UnStressMe: Explainable Stress Analytics and Self-tracking Data Visualizations

Eva Paraschou
Aristotle University of Thessaloniki
Thessaloniki, Greece
eparascho@csd.auth.gr

Sofia Yfantidou
Aristotle University of Thessaloniki
Thessaloniki, Greece
syfantid@csd.auth.gr

Athena Vakali
Aristotle University of Thessaloniki
Thessaloniki, Greece
avakali@csd.auth.gr

***Abstract*—Self-tracking technology for behavior monitoring is prevalent in various aspects of human life. It enables users' activities tracking with data produced "in the wild", namely capturing real-world physical activity, sleep patterns, and stress levels, among others. Advanced new sensors integrated into commercial self-tracking devices have empowered a new era of sensing data exploration and self-improvement recommendations, aiming to enhance physical and mental well-being. However, the collected data and related inferred knowledge are not always well-explained or well-presented and discourage users' commitment leading to sensing devices' abandonment. To sustain user engagement with self-tracking technology for well-being, this paper introduces a comprehensive framework and respective full-stack web service called "UnStressMe" for the analysis of diverse data modalities tracked in the wild, the prediction of future stress behavior and the production and provision of personalized, model-agnostic explanations and interactive visualizations. We showcase the utility of our framework through a mental health use case, paving the way for explainable, transparent, and human-centric self-tracking technology.**

***Index Terms*—Personal Informatics, XAI, ML, Visualizations, A/B testing, Wearable Technology**

I. Introduction

Nowadays, as pervasive technology is constantly and rapidly evolving, more and more people are using self-tracking (ST) devices, such as wearable activity trackers and smartwatches, in their everyday life, capturing and monitoring physical activity, health, and sleep, among others. However, 50% of them abandon their ST device within two weeks [1], and many others state that they face several effectiveness issues, including a mismatch between self-perception and tracked data, limited data utility, and high maintenance requirements [2]. Despite the ACM Principles for Algorithmic Transparency and Accountability emphasizing data utility through explainability [3], most commercial ST technology still adopts a basic approach to data serving. Specifically, it often displays raw information (i.e., data gathered by the source without further processing) in a textual format (e.g., step counts) to represent users' behavior [2]. Yet, what can a user truly understand from such nearly unprocessed, out-of-context data? Could raw information be converted into meaningful and essential knowledge to improve users' perceived utility and, ultimately, personal engagement?

It's rather challenging to answer these critical questions since the pervasive wearable analytics domain struggles with some important limitations: *[L1] significant lack of open data* to be exploited for model training due to privacy-sensitive personal data considerations; *[L2] complex and noisy processes* to represent and model even published, anonymized datasets; *[L3] lack of systematic guidelines on applying XAI models* which will extend traditional ML models to provide meaningful explanations and visualizations about wearable analytics (an interdisciplinary, rapidly changing domain); *[L4] very limited quantitative approaches*, such as A/B testing, exist since most user studies assessing the effectiveness of explainability on wearable analytics are based on self-reported surveys and interviews.

The above limitations have been partially addressed in prior work, especially in providing meaningful and interpretable knowledge through ML and XAI models. Most existing work has relied on textual explanations, and lack of descriptive visualizations has diminished interpretability [4], [5]. Even with works that have provided more generic visualizations, they mainly focus on static (i.e., questionnaires, interviews) rather than dynamic data from explainable models [6], [7]. To the best of our knowledge, there exists no earlier work that applies XAI models and visualizes their produced explanations in pervasive wearable analytics, especially in the emerging domain of mental health analytics.

This paper addresses the above limitations ([L1]-[L4]) by introducing UnStressMe, a framework with its full-stack web service peaking into the triggers behind users' behavioral or emotional responses, such as stress. Mental health analytics has been prioritized as an indicative use case, not only for its social importance but also as an emerging research domain aligned with wearable analytics trends [8]. However, the UnStressMe framework and pipeline can be generalized to other subdomains within ST (e.g., physical activity or sleep). In summary, the contributions of this work are as follows:

Newly-released, in-the-wild dataset: We introduce for the first time to the community and analyze the LifeSnaps dataset [9], a recently published, multi-modal, real-world dataset of ST data, self-reported user labels and psychological and behavioral surveys, tackling limitations *L1* and *L2*.

ML and XAI models for wearable analytics: We propose an ML model that predicts the next day's stress level and explains the rationale behind specific predictions' production, using the state-of-the-art Local Interpretable Model-Agnostic

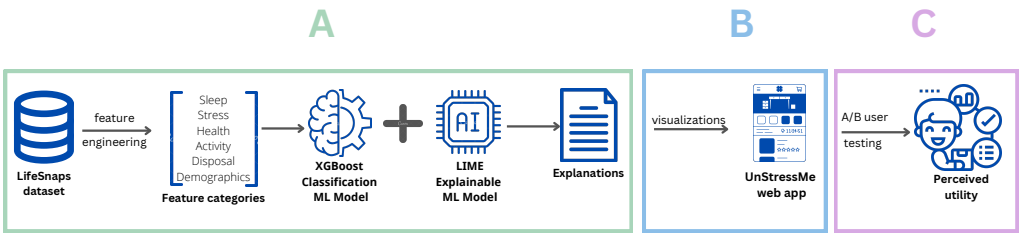

Fig. 1. The UnStressMe pipeline

Explanations (LIME) explainer [10], tackling limitation *L3*.

Visualizations & A/B user testing: We accompany the produced explanations with diverse, interactive visualizations in the UnStressMe web application and evaluate the complete pipeline through a user study adopting an A/B testing approach, tackling limitation *L4*.

II. The UnStressMe pipeline

This section presents the flow of the UnStressMe pipeline, as seen in Figure 1, from the introduction of the novel LifeSnaps dataset to the ML and XAI models and their integration into the UnStressMe web application.

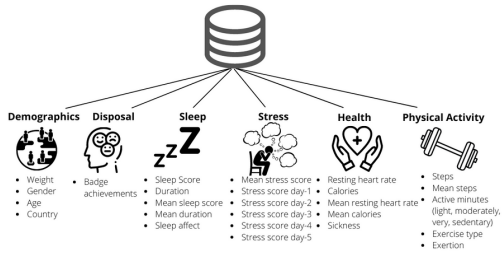

Fig. 2. Final selected features and categories

A. LifeSnaps Dataset

Our models are developed on the LifeSnaps dataset [9], a real-world, multi-modal dataset containing 71M rows of data collected from 71 users, in four countries, from May '21 to Feb. '22. LifeSnaps includes anonymized data from Fitbit Sense wearable devices in different granularities, such as stress score, sleep duration, calories, steps, active minutes, oxygen variation, and heart rate among others. To ensure the quality of our training data, we first apply common preprocessing steps (e.g., missing values imputation, type transformation, scaling, outlier removal, resampling, and duplicate removal) and feature engineering to identify the features most relevant to our mental health use case. Through this process, we introduce six feature categories from the raw Fitbit data: demographics, disposal, sleep, stress, health, and physical activity, totaling 29 features, which are displayed in detail in Figure 2.

B. ML and XAI Models in Persvasive Wearable Analytics

The next step of our UnSressMe pipeline, adapted to our use case, is to predict the next day's stress level. To this end, we utilize the Gradient Boosting algorithm, an ensemble classifier with high execution speed and model performance [11], to train our predictive model, whose parameters have been decided through hyperparameter tuning with a 5-cross validation based on the accuracy score on the validation set. Our predictive model reaches the final accuracy score of 92.3% on the held-out test set. The model's predictions are then passed to the second step of our pipeline, namely the computation of explanations. To this end, we apply the state-of-the-art LIME algorithm [10] that provides explanations with up to 80% confidence in our case. LIME can be applied to any black-box model to produce human-friendly explanations, and each prediction is reliable in its neighborhood.

Our proposed approach is based on the ML & XAI models, which are outlined above. In general, LIME computes for each classification instance the feature importance for the specific prediction, enabling us to identify the most crucial features for the stress prediction use case. To summarize, stress levels in our data are mainly affected by activity-related physical exertion and sleep pattern features. In addition, stress levels are partially affected by previous days' stress levels, capturing within-user differences, and by the users' usual stress levels, highlighting between-user differences in stress expression.

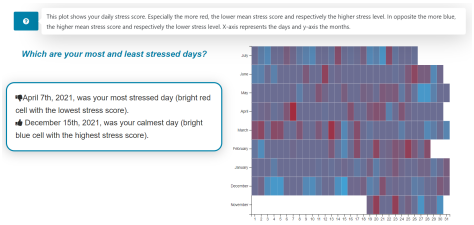

Fig. 3. The plot shows the user's stress distribution during a specific period

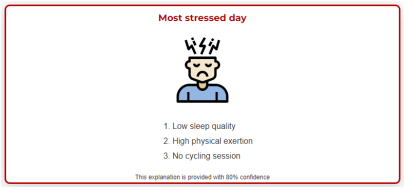

Fig. 4. The plot shows what happened on the user's most stressed day

C. UnStressMe Web Application

To conclude the UnStressMe pipeline, we develop a publicly available, full-stack web application (https://eparascho.github.io/), whose main goal is to provide human-friendly visualizations to explained predictions produced by Gradient Boosting and LIME. It creates person-specific interactive visualizations both on feature importance, as shown in Figure 3, and on LIME explanations, as shown in Figure 4. Such visualizations facilitate the understanding of a user's behavior and increase the results' transparency by supporting their interpretability. All the above are implemented with JavaScript and the Material UI and ObservableHQ libraries.

D. A/B User Testing Evaluation

To evaluate the perceived utility of the UnStressMe explanations and visualizations and the factors affecting user engagement, we resort to A/B user testing. We randomly distribute two versions of the web application to two different user groups. Version A constitutes a close copy of the original Fitbit stress tracking dashboard (https://fitbitstressmanagement.netlify.app/). Version B is described in detail in the previous subsection and constitutes a more explainable, human-centric, and transparent version of Version A. To assess their perceived utility, we also share a set of validated surveys capturing: 1) demographics and ST device usage, 2) stress assessment based on Perceived Stress Scale [12], and 3) explanation goodness based on Explanation Goodness Checklist [13]. The study has been reviewed and approved by the Institutional Review Board (IRB) with protocol number 151316/2022.

After collecting the results of the surveys, we apply Principal Component Analysis for reliability analysis, T-Tests, and Multivariate Analysis of Variance. Through the T-Tests, we reveal that gender plays a statistically significant role ($p = 0.01$) in how users perceive visualizations, while the MANOVA test reveals that the interest in activity tracking plays a weak statistically significant role in shaping users' points of view ($p = 0.068$). However, we have not found a statistically significant difference between the two application versions, possibly due to certain study design limitations.

III. Limitations & Conclusions

In this work, we study one of the most unexplored domains in wearable analytics: stress level prediction, explanation, and visualization. For this purpose, we utilize, for the first time in literature, LifeSnaps, an open, real-world dataset. To predict the next day's stress level, we find that the Gradient Boosting algorithm is the most suitable classifier for our purposes, with an accuracy of 92.3%, while we utilize the LIME algorithm to produce explanations. Moreover, to provide meaningful and interactive visualizations of the above explanations, we create the UnStressMe web application, uncovering discrepancies in perceived utility between users of different gender and digital competency. There are many limitations due to time restrictions, which provide grounds for future work. At first, a larger, more diverse sample for both data collection and the A/B user test could provide more generalizable information, while other explainers can be tried for more accurate and trustworthy explanations. Finally, the Fitbit API does not provide any information about stress in its endpoints, not allowing us to create a dynamic web application where users could interact with their data. Nevertheless, we believe this study to be a stepping stone to the ST domain and expressly stress interpretation.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 813162. The content of this paper reflects only the authors' view and the Agency and the Commission are not responsible for any use that may be made of the information it contains.

References

[1] H. Lee and Y. Lee, "A look at wearable abandonment," in *2017 18th IEEE Int. Conference on Mobile Data Management (MDM)*. IEEE, 2017, pp. 392–393.

[2] A. Lazar, C. Koehler, T. J. Tanenbaum, and D. H. Nguyen, "Why we use and abandon smart devices," in *Proc. of the 2015 ACM Int. joint conference on pervasive and ubiquitous computing*, 2015, pp. 635–646.

[3] A. U. P. P. Council, "Statement on algorithmic transparency and accountability," *Commun. ACM*, 2017.

[4] P. W. Woźniak, P. P. Kucharski, M. M. de Graaf, and J. Niess, "Exploring understandable algorithms to suggest fitness tracker goals that foster commitment," in *Proc. of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, 2020, pp. 1–12.

[5] A. Sano, S. Taylor, C. Ferguson, A. Mohan, and R. W. Picard, "Quantifyme: An automated single-case experimental design platform," in *Int. Conference on Wireless Mobile Communication and Healthcare*. Springer, 2017, pp. 199–206.

[6] D. Epstein, F. Cordeiro, E. Bales, J. Fogarty, and S. Munson, "Taming data complexity in lifelogs: exploring visual cuts of personal informatics data," in *Proc. of the 2014 conference on Designing interactive systems*, 2014, pp. 667–676.

[7] J. Niess, K. Knaving, A. Kolb, and P. W. Woźniak, "Exploring fitness tracker visualisations to avoid rumination," in *22nd Int. Conference on Human-Computer Interaction with Mobile Devices and Services*, 2020, pp. 1–11.

[8] D. A. Epstein, C. Caldeira, M. C. Figueiredo, X. Lu, L. M. Silva, L. Williams, J. H. Lee, Q. Li, S. Ahuja, Q. Chen *et al.*, "Mapping and taking stock of the personal informatics literature," *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 4, pp. 1–38, 2020.

[9] S. Yfantidou, C. Karagianni, S. Efstathiou, A. Vakali, J. Palotti, D. P. Giakatos, T. Marchioro, A. Kazlouski, E. Ferrari, and Š. Girdzijauskas, "Lifesnaps, a 4-month multi-modal dataset capturing unobtrusive snapshots of our lives in the wild," *Scientific Data*, vol. 9, no. 1, pp. 1–19, 2022.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proc. of the 22nd ACM SIGKDD Int. conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[11] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1937–1967, 2021.

[12] S. Cohen, T. Kamarck, R. Mermelstein *et al.*, "Perceived stress scale," *Measuring stress: A guide for health and social scientists*, vol. 10, no. 2, pp. 1–2, 1994.

[13] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018.