Predicting Stress and Providing Counterfactual Explanations: A Pilot Study on Caregivers

Kei Shibuya[12], Zachary D King[1], Maryam Khalid[1], Han Yu[1], Yufei Shen[3], Khadija Zanna[1], Ryan L Brown[4], Marzieh Majd[56], Christopher P Fagunders[7], Akane Sano[1]
[1]*Dept. of Electrical and Computer Engineering*, *Rice University*, Houston, U.S.
(Kei.Shibuya, zdk2, mk79, hy29, kz35, Akane.Sano)@rice.edu

***Abstract*—Caregiving for spouses with Alzheimer's disease or related dementias (ADRD) is one of the most stressful experiences. Evidence-based treatments for caregivers who have a high risk of mental health issues are needed. In this study, we designed models for predicting changes in perceived stress scale (PSS) (increase/not increase) in one week and generated some examples of counterfactual ('what-if') explanations to change the stress state for helping manage stress. Using self-report (positive and negative affect and sleep quality) and sensor data (heart rate, sleep, and steps) collected in 132 week-long study sessions from 57 participants, we compared explainable PSS change prediction models (Random Forest, XGBoost, LightGBM, EBM, and Neural Network) along with 'what-if' explanations. First, we developed machine learning models for classifying the change in PSS scores before and after the session period. Second, we identified feature importance using our explainable models. Our results showed that XGBoost performed the best with an accuracy of 0.79 and an F1 score of 0.78 for predicting changes in perceived stress. Our results also showed that minimum heart rate, mean steps per day, and negative affect are the most predictive features. Our preliminary counterfactual examples about sleep parameters would be able to provide suggestions for improving one's health. We discussed our ideas to provide better suggestions using DiCE.**

***Index Terms*—caregiver, healthcare, XAI, DiCE, PSS**

I. Introduction

The number of people with Alzheimer's disease (AD) and AD-related dementias (ADRD) in the United States is projected to grow from 5.7 million to 13.8 million by 2050 [1]. Caregiving for ADRD patients is challenging and burdensome, especially for those lacking financial resources, time, or knowledge to provide sufficient care while looking after their own health. Developing evidence-based treatments for high-risk caregivers is crucial.

Affect and stress inference using machine learning (ML) models and wearable devices have been well documented in literature. Gedam et al. reviewed over 55 papers related to stress detection from wearables [2] and showed dozens of combinations of sensors, devices, models, and stress constructs. Stress detection can be split into two groups those predicting distal outcomes (end-of-study stress scales) [3] and proximal outcomes (momentary stress) [4]. Most stress detection studies focus on stress levels at time points. Few researchers have looked at change (increase) in stress as the construct for detection, determing when and why stress is increasing more useful to give proactive care. While many of these studies have prioritized predicted accuracy using black-box models, healthcare is one of the sectors, which is increasingly recognizing the significance of explainability. According to a review of explainable AI (XAI) [5], this has led to a growing trend of using a combination of black-box and explainable approaches or interpretable/white-box models. Previous studies utilized XAI models (e.g., SHAP [6], LIME [7]) to detect/predict stress, revealing how strong features contribute to the outcomes. But it was difficult to decide which actions should be taken to change the outcomes. A new algorithm called diverse counterfactual explanation (DiCE) [8] has been developed to provide suggestions. DiCE provides 'what-if' explanations, indicating which features need to be changed and how much for better outcomes.

Our ultimate goal is to design a technology that can support people to take action to manage their health conditions. We focus on ADRD elderly caregivers in this study. Understanding the causes of poor health and what should do next for improvement is important. To provide explanations and suggestions to people who have a high risk of stress, we collected data from 57 ADRD caregivers. Then we developed perceived stress change prediction models for classifying changes in stress using binary labels, identified feature importance using explainable ML models, and generated explainable counterfactual examples using DiCE.

Among the models we developed, XGBoost model performed the best with an accuracy of 0.79 and an F1 score of 0.78 for predicting changes in perceived stress. Minimum heart rate, means of steps, and negative affect were the most predictive features. We also showed a preliminary example of providing suggestions on what users should do to improve their perceived stress using DiCE's counterfactual explanations.

II. Materials and Methods

A. Dataset

The experiment was conducted to collect data remotely from caregivers of AD/ADRD patients from March 2021 to

[2]*Biometrics Research Laboratories. NEC Corporation*, Tokyo, Japan, kei-shibuya@nec.com
[3]*Dept. of Electrical and Computer Engineering, The University of Texas*, Austin, U.S., shenyufei@utexas.edu
[4]*Dept. of Psychiatry and Behavioral Sciences, University of California San Francisco*, San Francisco, U.S., Ryan.Brown@ucsf.edu
[5]*Dept. of Psychiatry, Harvard Medical School*, Boston, U.S., mmajd@bwh.harvard.edu
[6]*Mood and Psychosis Research Program, Dept. of Psychiatry, Brigham and Women's Hospital*, Boston, U.S.
[7]*Dept. of Psychological Science, Rice University*, Houston, U.S., Christopher.Fagundes@rice.edu

May 2023 (IRB-FY2021-65). Participants joined the study for seven days each for three sessions one month apart over three months where they received online questionnaires and wore the Fitbit (Figure1). Inclusion criteria: AD/ADRD spousal caregivers were self-identified as the principal person taking care of the spouse with a physician-based diagnosis of AD, and devoting at least 4 hours daily to the care of the spouse for at least the last three months. The AD/ADRD spousal caregiver & AD/ADRD patient must have been married or (or self-defined as a long-term committed partner) for at least 3 years. AD/ADRD patients met clinical dementia rating criteria for mild or moderate dementia at the screening. AD/ADRD spousal caregivers were preliminarily screened for eligibility over the telephone. Exclusion criteria: Psychiatric disorders were evaluated by the Mini-International Neuropsychiatric Structured Clinical Interview (MINI SCID) [9] at screening. We excluded those who exhibit psychotic symptoms; or a history of suicide attempts within the last year, acute or uncontrolled medical illnesses (e.g., major surgery, metastatic cancer, class III heart failure, and inflammatory disorders), those with a BMI over 40, those using hormone-containing medications, including steroids or immune-modifying drugs, those on daily analgesics such as opioids.

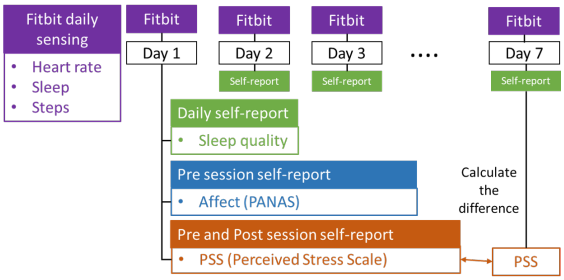

Fig. 1. A Structure of the dataset.

The modeling and analysis presented in this paper are from the data of 57 caregivers. The reported gender percentage was female 82.5% and the average age was 61.5 (SD=12.5) years old. The number of participants who joined 1-time session was 49, 2-time sessions was 43, and 3-time sessions was 40. We obtained 132 sets from three sessions.

1) Ground truth: Labels for our stress change prediction models are derived from the pre and post session scores of the Perceived Stress Scale (PSS) [10]. PSS includes 10 items rated on a Likert scale from 0 (never) to 4 (very often), in which the higher scores indicate higher levels of chronic stress. Binary ground truth was created based on the difference between pre-session PSS (the beginning of the session) and post-session PSS (the end of the session) scores (Figure 1). Depending on the change in PSS from pre to post, we labeled each session of each participant as "Increase" (Pre-PSS $<$ Post-PSS, n=56) or "Non-Increase" (Pre-PSS $\geq$ Post-PSS, n=76). Identifying even a 1-point change is crucial in preventing stress-related diseases, especially among high risk caregivers.

2) Measurements: To measure affect, the Positive and Negative Affect Schedule (PANAS) [11] was used. Participants were asked to answer the PANAS on pre-session and the question 'How would you rate your sleep quality last night?' (0 to 100) to measure subjective sleep quality every morning. Participants were asked to wear the Fitbit during each session. Intraday data on a minute-by-minute frequency were downloaded using a Fitbit developer account. The features extracted from the Fitbit are based on heart rate (average heart rate), sleep (sleep onset, offset, sleep/wake states), and step count (total number of steps) at each minute.

TABLE I
A LIST OF FEATURES FOR PSS CHANGE CLASSIFICATION

Method	Freq.	Modality	Features
Self report	Pre	Affect	Posi. and Nega. (mean)
	Daily	Sleep	Sleep quality (mean, std)
Sensor	Daily	HR	Mean, std, min, max, min
	Daily	Sleep	Sleep duration (mean, std), Sleep onset (mean, std), Sleep offset (mean, std), Sleep Regularity Index
	Daily	Steps	Mean, std, max, all

B. Methods

1) Feature Extraction: We extracted 5 different modalities of features from the data described in section II-A2. Table I shows a list of the features extracted from the Fitbit and self-report data. From Fitbit, we calculated 16 daily features. We calculated the average heart rate in a day for heart rate and obtained the total number of steps for steps. For sleep, 4 features were calculated for each day: sleep duration, sleep onset and offset (time of day the participant fell asleep and woke up, elapsed minutes from midnight), and sleep regularity index (SRI) which measures the consistency of the participants' sleep within previous two days [12]. Then we computed weekly statistical 20 features (mean, standard deviation, min, and max) from the daily features for our models.

2) Models: The following five algorithms were compared for the stress change classification task based on the prior work about XAI [5], (1) Random Forest (RF), (2) XGBoost [13], (3) Light Gradient Boosting Machine (LightGBM) [14], (4) Explainable Boosting Machine (EBM) [15], and (5) Neural Networks (NN) (2 layers, activation function is sigmoid). We used 10 folds cross-validation to train models with 80% of the data and to test it with the remaining. In addition, we used SMOTE [16] for over-sampling training data because the number of each ground truth group was unbalanced, and used grid search to optimize hyperparameters. (For RF, number of estimators: 10-100, max features: 'sqrt', 'log2', None, max depth: None-30). The performances of the models were evaluated using accuracy and F1 scores (Table IV).

3) Counterfactual Explanations: After evaluating and comparing classification models, we generated counterfactual explanations (CFs) using Diverse Counterfactual Explanation for ML (DiCE) [8]. DiCE provides CFs as an optimization problem by using the critical differences between features of the example and features of opposite examples. To provide

TABLE II
STATISTICS OF PSS AND FEATURES

Feature	Group	Min	Mean	Max	Std
PrePSS	Inc.	3.00	14.41	30.00	7.60
	Non-Inc.	5.00	20.79	36.00	6.71
PostPSS	Inc.	4.00	17.79	37.00	7.88
	Non-Inc.	3.00	17.59	33.00	6.33
Change in PSS	Inc.	1.00	3.38	12.00	2.71
	Non-Inc.	- 11.00	- 3.20	0.00	2.96
Sleep duration [mins]	Inc.	253.67	415.90	562.00	70.65
	Non-Inc.	180.75	406.26	547.33	78.95
Time in bed [mins]	Inc.	274.33	451.02	577.83	71.38
	Non-Inc.	229.17	444.35	585.83	77.85
Sleep onset [mins from 0:00]	Inc.	1203.71	1735.50	1431.29	90.53
	Non-Inc.	1070.33	1410.96	1727.50	128.36
SRI	Inc.	43.78	73.05	92.11	14.20
	Non-Inc.	28.52	69.92	94.38	14.55
Sleep quality	Inc.	19.00	70.24	98.14	16.42
	Non-Inc.	14.24	63.66	92.14	18.15
Steps [mean per day]	Inc.	2255.67	7301.45	21633.17	4387.17
	Non-Inc.	1611.17	6762.36	14506.00	2868.31

'what users should do next [8]' to decrease PSS or to keep the same PSS, we selected one participant in the 'Increase' group and generated CFs as preliminary results. We selected six features (SRI, sleep duration, time in bed, sleep onset, sleep quality, and steps) that are relatively modifiable and set these as 'features to vary' in DiCE. For example, changing sleep onset is relatively easier than changing heart rate. All features were used to generate CFs but only the features selected 'to vary' were allowed to change the values to obtain '(PSS) Not-increase' outcome. See statistics of these variables (Table II). We also set the range of three features to avoid unrealistic CFs based on the original outcome as follows: 5 to 8 hours for sleep duration, 8pm to 2am for sleep onset, and sleep quality was set over the original outcome. We generated 10 CFs and show some examples as preliminary results.

III. RESULTS

A. Stress and Feature Profiles

Table II shows statistics of PSS and modifiable features set 'to vary' in each group ('Inc.' vs 'Non-Inc.'). We conducted ANOVA tests after F-test. Homoscedasticity or heteroscedasticity was used on ANOVA based on results of F-test As a result, pre-PSS score was lower in 'Inc.' than in 'Non-Inc.' (F=(1, 74), p<.01), changes in PSS score were higher in 'Inc.' than in 'Non-Inc.' (F(1, 74)=192.8, p<.01). The participants were moderately stressed on average. About modifiable features, sleep quality was higher in 'Inc.' than in 'Non-Inc.' (F(1, 74)=16.4, p<.01). Table III shows Pre-Post stress level changes. We categorized stress levels as low (0-13), moderate (14-26) and high (27-40) according to PSS scale score [17]. Among 'Inc.' group, high Post-PSS level was 14.2%, moderate was 50.0%, and low was 35.7%.

B. Model Evaluations

Table IV shows the model performance in classifying the '(PSS) Increase' or '(PSS) Not-increase.' As a baseline score, 0.58 accuracy was computed by using the zero rule algorithm. XGBoost performed the best (ACC=0.79, F1=0.78), random forest performed second (ACC=0.78, F1=0.77), and LightGBM performed third (ACC=0.75, F1=0.75). Decision tree group performed well in this classifying task. Table V shows the top five feature importance of each model we evaluated. We found that features about sleep were important in each model: time in bed (XGBoost), sleep onset and sleep quality (randam forest), and sleep offset and sleep quality (LightGBM).

TABLE III
THE NUMBER OF PARTICIPANTS IN EACH PSS LEVELS

Pre \Post	Low ($\leq$ 13)	Mode	High ($\geq$ 24)
Low ($\leq$ 13)	31 (20)	10	0
Mode	9	59 (18)	2
High ($\geq$ 24)	0	10	11 (6)

Numbers inside the parentheses indicate how many participants 'Increase' PSS score while staying at the same PSS level.

TABLE IV
EVALUATION SCORES OF MODELS

Models	ACC	F1
Baseline	0.57	-
Random Forest	0.78 (0.07)	0.77 (0.07)
XGBoost	0.79 (0.00)	0.78 (0.00)
LightGBM	0.75 (0.00)	0.75 (0.00)
EBM	0.63 (0.00)	0.69 (0.00)
NN	0.60 (0.00)	0.55 (0.37)

C. Counterfactual Explanations (CFs) for Stress Reduction

We generated CFs using the random forest model because XGBoost has not worked on DiCE yet. Table VI shows the explanation of the original model for one participant in the PSS 'Increase' group and CFs to change the participant from 'Increase' to 'Not-increase' group. The features for the original outcome 'Increase' showed that this participant slept for 393 mins (sleep duration), stayed for 445 mins in bed (time in bed), went to sleep at 00:28 am (sleep onset), felt 61.5 pt of quality sleep (sleep quality: 0 - 100 scale), and walked 3697 steps per day (steps) on average.

Based on the results of CFs, we would be able to provide potential suggestions about 'what users should do next [8]. For example, if you increase sleep efficiency by reducing the time staying in bed while maintaining sleep duration and walking more, then your stress levels might be lower or might not be higher (Table VI, Counterfactual 1). If you sleep longer for about 7 hours starting at 1:30 am and walk more for one week, then your stress levels might be lower or not be higher (Table VI, Counterfactual 2).

IV. DISCUSSION

Our results suggested that XAIs for predicting changes in perceived stress. We also showed some examples of 'what-if' explanations about sleep habits to help lower stress. Generating these 'what-if' explanations suggested the possibility of

TABLE V
TOP FIVE CONTRIBUTION FEATURES OF EACH MODEL

Models	Feature Importance (Rank)				
	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}
Random Forest	Sleep onset_std	Sleep quality_std	Steps	Positive affect	Steps_mean
XGBoost	Heat rate_min	Steps_mean	Negative affect	Steps_std	Time in bed
LightGBM	Sleep offset_mean	Sleep quality_mean	Sleep onset_std	Heart rate_mean	Negative affect
EBM	Negative affect	Sleep onset_std	Heart rate_mean	Steps_std	Sleep quality_mean
NN	Sleep onset_std	Heart rate_max	Negative affect	Sleep quality_mean	Positive affect

TABLE VI
EXAMPLES OF DIVERSE COUNTERFACTUAL SET.

	Original	CF1	CF2
SRI	84.4	84.4 (100%)	84.4 (100%)
Sleep duration [mins]	393	393 (100%)	424 (114%)
Time in bed [mins]	445	412 (86%)	445 (100%)
Sleep onset	00:28am	00:28am (100%)	1:30am (115%)
Sleep quality	61.5	61.5 (100%)	61.5 (100%)
Steps	3697	12504 (521%)	16885 (700%)

Original indicates 'Original outcome: Increase PSS
CF1 indicates 'New outcome: Not-increase (Counterfactual 1)'
CF2 indicates 'New outcome: Not-increase (Counterfactual 2)'
Numbers inside the parentheses indicate 'change rate'

a technology that can provide actionable information to help people change their health outcomes.

Regarding limitations and future work in this paper, first, we generated 'what-if' explanations just only for one participant as preliminary results. We need to generate more explanations and then analyze them to provide general and better suggestions. Second, we set ranges of features selected as 'to vary' based on the original outcome scores. Setting ranges based on the opinions of experts (e.g., medical doctors, clinical psychologists, etc.) will be more helpful and meaningful for users. Third, we need to test how real users use this kind of model and CFs and whether users can trust them and make confident decisions based on the explanations in user studies. We believe that providing 'what-if' explanations in healthcare would be very useful and these findings could contribute to developing better technologies.

ETHICAL IMPACT STATEMENT

Our study ultimately aims to contribute to helping people who have a risk of high stress manage their stress by providing actionable data-driven insights. The stress change model and counterfactual explanations presented in this paper are still preliminary and require future careful evaluations including the assessments of model output errors (when models make errors and why and whether models work equally for broader populations including different demographic groups) as well as safety and effectiveness in using stress change classification models and suggested counterfactual explanations for helping manage stress.

REFERENCES

[1] A. Association *et al.*, "2018 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 14, no. 3, pp. 367–429, 2018.

[2] S. Gedam and S. Paul, "A review on mental stress detection using wearable sensors and machine learning techniques," *IEEE Access*, vol. 9, pp. 84 045–84 066, 2021.

[3] E. Howe, J. Suh, M. Bin Morshed, D. McDuff, K. Rowan, J. Hernandez, M. I. Abdin, G. Ramos, T. Tran, and M. P. Czerwinski, "Design of digital workplace stress-reduction intervention systems: Effects of intervention type and timing," in *Proc. of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–16.

[4] K. Hovsepian, M. Al'Absi, E. Ertin, T. Kamarck, M. Nakajima, and S. Kumar, "cstress: towards a gold standard for continuous stress assessment in the mobile environment," in *Proc. of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, 2015, pp. 493–504.

[5] F. Giuste, W. Shi, Y. Zhu, T. Naren, M. Isgut, Y. Sha, L. Tong, M. Gupte, and M. D. Wang, "Explainable artificial intelligence methods in combating pandemics: A systematic review," *IEEE Reviews in Biomedical Engineering*, 2022.

[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[7] 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 international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[8] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proc. of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 607–617.

[9] D. V. Sheehan, Y. Lecrubier, K. H. Sheehan, P. Amorim, J. Janavs, E. Weiller *et al.*, "The mini-international neuropsychiatric interview (mini): the development and validation of a structured diagnostic psychiatric interview for dsm-iv and icd-10," *J. of clinical psychiatry*, vol. 59, no. 20, pp. 22–33, 1998.

[10] S. Cohen, T. Kamarck, and R. Mermelstein, "A global measure of perceived stress," *J. of health and social behavior*, pp. 385–396, 1983.

[11] A. Mackinnon, A. F. Jorm, H. Christensen, A. E. Korten, P. A. Jacomb, and B. Rodgers, "A short form of the positive and negative affect schedule: Evaluation of factorial validity and invariance across demographic variables in a community sample," *Personality and Individual differences*, vol. 27, no. 3, pp. 405–416, 1999.

[12] A. J. Phillips, W. M. Clerx, C. S. O'Brien, A. Sano, L. K. Barger, R. W. Picard, S. W. Lockley, E. B. Klerman, and C. A. Czeisler, "Irregular sleep/wake patterns are associated with poorer academic performance and delayed circadian and sleep/wake timing," *Scientific reports*, vol. 7, no. 1, p. 3216, 2017.

[13] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[14] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[15] H. Nori, S. Jenkins, P. Koch, and R. Caruana, "Interpretml: A unified framework for machine learning interpretability," *arXiv preprint arXiv:1909.09223*, 2019.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *J. of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[17] E. A. P. State of New Hampshire, "Perceived stress scale," accessed: 27 July 2023.