Exploring UX Frustration Recognition using Stress Recognition Models and Lightweight Biosignal Features

Sergio De León Aguilar[1][0000−0003−4731−6873], Yuki Matsuda[2][0000−0002−3135−4915], and Keiichi Yasumoto[3][0000−0003−1579−3237]

[1] Nara Institute of Science and Technology, Ikoma City Nara 630-0192, Japan
sergio.de.leon@ubi-lab.com
[2] Okayama University, Okayama City Okayama 700-0082, Japan
cocolab@okayama-u.ac.jp
[3] Nara Institute of Science and Technology, Ikoma City Nara 630-0192, Japan
yasumoto@is.naist.jp

Abstract. Adaptive user interfaces provide personalized experiences traditionally through customization menus, user profiling, or multi-modal sensing. Stress and cognitive-load-aware AUIs have received significant attention due to their potential to improve accessibility and performance for less tech-savvy users. Particularly, the use of wearable-captured physiological signals allows practical implementations with high usability. However, attention to bad user experience due to heterogeneous skill levels is limited in the literature; studies are commonly found interpreting stress and frustrating user experiences (UX) as equal, and many approaches rely on self-reported or post-hoc analyses. In this study, we investigate the challenges lightweight machine learning models face when inferring UX-induced frustration from wearable physiological signals using stress-trained datasets, highlighting how task specificity achieves up to 19.3% performance gains in frustration recognition tasks, and label granularity makes 4-class classification viable at above-random chance performance (42.3%).

Keywords: UX · frustration · wearables · machine learning.

1 Introduction and Motivation

Adaptive User Interfaces (AUIs) are commonly designed for homogeneous populations such as trained operators, students, or users with specialized expertise. While early AUIs emphasized static personalization (e.g., adapting layouts to user expertise), recent HCI research has explored dynamic adaptation based on users' accessibility needs to reduce cognitive load and prevent confusion during challenging tasks [3]. Despite performance gains, most systems overlook the wide variation in technological familiarity, perceptual abilities, and emotional reactivity found in the general public, which intensifies marginalization of vulnerable or less tech-savvy users [6]. To address these considerations, several studies leverage physiological stress detection to identify struggling behavior during interaction [5]. Although promising, these approaches are often constrained by limited subject diversity and by ambiguously defined stress labels that may conflate multiple emotional states.

In this study, we evaluate stress-trained multimodal traditional machine learning models on smartphone user experience (UX) induced realistic frustration events. We demonstrated that by increasing task specificity, accuracy improvements as great as 19.3% can be achieved, and 4-class classification becomes viable by reducing emotion-ambiguity.

2 Related Work

Recent advances in physiological and behavioral sensing have enabled AUIs to adapt interface complexity, pacing, and visual emphasis based on cognitive load. Vasiljevas *et al.* showed that gaze behavior can continuously estimate cognitive load during interaction, enabling anticipatory difficulty adaptation [7]. Viviani *et al.* reported strong correlations between fine motor behavior in smartphone typing and cognitive function, with implications for learning and accessibility [8].

Although stress recognition is widely studied in mental health literature, wearable-based adaptation remains an emerging approach. Most stress recognition datasets rely on coarse temporal windows and post-hoc analysis, with few approaches supporting semi-real-time inference [4]. Liapis *et al.* demonstrated mobile-compatible stress classification for UX evaluation using physiological signals; however, their approach does not address subject or task heterogeneity, limiting generalization in realistic settings.

In contrast, our work directly examines the feasibility of real-time, on-device UX-induced frustration recognition by evaluating stress-trained models against both standardized stress tasks and a realistic, frustration-inducing smartphone interaction, thereby addressing the combined challenges of task specificity, emotion ambiguity, and practical deployment constraints.

3 Experiment and Results

To evaluate the performance of traditional stress recognition methods on a frustration-inducing realistic smartphone interface, we modified an open source calculator application[4] to reproduce UX friction points identified by elderly subjects from a previous experiment on AR-based guidelines for the general public [2]. Implemented friction points include reduced readability, random input delays, missed interactions, and hard-to-access operator keys. To reduce the presence of motion artifacts and increase reproducibility, video-based tasks were selected as the baselines to be compared against. 21 subjects participated in our experiment (6F/15M).

PPG and EDA physiological reactions were sampled with a Shimmer3 GSR+ wearable device at 51.2Hz during the execution of arithmetic calculations with

[4] `https://github.com/violiarns28/calculator`

and without UX bugs present, and during video-based stress and relaxing elating video clips according to the multi-modal StressID dataset protocol [1]. Every phase was preceded by an emotional reset via a deep breathing and relaxation activity as seen in Fig. 1.

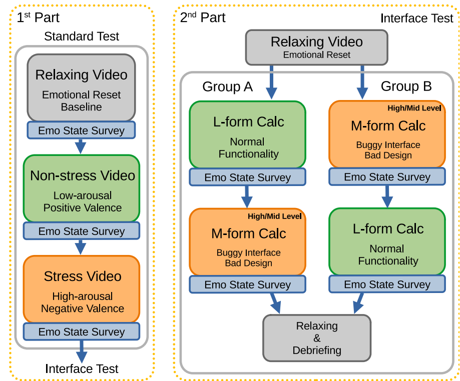


Fig. 1: UX frustration recognition protocol

Self-reported direct emotion (stressed and relaxed) and a 3-axis SAM manikin survey were conducted to produce our ground-truth. Three classification schemes were tested: non-stressed vs stressed (binary); relaxed, stressed, and undefined (ternary); and relaxed, amused, stressed, and undefined (quaternary).

For evaluation, each model would be trained on all tasks from StressID's protocol and only on their video-based tasks. Inference performance would be tested on our experimental data video tasks (stress) and calculator tasks (UX frustration). Feature oversampling and subject standardization followed StressID's updated feature pipeline.

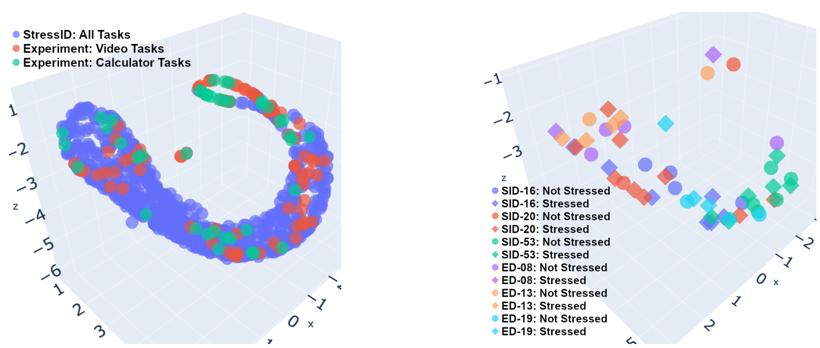
We observe classification performance being severely affected by experimental setup—binary classification performance is reduced on unseen data, our experiment video tasks—, performance is improved by training the models only on the relevant StressID's video tasks, as seen in the difference between validation and video task ($VTasks$) rows in Table 1. To visualize the factors affecting feature clustering in our experiment and StressID's data, we run a t-SNE analysis. As seen in Fig. 2, regardless of a uniform distribution of all tasks in the feature space, the greatest clustering depends on the subject rather than label. To increase specificity, we added a new classification schema, in which by considering the "Dominance" axis and redifining "Arousal"'s relevance, we successfully achieved above-random classification power (42.3%) in the hardest 4-label classification task for the calculator task ($CTasks$).

4 Conclusion and Limitations

Though the authors recognize that generalization may be affected by the reduced sample size and the need for further tasks comparison to better discern the nature of UX-induced frustration, this report reveals the importance of label specificity on under-studied emotion recognition tasks. Reducing emotion ambiguity can achieve great improvements (19.3%) in difficult classification tasks such as 4-label classification, showing the potential of lightweight machine learning models for practical AUIs on the edge.

Table 1: Baseline models' accuracy performance comparison between models trained with all StressID's tasks and trained only with the relevant video tasks.

		KNeighbors			MLP			RandomForest			SVC		
Cls.		All	Video	Diff.	All	Video	Diff.	All	Video	Diff.	All	Video	Diff.
Bin.	Validation	86.5	83.8	-2.7↓	86.1	83.1	-3.0↓	84.4	87.1	2.7↑	86.4	89.9	3.5↑
	VTasks	51.1	66.7	15.6⋆	49.4	56.7	7.3⋆	55.6	54.4	-1.2↓	52.2	48.9	-3.3↓
	CTasks	50.0	54.2	4.2↑	45.4	50.5	5.1↑	44.9	50.0	5.1↑	43.1	50.5	7.4⋆
Ter.	Validation	79.1	86.1	7.0⋆	77.0	87.4	10.4⋆	78.3	88.3	10.0⋆	78.4	91.3	12.9⋆
	VTasks	44.3	47.2	2.9↑	34.2	43.2	9.0⋆	42.0	47.9	5.9↑	41.2	32.6	-8.6×
	CTasks	31.5	39.9	8.4⋆	40.0	30.8	-9.2×	32.5	31.6	-0.9↓	37.6	35.7	-1.9↓
Qad.	Validation	79.7	82.0	2.3↑	76.3	80.3	4.0↑	75.2	80.1	4.9↑	79.7	86.3	6.6↑
	VTasks	21.0	23.9	2.9↑	32.7	33.3	0.6↑	24.6	30.0	5.4↑	29.2	23.7	-5.5↓
	CTasks	34.1	42.3	8.2⋆	10.9	30.2	**19.3**⋆	22.1	33.0	**10.9**⋆	23.1	27.6	4.5↑



(a) All tasks grouping. (b) All tasks, 6 subjects subset.

Fig. 2: t-SNE analysis on StressID (SID) and our experiment's data (ED).

References

1. Chaptoukaev et al, H.: Stressid: a multimodal dataset for stress identification. In: Advances in Neural Information Processing Systems. vol. 36, pp. 29798–29811 (2023)
2. De León Aguilar, S., Matsuda, Y., Yasumoto, K.: Mobile Augmented Reality Interface for Instruction-based Disaster Preparedness Guidelines. Sensors and Materials **36**(10), 4585–4606 (2024)
3. Heumader, P., Miesenberger, K., Murillo-Morales, T.: Adaptive user interfaces for people with cognitive disabilities within the Easy Reading framework. In: Lecture Notes in Computer Science. vol. 12377, pp. 53–60 (2020)
4. Jaiswal, D., Chatterjee, D., B s, M., Ramakrishnan, R.K., Pal, A.: GSR Based Generic Stress Prediction System. In: Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing. pp. 433–438 (2023)
5. Liapis, A., Faliagka, E., Katsanos, C., Antonopoulos, C., Voros, N.: Detection of Subtle Stress Episodes During UX Evaluation: Assessing the Performance of the WESAD Bio-Signals Dataset. In: Human-Computer Interaction. pp. 238–247. INTERACT '21 (2021)

6. Sin, J., L. Franz, R., Munteanu, C., Barbosa Neves, B.: Digital Design Marginalization: New Perspectives on Designing Inclusive Interfaces. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–11 (2021)
7. Vasiljevas, M., Damaševičius, R., Maskeliūnas, R.: A Human-Adaptive Model for User Performance and Fatigue Evaluation during Gaze-Tracking Tasks. Electronics **12**(5) (2023)
8. Viviani, L., Liso, A., Craighero, L.: Mobile Typing as a Window into Sensorimotor and Cognitive Function. Brain Sciences **15**(10) (2025)