Proposing Design Recommendations for an Intelligent Recommender System Logging Stress

Aku Visuri
Center for Ubiquitous Computing, University of Oulu, Finland
aku.visuri@oulu.fi

Romina Poguntke
VIS, University of Stuttgart, Germany
romina.poguntke@vis.uni-stuttgart.de

Elina Kuosmanen
Center for Ubiquitous Computing, University of Oulu, Finland
elina.kuosmanen@oulu.fi

Every submission will be assigned their own unique DOI string to be included here.

Abstract
The connection between stress and smartphone usage behavior has been investigated extensively. While the prediction results using machine learning are encouraging, the challenge of how to cope with data loss remains. Addressing this problem, we propose an Intelligent Recommender System for logging stress based on adding a subjective user data-based validation to predictions made by intelligent algorithms. In a user study involving 731 daily stress self-reports from 30 participants we found discrepancies between subjective and smartphone usage data, i.e. battery, call information, or network usage. Despite the good prediction accuracy of 65% using a Random Forest classifier, combining both information would be beneficial for avoiding data and improving prediction accuracy. For realizing such a system (i.e., a mobile application), we propose three design recommendations, based on the capabilities of frequently used machine learning classifiers, enabling users to annotate their daily stress levels with a predict-and-validate methodology.

Author Keywords
Smartphones; Sensor Data; Stress Recognition.

ACM Classification Keywords
H.5.m [Information Interfaces and Presentation (e.g. HCI)]: Miscellaneous

Group	NMS	NA	NS
LS	63%	65%	-
HS	48%	65%*	82%*

Table 1: Stress prediction accuracy for high stress (HS) and low stress (LS) participants, using the NMS (none-mild-severe), NA (none-any), and NS (none or mild-severe) categorization. For results marked with '*' significant overfitting occurred, meaning that the algorithm was significantly biased towards one of the alternatives.

RW	Sensors	Accuracy
[9]	Contacts, Visited websites, Location, App usage.	66%
[11]	Calls, SMS, Location, Communications, Screen.	75%
[12]	Location, Accelerometer, Microphone, SMS, Calls.	70%
Our Study	Calls, Battery, Network, Screen.	65%

Table 2: Overview for related work and our experiment regarding mood and stress prediction accuracy based on smartphone sensor data.

Introduction and Background

The primary challenge of smartphone tools aiming at collecting longitudinal information about self-behavior is ensuring continuous data collection [13, 14]. When data collection ceases or suffers a temporary break, insights generated by such applications lose accuracy. Applications that require less effort from the user ensure higher initial user adoption and application usage lifetimes tend to be longer [5].

Stress Recognition based on Mobile Sensor Data

The rapid proliferation of the smartphone has enabled conducting various scientific studies using the rich array of onboard sensors. Inferring affective states from mobile sensors and usage patterns, such as communication history and application usage has also facilitated studies on identifying when users are stressed [9]. Sano and Picard [11] detected stress using mobile usage data, namely location, calls, messages, screen states (on/off) and assessed the participants' stress level subjectively through surveys. By analyzing both data sets, screen state, mobility, call or activity level information yielded 75% accuracy in a binary classification, in comparison to 87.5% accuracy obtained using survey data. Adding weather conditions and personal traits, Bogomolov et al. [3] developed a person-independent 2-class daily stress recognition model with an accuracy of 72.28%. Likewise, Ferdous et al. [4] took smartphone application usage 'as a predictor of perceived stress levels at workplace' resulting into a model with 75% average accuracy and 85.7% average precision.

Towards an Intelligent Recommender System

Designing intelligent stress detection recommender systems becomes increasingly popular for reliably identifying stress. Hereby, the requirement to provide reliable predictions based on sufficient data remains one the core challenges. By this work we contribute to a better understanding of how to design such future systems that avoid data loss and imporove prediction accuracy by proposing three design recommendations considering user needs already at the design stage. By proposing an intelligent recommender system that ensures longitudinal data collection helping the user recollect and annotate past emotional states, e.g. when he or she forgot or neglected to log, we envision to improve stress detection being performed in research studies but also for users interested in the 'quantified self'. We inform the design by the results from our user study indicating that combing subjectively assessed data and smartphone usage pattern would yield huge advantages for overcoming data loss and improving prediction acccuracy.

User Study

We conducted a user study for which we collected daily stress levels via voluntary self-reports from 30 participants ($M = 26.5, SD = 4.61$, 6 females, 23 males, one not specified). The daily stress level was assessed once per day (between 6PM and 12PM) using a three-tier scale ('None', 'Mild', and 'Severe'). In addition to the self-reports we tracked daily device usage patterns from battery information, call-related information, network usage, and usage statistics. Participants logged their stress levels altogether 731 times.

Results

With the data collected in our experiment, we sampled different machine learning algorithms. We clustered our participants into high stress (HS) (N = 18) and low stress (LS) (N = 12) groups, similarly to the Perceived Stress Scale (PSS) score as performed in [11]. We experimented further by sampling the accuracy for the 3-tier prediction ('None', 'Mild', or 'Severe' (NMS) stress), for 'None'-'Any' (NA), and 'None-Mild'-'Severe' (NS). The prediction accuracies are

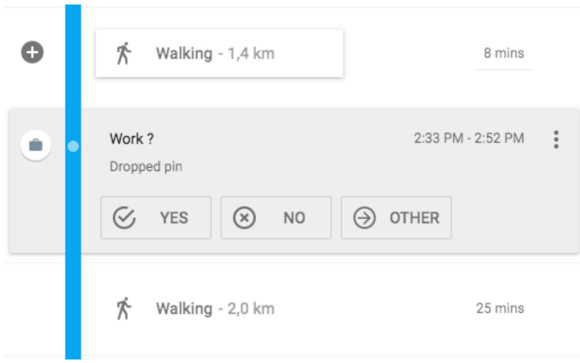

Figure 1: The Google Timeline[1] inquires about past events where the prediction algorithms lacks sufficient certainty.

listed in Table 1. Detecting differences between no stress and any level of stress for the low stress (LS) - None or any stress (NA) participant group is deemed as the best result for our experiment with the accuracy of 65%.

Our results are in line with other findings revealing that predicting stress is most accurate when measuring low to medium levels, but the efficiency in solely relying in the predicted values is mediocre, at best [1]. Our study - alongside the cited work - functions as a motivation for design-orientated solutions aiming to address the limitations of such predictive systems. While broadly speaking the stress prediction accuracy of intelligent systems is still lacking (ranging from 65% to 75% percentage), novel design choices could be used to derive these predictions to still be meaningful to the user.

Informative Classifier Characteristics
Commonly best performing ML classifier in modelling according to mobile sensing data is the Random Forest [15, 8]. The Random Forest algorithm can offer interesting details on its performance and the factors behind the performance, which are often used merely in the analysis phase. The Random Forest algorithm functions by creating a number of sub-trees (in our case best performing N = 1500), with a subset of the attributes used for the algorithm (in our case K = 5 out of the 23 available attributes). A voting mechanism is then used across all the generated trees and the popularity vote is selected as the algorithm's prediction. Two useful sets of detail can be extracted from the generated model and its predictions. **Attribute importance** ranks the used attributes according to their importance in making a correct prediction, and is defined by the *average impurity decrease*, signifying the deterioration of accuracy when each attribute is removed from the set.

Second detail provided by the classifier is its **confidence level** on its own predictions. The confidence level indicates *how many of the subtrees' predictions were correct* and is reported as a percentage from 0 to 100. These details can be further leveraged in application design to provide transparency to the predictions, and to offer more detail to the user about his own behaviour and on potential factors influencing her stress levels.

Design Recommendations
Envisioning Intelligent Recommender Systems for logging stress to be usable for users, we consequently introduce three core aspects for designing applications that support stress detection inspired by prior work [10, 11, 9, 12] and our user study. The capabilities of intelligent mechanisms are not properly appropriated for user-focused design, as several potential mechanisms available in e.g., machine learning classifiers, remain unused. Thus, we propose three mechanisms for applications; data validation, decision-making transparency, and instilling confidence about the recommendations to the user. In Figure 2 we depict how

these recommendations could be transferred and included in a mobile application. The highlighted sections (1, 2, and 3) in Figure 2 are annotated in the left margin.

Validation

With an prediction accuracy capping at 75%, any application should not make an attempt perform mood predictions without **validating** the results from the user. Since the validation process requires user input, the main methodology in a mood sensing or logging application should rely on self-reports, which are still considered the most accurate method for tracking low to medium levels of stress [1]. The core concept of designing intelligent mood sensing applications is **preventing** users' forgetfulness in logging with the use of intelligent mechanisms, e.g. by removing part of the effort in recollecting past events, and reducing likelihood of erroneous past entries. An exemplary implementation could be a timeline of the users logged moods, with existing gaps in data pre-filled with the predicted values, but still waiting for user validation before being finally logged.

Transparency

Another common challenge in implementing intelligent methods is communicating the decision-making to users appropriately. From both, the feature extraction and ranking capabilities of machine learning models, and the confidence value provided by the Random Forest algorithm, some details can be presented the user comprehensively. Many other potential software sensors exist, which have not been extensively investigated for having influence on stress level. For example, calendar event details, social media activity, or physical activity levels extracted from other applications (*e.g.*, Google Fit) can potentially have an influence on experienced stress. Based on the association between highly influential attributes and stress levels, hints of attributes factoring to user experiencing low or high levels of stress can be offered. This both helps users in validating the prediction, and offers insights in potential stressors - being referred to as a very influential method for ensuring long-term use and better user experience [2].

Instilling Confidence

Lastly, users confidence in the predictions can be further improved by positive reinforcement. Showing and reminding users of correct predictions characterizes the system as a reliable method for helping in self-reporting efforts. Even if an entry is already logged by the user, the algorithm can continuously learn and self-reflect, and attempt to predict whether *if it would have been correct* - even when there was no distinct need for a prediction.

Discussion

As previous work found, stress levels of users are associated with smartphone activities and usage such as activity level, SMS, battery usage, and screen on/off patterns [11]. Hence, we focused on the analysis of these parameters in our user study. In addition to commonly used usage characteristics and sensor information, our experiment also included the use of battery statistics, often cited as a source of stress, and also potentially as indication of laborious or hectic days, which can be associated with increased stress levels. Hosio *et al.* [6] revealed how study participants greatly valued their smartphone battery life when it reached low or critical thresholds, and discusses how the use of smartphone battery is of vital importance to users, and any problems of challenges in battery lifetime can be experienced as unwanted or stressful. In our study, the battery details were most important attributes for prediction accuracy (ranked top5 in the extracted feature importance). The diminished accuracy was likely due to the minimal sensor selection (Table 2). However, as accessing this set of sensors is very power efficient, it is a good minimum start-

1 VALIDATION
Each missed entry is marked with a '?' on the stress level diagram. By selecting each of missed entries, user is asked to **validate** the prediction. If the prediction is considered incorrect, the user has the option to directly correct the entry to another value.

2 TRANSPARENCY
The reasoning for the algorithm's predictions is shown to the user via the **confidence** value, indicating how certain the model is that is is correct, as well as the **associated attributes**, showcasing which daily activities most likely influenced the prediction (and are associated with different stress levels).

3 INSTILL CONFIDENCE
If the system can provide observable results, the user will be more likely to **trust** the system. In addition to the validated predictions, each logged entry is also predicted on the background, and correct predictions are marked in the stress level diagram.

ing point for application leveraging mood predictions in its design. In addition to prediction results, and prediction accuracy, other factors should be provided to the user in order to create transparency between the prediction methodology and its results. This makes it effortless for the user to understand the mechanisms behind the intelligence, and to understand that - in this case - the intelligent algorithms are not fool-proof. Thus, we propose an intelligently supported design for applications relying on mood logging. The core concept is a method where intelligent predictions are used to fill gaps in logging (*i.e.*, past entries), where the user verifies predictions made from sensed data. Such a system could support developing a 'gold standard' for a 100% accuracy stress detection based on mobile sensor data, that is currently missing [7]. Similarly, Google Timeline (cf. Fig. 1) realizes a related approach prompting their users to verify locations that had been tracked with a certain amount of uncertainty.

Although taking the user into the design loop, this method is less unobtrusive than the pure sensing based on smartphone usage patterns. Future research will have to focus on improving the algorithms accuracy verified by as little user validated data as possible to limit the amount of prompts the user has to fill in. Another challenge to be dealt with will be the memory biases that could influence annotating formerly experiences affective states. For this, future work be interested in providing contextual information or use information involving life logging to make the user fully engage with the feelings he or she had been experiencing throughout the inquired activities. In summary, we believe that establishing Intelligent Recommender Systems for logging stress help to built future systems reassuring that the inferred stress detection based on smartphone usage patterns is more reliable and therefore beneficial, particularly

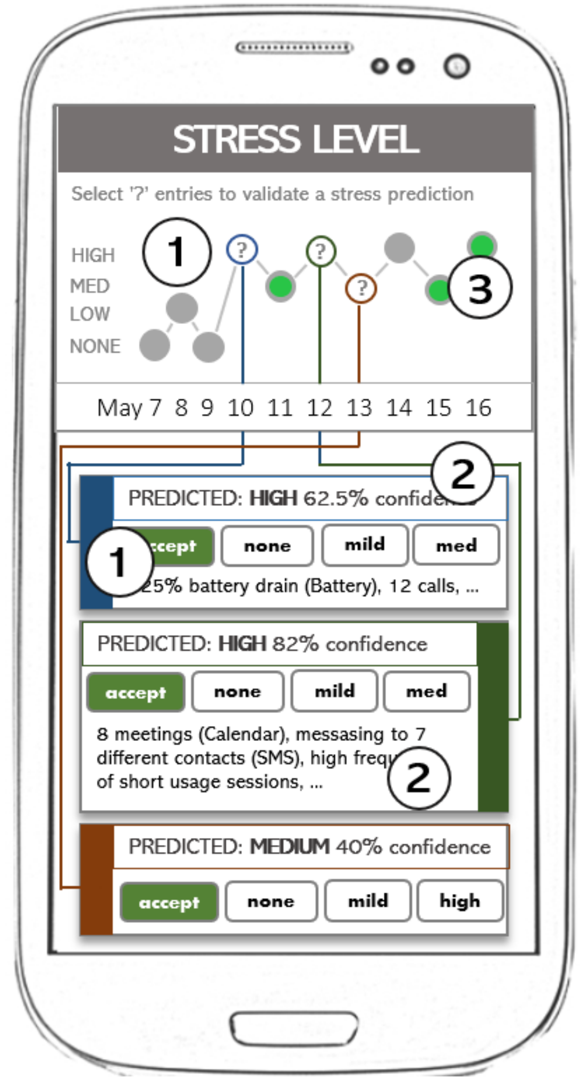

Figure 2: A wireframe sketch of how a design based on our guidelines could look like, with the three guidelines annotated by 1) Validation 2) Transparency and 3) Instilling confidence.

when user needs are addressed as suggested in our design recommendations.

Conclusion

Addressing the problem of how to overcome data loss when using smartphone usage data for predicting affective states, we have proposed combining subjective and objective data collection through an Intelligent Recommender System. For the realization of such a user validation-based system, we proposed three design recommendations addressing user needs facility its usability. In this work, we have presented the stress level prediction results from a user study involving 30 participants using a Random Forest classier. Apart from logging smartphone usage patterns, we collected 731 stress self-reports parallel to the objective smartphone usage pattern logging. Our findings indicate that battery-related variables have a strong association with perceived stress levels - something that has not been previously reported. Based on our results reavaling discrepancies between subjective and objective data, we finally introduced three design recommendations for designing applications that collect user's emotional states, e.g., stress and combine them with objective data to overcome data loss and considerably improve stress detection accuracy. By this work, we envision to support intelligent recommender systems logging stress and being used, particularly for research settings to validate stress detection based on smartphone usage data by users' annotations.

REFERENCES

1. Phil Adams, Mashfiqui Rabbi, Tauhidur Rahman, Mark Matthews, Amy Voida, Geri Gay, Tanzeem Choudhury, and Stephen Voida. 2014. Towards personal stress informatics: Comparing minimally invasive techniques for measuring daily stress in the wild. In *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 72–79.

2. Frank Bentley, Konrad Tollmar, Peter Stephenson, Laura Levy, Brian Jones, Scott Robertson, Ed Price, Richard Catrambone, and Jeff Wilson. 2013. Health Mashups: Presenting statistical patterns between wellbeing data and context in natural language to promote behavior change. *ACM Transactions on Computer-Human Interaction (TOCHI)* 20, 5 (2013), 30.

3. Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex (Sandy) Pentland. 2014. Daily Stress Recognition from Mobile Phone Data, Weather Conditions and Individual Traits. In *Proceedings of the 22Nd ACM International Conference on Multimedia (MM '14)*. ACM, New York, NY, USA, 477–486. DOI: http://dx.doi.org/10.1145/2647868.2654933

4. R. Ferdous, V. Osmani, and O. Mayora. 2015. Smartphone app usage as a predictor of perceived stress levels at workplace. In *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. 225–228. DOI: http://dx.doi.org/10.4108/icst.pervasivehealth.2015.260192

5. Thomas Fritz, Elaine M Huang, Gail C Murphy, and Thomas Zimmermann. 2014. Persuasive technology in the real world: a study of long-term use of activity sensing devices for fitness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 487–496.

6. Simo Hosio, Denzil Ferreira, Jorge Goncalves, Niels van Berkel, Chu Luo, Muzamil Ahmed, Huber Flores, and Vassilis Kostakos. 2016. Monetary assessment of battery life on smartphones. In *Proceedings of the*

2016 CHI Conference on Human Factors in Computing Systems. ACM, 1869–1880.

7. Karen Hovsepian, Mustafa al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. 2015. cStress: Towards a Gold Standard for Continuous Stress Assessment in the Mobile Environment. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 493–504. DOI: http://dx.doi.org/10.1145/2750858.2807526

8. Simon Klakegg, Jorge Goncalves, Chu Luo, Aku Visuri, Alexey Popov, Niels van Berkel, Zhanna Sarsenbayeva, Vassilis Kostakos, Simo Hosio, Scott Savage, and others. 2018. Assisted Medication Management in Elderly Care Using Miniaturised Near-Infrared Spectroscopy. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 69.

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

10. Leanne G Morrison, Charlie Hargood, Veljko Pejovic, Adam WA Geraghty, Scott Lloyd, Natalie Goodman, Danius T Michaelides, Anna Weston, Mirco Musolesi, Mark J Weal, and others. 2017. The effect of timing and frequency of push notifications on usage of a smartphone-based stress management intervention: An exploratory trial. *PloS one* 12, 1 (2017), e0169162.

11. Akane Sano and Rosalind W Picard. 2013. Stress recognition using wearable sensors and mobile phones. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 671–676.

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

13. Niels van Berkel, Chu Luo, Denzil Ferreira, Jorge Goncalves, and Vassilis Kostakos. 2015. The curse of quantified-self: an endless quest for answers. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. ACM, 973–978.

14. Aku Visuri, Niels van Berkel, Chu Luo, Jorge Goncalves, Denzil Ferreira, and Vassilis Kostakos. 2017a. Challenges of quantified-self: encouraging self-reported data logging during recurrent smartphone usage. In *Proceedings of the 31st British Computer Society Human Computer Interaction Conference*. BCS Learning & Development Ltd., 81.

15. Aku Visuri, Niels van Berkel, Chu Luo, Jorge Goncalves, Denzil Ferreira, and Vassilis Kostakos. 2017b. Predicting interruptibility for manual data collection: a cluster-based user model. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 12.