine the informative value of PeerMA derived features on the classification task.

---

https://scikit-learn.org/
For each algorithm (DT, RF, XGB), we run a two-sided Student’s t-test to assess the statistical significance between the differences among the pairs DS.1 vs. DS.2, and DS.2 vs. DS.4 across all dyads. Because we tested DS.2 twice, we applied a Bonferroni correction, with a p-value of 0.025. The differences are statistically significant for the DT, but not for the RF and XGB classifiers. RF and DT are the most accurate when using pDFs only to predict the individual’s well-being state. RF and XGB performed better than the DT when combining sDFs and pDFs.

Fig. 4: Average classification accuracy across all dyads comparing four ML algorithms predicting well-being

Fig. 5: Classification accuracy of a Decision Tree for each dyad comparing three datasets DS.1, DS.2 and DS.4.
Fig 5 shows the results of the binary classification task for every dyad using the DT classifier, and the four target variables: stress, fatigue, anxiety, and well-being. In this sample, the difference across all dyads between the average classification accuracy using DS.1 (sDFs Only) vs. DS.2 (sDFs+pDFs) is not statistically significant in any of the four targets. However, the difference is statistically significant between DS.2 and DS.4 (pDFs only) in all the four targets (in both cases using a t-test with an adjusted p-value of 0.025).

**Results for RQ.1:** To answer RQ.1, in this study, the pDFs added predictive value to the binary classification task in all ML models (shown in Fig 4 and Fig 5). One important aspect to notice is that, despite the small number of training samples in our dataset (Table 5), the algorithms discriminated better the target variable class in the presence of pDFs. Considering the anecdotal evidence for these results, we observe in Fig 5 that for S1P1, S7P2, both DS.2 and DS.4 yield particularly high accuracy compared to DS.1 across all the targets. In both cases (S1P1 and S7P2), the peer is a parent of the participant. Similarly, DS.4 was particularly informative for S11P1 and S12P2 across all the targets compared to both DS.1 and DS.2. In these cases, the relationship between the peer and the subject was friendship.

<table>
<thead>
<tr>
<th>Feature Cat.</th>
<th>Rank</th>
<th>Feature Cat.</th>
<th>Rank</th>
<th>Feature Cat.</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>network</td>
<td>1</td>
<td>touches</td>
<td>7</td>
<td>stress (v)</td>
<td>13</td>
</tr>
<tr>
<td>hour of day</td>
<td>2</td>
<td>well-being (v)</td>
<td>8</td>
<td>fatigue (p)</td>
<td>14</td>
</tr>
<tr>
<td>fatigue (v)</td>
<td>3</td>
<td>anxiety (v)</td>
<td>9</td>
<td>stress (p)</td>
<td>15</td>
</tr>
<tr>
<td>battery</td>
<td>4</td>
<td>activity</td>
<td>10</td>
<td>confidence (p)</td>
<td>16</td>
</tr>
<tr>
<td>day of week</td>
<td>5</td>
<td>presence</td>
<td>11</td>
<td>anxiety (p)</td>
<td>17</td>
</tr>
<tr>
<td>applications</td>
<td>6</td>
<td>confidence (v)</td>
<td>12</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6: DS.2: Most predictive features of the DT for well-being

<table>
<thead>
<tr>
<th>Feature category</th>
<th>Avg.</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>well-being (v)</td>
<td>3.3</td>
<td>2.4</td>
</tr>
<tr>
<td>fatigue (v)</td>
<td>4.6</td>
<td>2.7</td>
</tr>
<tr>
<td>anxiety (v)</td>
<td>4.6</td>
<td>3.2</td>
</tr>
<tr>
<td>stress (v)</td>
<td>4.9</td>
<td>2.2</td>
</tr>
<tr>
<td>stress (p)</td>
<td>5.5</td>
<td>3.3</td>
</tr>
</tbody>
</table>

Table 7: DS.4: Most predictive features of the DT for well-being

Results for RQ.2: To answer RQ.2, we present the relative importance of the pDFs from the DT algorithm only (due to space limitations), and the well-being prediction (Fig 5, derived from stress, fatigue, and anxiety, Section 4.1). For each dyad and sub-dataset (DS.1-5), we took the most accurate models (40 in total) from the outer cross-validation loop (Section 5.2) and retrieved the array of features’ importance. Then, we examined the average importance of the features across the 40 models in DS.2 and DS.4 (leaving out DS.1, 3, 5 as they do not have pDFs).

We examined the composition of the top 30 features (equals the # of pDFs) based on their average importance to count how many of them are pDFs. Table 6 shows the results for DS.2 (sDFs+pDFs) where we ranked the features by dividing its number of occurrences across the 10 dyads over its average position. Then, Table 7 shows the results for DS.4 (pDFs only). We show the average position, and standard deviation of each feature across all 10 dyads.

Summarizing both Table 6 and Table 7, the most predictive pDFs relate to the value of the assessment of fatigue, well-being, and anxiety, in this order when combined with sDFs, and well-being, fatigue, and anxiety when using only pDFs. Fatigue and anxiety are states that individuals express more openly and even involuntarily [18], contrary to stress, which tends to be affected by stereotypes [20]. Consequently, if peers can accurately estimate some states of the individual, the algorithms leverage this information to improve its predictive accuracy. This result has potential benefits when predictions are derived solely based on data acquired from the smartphone sensors. In our dataset, we derived well-being from a combination of stress, fatigue, and anxiety; consequently, its high predictive value could be an indication of how clearly it (i.e., “computed well-being”) summarizes those three variables.

7. Conclusion and Future Work

We presented results from a first of its kind study leveraging the PeerMA method to create PeerMA derived features (pDFs) to be used in machine learning. Our preliminary results show that three models yield higher well-being classification accuracy when trained using pDFs, compared to being trained with only smartphone-sensors derived features (sDFs). We looked inside the Decision Tree model and found evidence that pDFs were among the features that contributed to higher discrimination power in the classification task. The results are not generalizable; however they are promising for the field of personal health supported by mobile computing, and pervasive technologies.

A limitation of this work is the small dataset which can be overcome in longitudinal studies, or by conducting the experiments with global models using data from various subjects and techniques such as leave-one-(subject)-out. As future work, we plan to a) examine other pDFs beyond (min, max, mean) employed here; b) explore other imputation methods for missing data; c) examine other evaluation metrics beyond classification accuracy (e.g., precision, recall,
F-Score). Finally, we plan to examine the effect of pDFs in a global model that combines data from all participants and peers into a single dataset.

Acknowledgements

We thank the following colleagues for their valuable comments: M. Muszynski, S. Ferdowski, V. Manea, and M. Ben Moussa. We also thank the Swiss FCS (2016-19), H2020 WellCo (769765), and The University of Costa Rica.

References