Investigating the Role of Context in Perceived Stress Detection in the Wild

Varun Mishra[*]
IBM T.J. Watson Research Center &
Dartmouth College
varun@cs.dartmouth.edu

Tian Hao
Center for Computational Health
IBM T.J. Watson Research Center

Si Sun
Center for Computational Health
IBM T.J. Watson Research Center

Kimberly N. Walter
Center for Computational Health
IBM T.J. Watson Research Center

Marion J. Ball
Center for Computational Health
IBM T.J. Watson Research Center

Ching-Hua Chen
Center for Computational Health
IBM T.J. Watson Research Center

Xinxin Zhu
Center for Computational Health
IBM T.J. Watson Research Center

[*] Work was done while author was an intern at IBM.

UbiComp/ISWC'18 Adjunct, October 8–12, 2018, Singapore, Singapore

ACM ISBN 978-1-4503-5966-5/18/10…$15.00
https://doi.org/10.1145/3267305.3267537

ABSTRACT

The advances in mobile and wearable sensing have led to a myriad of approaches for stress detection in both laboratory and free-living settings. Most of these methods, however, rely on the usage of some combination of physiological signals measured by the sensors to detect stress. While these solutions work great in a lab or a controlled environment, the performance in free-living situations leaves much to be desired. In this work, we explore the role of context of the user in free-living conditions, and how that affects users' perceived stress levels. To this end, we conducted an 'in-the-wild' study with 23 participants, where we collected physiological data from the users, along with 'high-level' contextual labels, and perceived stress levels. Our analysis shows that context plays a significant role in the users' perceived stress levels, and when used in conjunction with physiological signals leads to much higher stress detection results, as compared to relying on just physiological data.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing**; • **Applied computing** → *Health care information systems*; *Health informatics*;

KEYWORDS

context aware, Stress detection, mobile health (mHealth), wearable sensing, mental health

ACM Reference Format:
Varun Mishra, Tian Hao, Si Sun, Kimberly N. Walter, Marion J. Ball, Ching-Hua Chen, and Xinxin Zhu. 2018. Investigating the Role of Context in Perceived Stress Detection in the Wild. In *Adjunct Proceedings of the 2018 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2018 International Symposium on Wearable Computers (UbiComp/ISWC'18 Adjunct), October 8–12, 2018, Singapore, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3267305.3267537

INTRODUCTION

Stress can be defined as a physiological response to external stimuli – physical, mental, social or emotional. While short periods of stress can actuate positive changes in an individual's life, continuous and sustained exposure to stress can lead to chronic health outcomes [7, 10]. Hence, timely detection of an individual's stress can help them effectively manage their stress and in turn improve their physical and mental health.

Over the years there have been several works using physiological sensors towards stress detection, in controlled, semi-controlled and uncontrolled (or natural) conditions. The most prominent of which is *cStress*, where the authors use a combination of Electrocardiograph (ECG) and Respiration (RIP) sensors to measure binary stress in the lab and the field [6] settings. The *cStress* system performs well in the lab setting with high recall and an F1 score of 0.81, however, in the field, the performance drops to a median F1 score of 0.71. Building on the *cStress* model, Sarkar et al. conducted an independent field study with 38 participants [12], and were able to detect stress with an F1 score of 0.72.

Gjoreski et al. used the Empatica E3 wrist device [2] to collect Blood Volume Pulse (BVP), Skin Temperature (ST), Heart Rate Variability (HRV), Galvanic Skin Response (GSR), and physical activity level using the accelerometer [3]. Similar to *cStress*, the authors conducted a lab and field study and showed that they could detect binary stress with an F1 score of 0.80 and 0.63 in the lab and field settings respectively.

Egilmez et al. show that in a lab setting, they could achieve F1 scores of 0.88 using just wrist based heart-rate and GSR sensors [1]. Sano et al. conducted a field study where they used wrist-worn sensors to collect accelerometer and GSR data, along with smartphone usage data (including calls, SMS, location and screen on/off) from 18 participants [11]. The authors report results for binary classification of stress with an accuracy of 75%.

A trend observed in all of the above mentioned works is the high stress detection accuracy in a lab or controlled setting, which doesn't translate to similar results in the field setting. One key factor for such a discrepancy is the fact that researchers use models trained in a constrained lab scenario to detect stress in daily living conditions, which has several unknowns, variables and confounding factors. Further, the current usage of just physiological signals assumes that whenever people *perceive* stress, it would be reflected in their physiological signals, which is not always true.

We believe that knowledge of the *context* of an individual would help simplify the task of stress detection by helping address the points above. Contextual information can help provide some a-priori knowledge, which can lead to stress detection with an expected knowledge, or detecting in a particular type of situation, e.g., A stress-detection model trained during mental arithmetic task in the lab might not be appropriate while the user is driving, but might be more appropriate while the user is working or studying.

Further, recent developments in context detection has shown much promise in being able to detect a person's context or contextual behavior using smartphones and wearables with reasonable accuracy [13], suggesting that in the future, fine-grained contextual information would be readily available to incorporate in to different mental and behavioral sensing tasks, including *Stress*.

To test out our initial hypothesis about the usefulness of context, we conduct an *in-the-wild* study, where we collect physiological signals (R-R intervals) from a commercial wearable (Moto 360 [9]) along with self reports for *stress*, which is on a scale of 0–5. Further, we identified several 'high-level' daily activities[1], which we also asked the participants to self-report. As this is an exploratory analysis to observe the effect of context in stress detection, we decided to obtain the context (i.e., the high-level activity) labels as a self-report. Based on the evaluations of this study, we hope to employ 'passive' context detection in future.

[1] Since all participants in the study were employees of a large information technology corporation, the choice of the activities were based on what a regular employee in a technology firm experiences. These activities might or might not be applicable to a different population group.

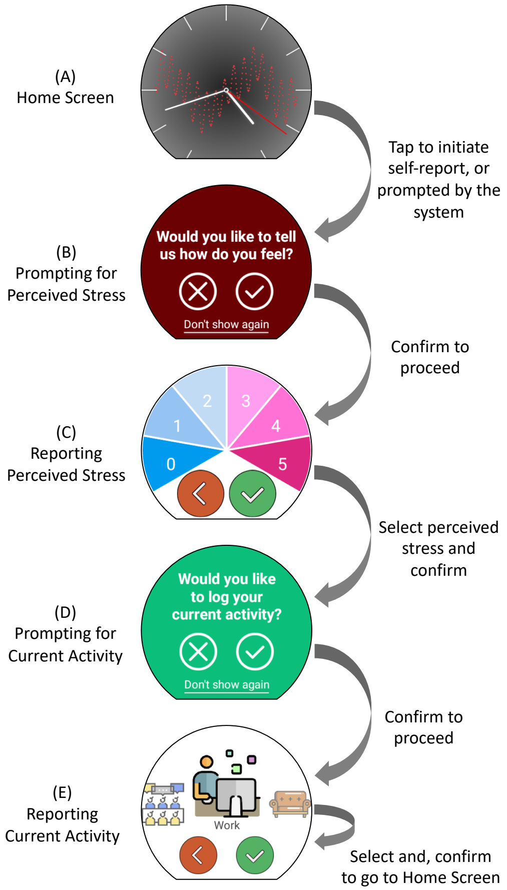

(A)
Home Screen
Tap to initiate self-report, or prompted by the system
(B)
Prompting for Perceived Stress
Would you like to tell us how do you feel?
Don't show again
Confirm to proceed
(C)
Reporting Perceived Stress
0
1
2
3
4
5
Select perceived stress and confirm
(D)
Prompting for Current Activity
Would you like to log your current activity?
Don't show again
Confirm to proceed
(E)
Reporting Current Activity
Work
Select and, confirm to go to Home Screen

Figure 1: The flow of a self-report prompt as presented to the user. During the study, the activity prompt always succeeded the stress prompt, to make sure that the participants' perception of stress was not affected by the activity they were doing.

Our analysis shows that contextual features have a significant effect on the users' perceived stress levels, and when used in conjunction with physiological features lead to better prediction results, in both, binary classification (F1 score of 0.769), and regression based detection models ($r = 0.72$), as compared to using just physiological features or just contextual features.

DATA COLLECTION

We conducted an *in-the-wild* study, with $n = 30$ participants (11 females, 19 males), all over 25 years of age (25-34 years:15; 35-44 years: 8; 45-54 years: 3; over 55 years: 2). All recruited participants were employees of a large information technology corporation. At the start of the study, the participants completed the consent form, along with a brief demographic survey. After signing the consent form, the participants were asked to wear a commercially available smartwatch (Moto 360 [9]) for at least 3 days. The watch was pre-installed with an app to passively collect continuous HRV and accelerometer data from onboard sensors (while the Moto 360 does not have an HRV sensor, we use the approach described by Hao et al. to extract R-R intervals from the onboard Photoplethysmography (PPG) sensor [5]). The app also prompted the participants with self-reports about their in-situ perceived stress levels and context information. Each self-report prompt consisted of two questions; the first question asked the participants to rate their stress level on a scale of 0–5 (0 being extremely relaxed and 5 being extremely stressed), the second question asked about the associated context, i.e., the type of activity they were engaged in during that time. The participants had to choose from 10 common daily activities – sleep, dine, socialize, meeting, work, rest, drive, housework, exercise, and entertainment. During the study, the stress prompt *always* preceded the activity prompt, to make sure that the participants' perception of stress was not affected by the activity they were doing.

During the initial visit, the participants were trained on how to use the app, respond to prompts, and use the stress scale (0–5) to report their stress levels.. The participants were instructed to wear the device for the whole day, except while bathing, swimming and sleeping. The flow of the self-report prompts as presented to the participants, is shown in Figure 1

Net Data Collected

While we collected data from 30 participants, we had to filter out some users due to data quality/quantity issues. Two users seemed to have data quality issues, where one user marked all 'stress' self-reports as 0, while another just marked 0 for all but one 'stress' self-report. Further, we had to remove 5 additional participants from the analyses due to low response count to self-reports, resulting in a total of 23 participants that were finally included in the analyses. Further, since we had asked the participants to not wear the watch while sleeping, we removed the self-reports where the reported activity was 'sleeping' (count = 12), and self-reports where the participants responded between 12 a.m. and 6 a.m. (count = 14).

Thus, we ended up with a total of 112 days worth of data from 23 participants (mean = 4.86), and a total of 1176 self-reports containing both stress and activity labels (mean = 51.13).

METHOD

In this work, we evaluate the effect of contextual features on the users' perceived stress level. We consider three contextual features – (a) the self-reported activity, (b) the time of the day, and (c) the day of the week. The self-reported activity is considered as a categorical value, the time of the day is broken down into three categories, (1) 6 a.m. – 12 p.m., (2) 12 p.m. – 6 p.m., and (3) 6 p.m. – 12 a.m. and the day of the week is also represented as a categorical value ranging from (0) to (6), representing Monday to Sunday.

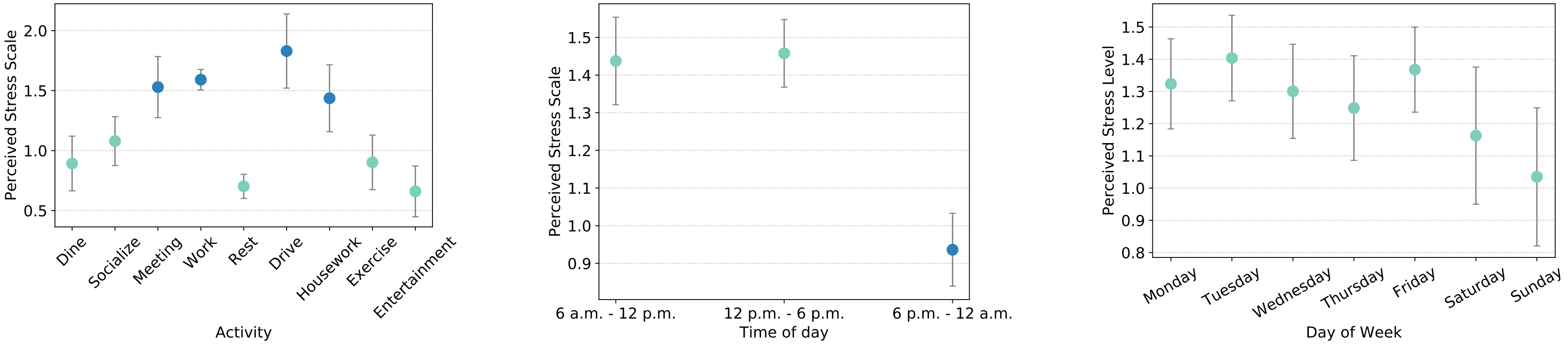
Further, we use the contextual features and the R-R interval features, individually and in combination, to train machine learning models for stress detection. We report results for both (a) classification between binary stress, and (b) regression for a continuous stress scale of 0–5.

For the binary classification task, we calculated the median of the self-reported stress levels for each participant. If the self-reported score for a participant was lower than their median score, we labeled it as '0', i.e., *not stressed*, else we labeled it as '1', i.e., *stressed*.

For using physiological data for stress detection, we performed a thorough data-cleaning and normalization pipeline (as recommended by Mishra et al. for commodity devices [8]), to remove any participant specific effects on the data. Next, we compute R-R interval features for every self-report. Assuming a self-report was answered at time t. We look at a window of time Δt before that self-report, and divide it into 60-second intervals. We compute several time-domain features, as used by Mishra et al. [8] – mean, median, standard deviation, max, min, 80^{th} percentile, 20^{th} percentile and root mean square of successive differences (RMSSD), for each 60-second interval in the $[t - \Delta t, t]$ time window, and label each instance with the self-reported stress score or the binarized stress score at time t. In this work, we choose $\Delta t = 10$ minutes, which is in-line with previous work by Gjoreski et al. [4].

EVALUATION

We start our evaluations by exploring the effect of contextual features on the perceived stress levels. To this end, we conduct three separate one-way ANOVA tests with the perceived stress score as the dependent variable and the reported activity, time of day, and day of the week as independent categorical values respectively. The results of the first analysis (i.e., the effect of activity on the perceived stress levels) show a significant effect, with $F(8, 1167) = 24.48, p < 0.001$. In order to find which activities affect participants' perceived stress, we perform a Tukey post-hoc test (by setting $\alpha = 0.05$). As shown in Figure 2a, the test revealed that activities like Meeting, Work, Driving, and Housework result in significantly higher perceived stress levels.



(a) Effect of activity context on perceived stress. Blue markers represent significantly higher stress levels, $F(8, 1167) = 24.48, p < 0.001$.

(b) Effect of time of day on perceived stress. Blue marker represents significantly lower stress levels, $F(2, 1173) = 31.65, p < 0.001$.

(c) Effect of day of the week on perceived stress. No statistical significance is observed, $F(6, 1169) = 1.82, p = 0.09$.

Figure 2: Mean and 95% CI, showing the effect of different contextual features on the perceived stress levels

	Context	Physiological	Context & Physiological
F1 Score	0.613	0.502	**0.769**
Precision	0.567	0.588	**0.765**
Recall	0.667	0.438	**0.774**
False Positive Rate	0.391	0.243	**0.189**

Table 1: Classification results for binary stress detection, using just contextual features, just physiological features and a combination of contextual and physiological features.

The results of the second analysis (i.e., the effect of time of day on perceived stress levels) show a significant effect, with $F(2, 1173) = 31.65, p < 0.001$. Next, we perform Tukey post-hoc test (with $\alpha = 0.05$) and observe the perceived stress levels between 6 p.m. − 12 a.m. are significantly lower than the rest of the day, as shown in Figure 2b

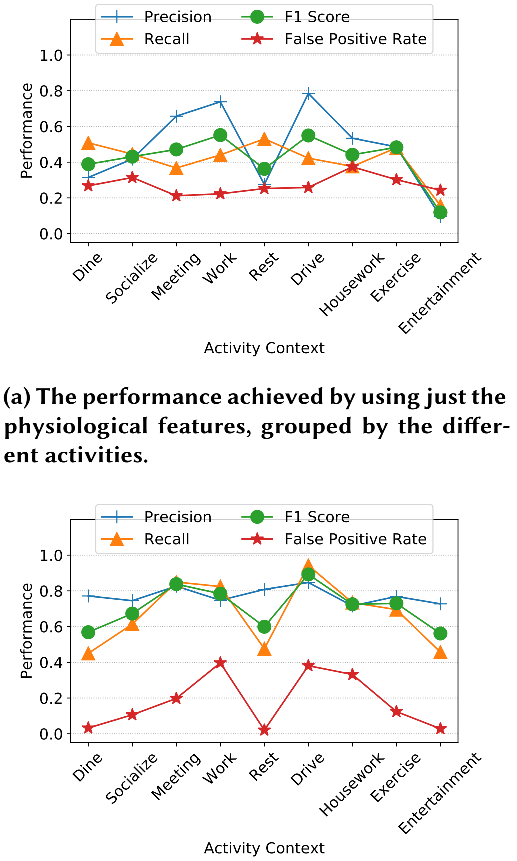
Finally, the third analysis (i.e., the effect of day of the week on perceived stress levels) does not show any significant effect, with $F(6, 1169) = 1.82, p = 0.09$. The variation of perceived stress levels based on day of the week is shown in Figure 2c.

Having established that contextual information (activity and time of day) has a significant effect on the participants' perceived stress level, we move towards detection of stress. As mentioned previously, we evaluate stress detection by, first, a binary classification between stressed and not stressed, and second, regression on a continuous stress scale of 0–5. We use Random Forests for both, the classification[2] and regression evaluations.

We report 10-fold cross-validation results (F1 score, Precision, Recall, False Positive Rate) for a binary classification of stress in Table 1.

It is interesting to see that just the contextual features result in higher classification performance (F1 score) than using just physiological features. The poor performance of just the physiological features

[2] In the case of binary classification, the Random Forest model outputs the probability of the instance belonging to the positive class (i.e., *stressed*). If this probability is greater than a particular threshold (default = 0.5), then that instance is classified as 1 or *true*, else it is classified as 0 or *false*



(a) The performance achieved by using just the physiological features, grouped by the different activities.

(b) The performance achieved by using the combination of contextual and physiological features, grouped by the different activities.

Figure 3: Performance achieved by M_p and M_{c+p} grouped by the different activities.

is expected since the evaluations are *in-the-wild* using just the R-R interval signals from a commercial smartwatch. We, however, notice that using the combination of Contextual and Physiological features leads to a major boost in performance – higher F1 score, precision, recall, all with lower false positive rates.

To further investigate the big performance gap between the classification model using just physiological features (M_p) and the classification model using combination of contextual and physiological features (M_{c+p}), we look at the performance metrics achieved by both the models, segregating them by activity, as shown in Figure 3. In Figure 3a, it is interesting to observe that M_p has relatively better precision in detecting *stress* during Meeting, Work, and Drive activities, all three of which have been shown to relate with significantly higher 'perceived stress' levels. This suggests that out of all the instances classified as *stressed* (in the above mentioned activities), the proportion of instances that were actually *stressed* was high. The recall (i.e., the proportion of actual *stressed* instances that were classified as *stressed*), however, is poor across all activities, resulting in low F1 scores. This suggests that without any knowledge of the context, just the physiological features are able to accurately detect some of the *stress* events, but they miss out on many others. This is because the models are trying to detect 'perceived stress', which might not always translate to a physiological response.

When we add the contextual features, however, the model (M_{c+p}) gets some additional knowledge on how to better detect the 'perceived stress' levels. As we can observe in Figure 3b, while the precision is high across all activities, the recall in certain activities, e.g., Dine, Rest, and Entertainment, is low, along with extremely low False Positive Rate (< 0.05). It is interesting to note that the all three activities (Dine, Rest, and Entertainment) have been associated with lower 'perceived stress' levels, based on earlier evaluations. One possible reason for such a trend could be that based on the contextual and physiological features the model outputs a low probability of an instance (in the above-mentioned activities) being *stressed*, and with our default classification threshold (T_d) of 0.5, the model fails to classify several stressful instances. We believe that if we tune or optimize the classification threshold for each context, we would be able to achieve better results during those activities. It is important to note that we are not discussing about building separate models for each activity, but in the same M_{c+p} model use different thresholds based on the activity of the user for an instance to be classified as *stress*.

We optimize for F1 score (to have a balance between precision and recall, as F1 score is the harmonic average between recall and precision of inferring stress arousal). To this end, we plot the variation of F1 scores across different activities for different threshold, as shown in Figure 4. It is clearly evident from the plot that a default threshold of 0.5 is not ideal. Further, it shows that our previous reasoning for lower performance in some activities (Dine, Rest, and Entertainment), was indeed true. For these three activities, a custom threshold of 0.36, 0.40 and 0.26 results in an increased F1 score of 0.641, 0.637, and 0.674, respectively. A detailed representation of the optimal threshold for individual activities,

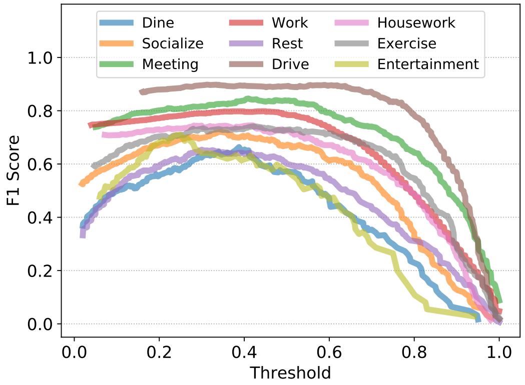


Figure 4: F1 score for individual activities for different classification thresholds (T).

	Custom Threshold	Updated F1	Previous F1
Dine	**0.36**	**0.641**	0.563
Socialize	**0.42**	**0.718**	0.680
Meeting	**0.43**	**0.839**	0.835
Work	**0.50**	**0.786**	0.786
Rest	**0.40**	**0.637**	0.591
Drive	**0.56**	**0.900**	0.890
Housework	**0.34**	**0.747**	0.705
Exercise	**0.36**	**0.748**	0.729
Entertainment	**0.26**	**0.674**	0.584

Table 2: Representation of the optimal thresholds for each activity and the resulting improved F1 score in contrast to the previous F1 score when the model had a default threshold = 0.5

with the updated F1 scores in contrast to the previous F1 scores is shown in Table 2. We observe that a minor change in the classification process (of adjusting the threshold according to the activity) would in higher F1 in each activity. While we did not build a new model incorporating individual threshold based classifications, we expect it to achieve better results (F1 score, precision, and recall) as compared to the results obtained by model M_{c+p} (Table 1). We leave the building and evaluation of the new model for future work. In the future, we also plan on testing different classifiers for different activities or for different times of day, etc.

Having done some exploratory analysis on how combining contextual features with physiological features affect binary classification, we move further to a regression based detection of stress. In this case instead of a binary classification between *stressed* and *not stressed*, we use the perceived stress scores reported by the participants directly as the outcome class, and use Random Forest regression for detection on a scale of [0,5]. The results are similar to that of binary classification, where the combination of contextual and physiological features result in the highest correlation coefficient of $r = 0.72, p < 0.001$. We show the comparison between correlation coefficients and Mean Absolute Error (MAE) for the different models in Table 3. We compare these results with that of a baseline classifier – which predicts the median stress score for each instance, regardless of the features.

DISCUSSION AND CONCLUSION

All our evaluations clearly highlight the importance of contextual information in stress detection. Our results show that using a combination of contextual and physiological features always leads to better detection results, in both, binary classification of stress (Table 1 and regression based detection 3, as compared to using just physiological signals or just contextual features. While results between different studies cannot be compared directly, the prediction results we obtain (by incorporating contextual features) are higher than the results obtained by other *in-the-wild* studies. Further, we hope that more sophisticated modeling techniques, or optimization of the current models can definitely help boost the results even further. We discuss one such case of optimization, where we show that different classification thresholds (based on the users' activity) could lead to better binary prediction scores. In the future, we intend to do more evaluations and optimization, to be able to better the current results and attempt to come closer to the *in-lab* results by other researchers.

It is also interesting to note that R-R interval was the only physiological signal used. We hope that adding another source of physiological signal, e.g., RIP, or GSR, as used in previous works, would lead to an improvement in performance [1, 3, 6]. Finally, our current evaluations have given us enough confidence to accept that context is indeed an important factor, and in the future, we intend on conducting studies where along with physiological signals, the contextual information will be collected passively.

	Correlation Coefficient	Mean Absolute Error
Physiological Features	$r = 0.40, p < 0.001$	0.76
Contextual Features	$r = 0.37, p < 0.001$	0.78
Contextual + Physiological Features	$\mathbf{r = 0.72, p < 0.001}$	**0.53**
Baseline Classifier	$r = -0.01, p = 0.09$	0.89

Table 3: Results from the regression based model for different sets of features.

REFERENCES

[1] Begum Egilmez, Emirhan Poyraz, Wenting Zhou, Gokhan Memik, Peter Dinda, and Nabil Alshurafa. 2017. UStress: Understanding college student subjective stress using wrist-based passive sensing. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 673–678. https://doi.org/10.1109/PERCOMW.2017.7917644

[2] Maurizio Garbarino, Matteo Lai, Simone Tognetti, Rosalind Picard, and Daniel Bender. 2014. Empatica E3 - A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *Proceedings of the 4th International Conference on Wireless Mobile Communication and Healthcare - "Transforming healthcare through innovations in mobile and wireless technologies"*. ICST. https://doi.org/10.4108/icst.mobihealth.2014.257418

[3] Martin Gjoreski, Hristijan Gjoreski, Mitja Luštrek, and Matjaž Gams. 2016. Continuous Stress Detection Using a Wrist Device: In Laboratory and Real Life. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct (UbiComp)*. ACM Press, 1185–1193. https://doi.org/10.1145/2968219.2968306

[4] Martin Gjoreski, Mitja Luštrek, Matjaž Gams, and Hristijan Gjoreski. 2017. Monitoring stress with a wrist device using context. *Journal of Biomedical Informatics* 73 (sep 2017), 159–170. https://doi.org/10.1016/j.jbi.2017.08.006

[5] Tian Hao, Henry Chang, Marion Ball, Kun Lin, and Xinxin Zhu. 2017. cHRV uncovering daily stress dynamics using bio-signal from consumer wearables. *Studies in Health Technology and Informatics* 245 (2017), 98–102. https://doi.org/10.3233/978-1-61499-830-3-98

[6] 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. *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)* (2015), 493–504. https://doi.org/10.1145/2750858.2807526

[7] Bruce S. McEwen and Eliot Stellar. 1993. Stress and the individual: mechanisms leading to disease. *Archives of Internal Medicine* 153, 18 (1993), 2093–2101.

[8] Varun Mishra, Gunnar Pope, Sarah Lord, Stephanie Lewia, Byron Lowens, Kelly Caine, Sougata Sen, Ryan Halter, and David Kotz. 2018. The Case for a Commodity Hardware Solution for Stress Detection. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct (UbiComp'18)*. ACM. https://doi.org/10.1145/3267305.3267538

[9] Motorola. 2018. Moto 360. (2018). https://www.motorola.com.au/products/moto-360

[10] R. Rosmond and P. Björntorp. 1998. Endocrine and metabolic aberrations in men with abdominal obesity in relation to anxio-depressive infirmity. *Metabolism* 47, 10 (1998), 1187–1193.

[11] Akane Sano and Rosalind W. Picard. 2013. Stress Recognition Using Wearable Sensors and Mobile Phones. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 671–676. https://doi.org/10.1109/ACII.2013.117

[12] Hillol Sarker, Inbal Nahum-Shani, Mustafa Al'Absi, Santosh Kumar, Matthew Tyburski, Md Mahbubur Rahman, Karen Hovsepian, Moushumi Sharmin, David H. Epstein, Kenzie L. Preston, C. Debra Furr-Holden, and Adam Milam. 2016. Finding Significant Stress Episodes in a Discontinuous Time Series of Rapidly Varying Mobile Sensor Data. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)*. ACM Press, 4489–4501. https://doi.org/10.1145/2858036.2858218

[13] Yonatan Vaizman. 2017. Context Recognition In-the-Wild: Unified Model for Multi-Modal Sensors and Multi-Label Classification. *PACM Interact. Mob. Wearable Ubiquitous Technol* 1, 1 (jan 2017), 1–22. https://doi.org/10.1145/3161192