What Does an AI Model Learn About Stress? A SHAP-Based Behavioural Interpretation of a Wearable Stress Classifier

Sarah Aziz[a], Nada Aboubakr[ab], Arfan Ahmed[a]

[a] AI Center for Precision Health, Weill Cornell Medicine-Qatar, Doha, Qatar

[b] Division of Information and Computing Technology, College of Science and Engineering, Hamad Bin Khalifa University, Qatar Foundation, Doha, Qatar

Corresponding author: Sarah Aziz (saa4038@qatar-med.cornell.edu), ORCID: 0000-0002-0861-9743

Abstract. Wearable devices allow passive monitoring of stress-related behaviour in university students, but classifiers built on these data are not widely used in clinical practice because their decisions are not interpretable. In this paper we apply SHAP (SHapley Additive exPlanations) with TreeExplainer to an XGBoost stress classifier trained on Fitbit data and BFI-10 personality scores from 86 university students in Qatar and interpret its predictions on a held-out test set (n=61). The model reaches AUC-ROC 0.640 (95% CI 0.48-0.78), accuracy 0.672 (0.56-0.79) and F1 0.593 (0.41-0.74) on 38 features. The two features with the highest mean absolute SHAP value are exam-period context (0.156) and morning step count five days before assessment (0.154). Conscientiousness contributes to an independent personality signal. The interpretation indicates that the classifier relies on a small set of behavioural signals detectable five days before self-reported stress, suggesting a potential early-alert window for university counselling systems.

Keywords: Explainable AI; SHAP; Wearable sensing; Stress detection; XGBoost; Digital mental health.

1 Introduction

Academic stress is a common problem in university students and is associated with anxiety, sleep loss and lower academic performance, especially during examination periods [1]. Wearable devices such as Fitbit can passively collect daily activity, sleep and heart-rate signals at minute-level resolution and at very low cost to the participant [2]. A growing body of work has shown that machine-learning models trained on these signals can identify stressed students above chance level, but they are rarely used in real clinical or counselling settings. The main reason is that the predictions of a wearable stress model are not, on their own, interpretable: a clinician cannot see which features the model is using to label a student as stressed and therefore cannot easily trust the prediction or act on it.

Existing wearable stress prediction approaches rely on either physiological signals (e.g., electrodermal activity and heart-rate variability) or behavioural data (e.g., physical activity, sleep, and heart rate) from consumer wearables. Recent reviews report limitations including small, homogeneous cohorts, inconsistent stress-labelling protocols, limited external validation, and a lack of model interpretability [3, 4]. These gaps highlight the need for explainable AI approaches capable of supporting trustworthy stress monitoring in university populations.

Personality influences how individuals respond to stress and complements behavioural information captured by wearables. Exposure–reactivity models suggest that traits such as Conscientiousness affect stress responses, motivating the inclusion of personality in wearable-based stress prediction [5].

This problem is often called the interpretability gap, and it has become one of the central concerns of medical AI. To address it, post-hoc explanation methods have become popular in healthcare research. SHAP (SHapley Additive exPlanations), introduced by Lundberg and Lee[6], is one of the most widely used approaches because it assigns each feature a clear, additive contribution for every individual prediction. SHAP has been applied to a range of clinical tasks, including the prevention of intraoperative hypoxemia [7], but its use in wearable-based mental health monitoring is still limited.

In this paper we apply SHAP to an XGBoost stress classifier, as in our main companion paper [8], that was trained on Fitbit data and BFI-10 personality scores from 86 university students, and we interpret its predictions on a held-out test set (n=61) . The contribution of this paper is not a new model or a higher accuracy score, but the systematic interpretation of an existing classifier at both the global and the individual level. Our primary research question is which behavioural, demographic, and personality features drive stress prediction and whether these predictors are clinically and psychologically plausible. We hypothesise that (H1) Behavioural features contribute more to stress prediction than demographic features, (H2) Behavioural features derived from the days preceding stress assessment contribute meaningfully to prediction, supporting an early-alert window, and (H3) Conscientiousness contributes complementary predictive information beyond wearable-derived behavioural features.[5]. Accordingly, our aims are to (i) identify which features the model relies on most to predict stress, (ii) test whether these features are clinically plausible, (iii) examine whether the model's reasoning at the population level also holds for individual participants, and (iv) characterise the interaction between the strongest behavioural feature and the strongest personality feature in the model. The rest of the paper is organised as follows. Section 2 describes the dataset, the classifier and the SHAP analysis pipeline. Section 3 presents the importance of global features, two case-level force plots and a feature dependence plot. Section 4 discusses the clinical implications and limitations, and Section 5 concludes.

2 Methodology

2.1 Dataset

The dataset, full study design and data-collection protocol are described in a related manuscript [8]. The dataset comprised 86 university students from two institutions (Qatar University, n = 46; Weill Cornell Medicine–Qatar, n = 40), yielding 194 DASS-21 assessments across pre-examination, examination, and post-examination periods. Participants had a mean age of 21.0 years, and 44.2% were female. Using DASS-21 severity cut-offs, 79 assessments (40.7%) were classified as above the Normal category and were subsequently grouped as “stressed” for binary classification. Data were split at the participant level into 133 training samples (60 participants) and 61 test samples (26 participants), ensuring no participant appeared in both partitions.

Participants were recruited by convenience sampling from the two participating universities. Eligibility criteria included age ≥18 years, undergraduate enrolment, ownership of a Fitbit-compatible smartphone, and English proficiency. Exclusion criteria comprised conditions preventing wearable use (e.g., skin allergies or open wounds), current psychiatric treatment or diagnosed mental health disorders, and pregnancy. These criteria were selected to ensure participants could safely wear the Fitbit device and complete the study procedures while minimising potential confounding from conditions known to influence stress or wearable-derived physiological measures.The study followed a longitudinal prospective observational design. No a priori power analysis or protocol preregistration was performed; consequently, the findings should be interpreted as exploratory.

Across the 270 DASS-21 assessments collected in the parent study, stress severity was distributed as Normal (57.0%), Mild (16.7%), Moderate (15.2%), Severe (8.9%), and Extremely Severe (2.2%). Within the final analytic subset (194 assessments), 79 (40.7%) were above the Normal threshold. For binary classification, the four non-Normal categories were combined into a single "stressed" class. Although this groups participants with varying symptom severity, Mild and Moderate cases constituted the majority of stressed observations, whereas Severe and Extremely Severe cases were uncommon. Consequently, the classifier and corresponding SHAP explanations should be interpreted as distinguishing normal from elevated stress rather than estimating the severity of stress symptoms.

The DASS-21 has established internal consistency and convergent validity in its original English-language version [9]and subsequent validation in university and Arabic-speaking population. The BFI-10 has shown acceptable test–retest reliability and convergent validity with longer Big Five inventories despite its brevity.

2.2 Feature engineering and selection

For each DASS-21 assessment, we extracted a 7-day window of daily Fitbit features starting on the day immediately before the assessment day (Day 1) and going back to Day 7. For each day we computed 24 features: activity aggregates (steps, distance, calories, sedentary minutes, active minutes), time-of-day (circadian) measures of steps and active minutes for night, morning, afternoon and evening periods, and sleep measures (sleep efficiency, minutes asleep, minutes awake, time in bed). The 7-day window produces 168 wearable features. Personality (five BFI-10 traits) and demographic (age, gender, university, exam period) features were calculated in start of study.

Feature selection was applied only to the 168 wearable features and performed exclusively on the training set using a three-method consensus approach based on F-score, mutual information and permutation importance. The top 40 from each method wereretained, with features selected by at least two methods included. This reduced the wearable feature space to 29, resulting in a final 38-feature set (29 Fitbit + 5 personality + 4 demographic). The 7-day window was selected to align with the DASS-21 recall period, which assesses symptoms experienced over the previous week, ensuring temporal consistency between wearable-derived features and self-reported stress outcomes. Personality and demographic features were collected at baseline as stable individual characteristics rather than dynamic state variables. The DASS-21 and BFI-10 have demonstrated established validity and reliability for assessing psychological symptoms and personality traits, respectively.

2.3 Classifier

An XGBoost gradient boosting classifier was trained on the 38 features. Hyperparameters were chosen to limit overfitting on the small training set (n_estimators=300, max_depth=3, learning_rate=0.02, subsample=0.6, colsample_bytree=0.5, min_child_weight=10, gamma=0.5, reg_alpha=1.0, reg_lambda=2.0). The Normal/Stressed class imbalance (59.3% vs 40.7%) was handled with balanced class weighting (scale_pos_weight = 1.42). Early stopping with a patience of 15 rounds was applied on test log loss. All features were standardised to zero mean and unit variance using parameters fit on the training set only.

2.4 SHAP analysis

SHAP values were computed on the 61 test samples using the TreeExplainer algorithm of Lundberg and Lee [6], implemented in the shap Python library [10]. A SHAP waterfall plot was used to visualise the additive contribution of each feature to an individual prediction, showing how features push the model output toward either the *Stressed* or *Normal* class. Global feature importance was computed as the mean of the absolute SHAP values across the test set. To enable qualitative comparison of model behaviour, we generated two side-by-side waterfall plots: one for a correctly classified high-stress participant and one for a correctly classified low-stress participant. This allows direct comparison of how feature contributions differ across contrasting prediction outcomes. A SHAP dependence plot was additionally used to examine the relationship between the morning step count five days before assessment and its corresponding SHAP value, with points coloured by Conscientiousness.

3 Results

3.1 Model performance

On the held-out test set of 61 participants (24 Stressed and 37 Normal), the XGBoost classifier reaches AUC-ROC 0.640, accuracy 0.672, F1 0.593, sensitivity 0.620 and specificity 0.706. Table 1 reports these values with 95% bootstrap confidence intervals computed from 2,000 resamples. The confidence intervals were wide (AUC-ROC: 0.48–0.78), reflecting the limited sample size of the test set. Importantly, the AUC-ROC confidence interval includes 0.5, indicating that the classifier's discriminative ability cannot be confirmed as better than chance in this dataset. Therefore, these results should be interpreted as a feasibility analysis rather than a definitive assessment of predictive performance. The subsequent SHAP analyses describe the features influencing this fitted model and should be interpreted as model-dependent explanations rather than validated stress biomarkers.

Table 1: Test set performance with 95% bootstrap confidence intervals (2000 resamples), n (no. of participants) =61

Metric	Value (95% CI)
Accuracy	0.672 (0.56-0.79)
AUC-ROC	0.640 (0.48-0.78)
Average Precision	0.606 (0.40-0.77)
F1-Score	0.593 (0.41-0.74)
Sensitivity	0.620 (0.42-0.81)
Specificity	0.706 (0.56-0.85)

3.2 Global feature importance

Figure 1 shows the SHAP beeswarm plot for the top 15 features ranked by mean absolute SHAP value. Two features clearly stand out, with a large gap to the next group: exam-period context (mean |SHAP| = 0.156) and morning step count five days before the assessment (0.154). The next three features are early-week caloric expenditure (0.067), total day-1 calories (0.047) and Conscientiousness (0.047). Across the full feature set, Fitbit-derived features account for 64.7% of the total SHAP magnitude, demographic features account for 24.7% and personality features account for 10.6% (Figure. 4). Conscientiousness is the only personality trait in the top five features.

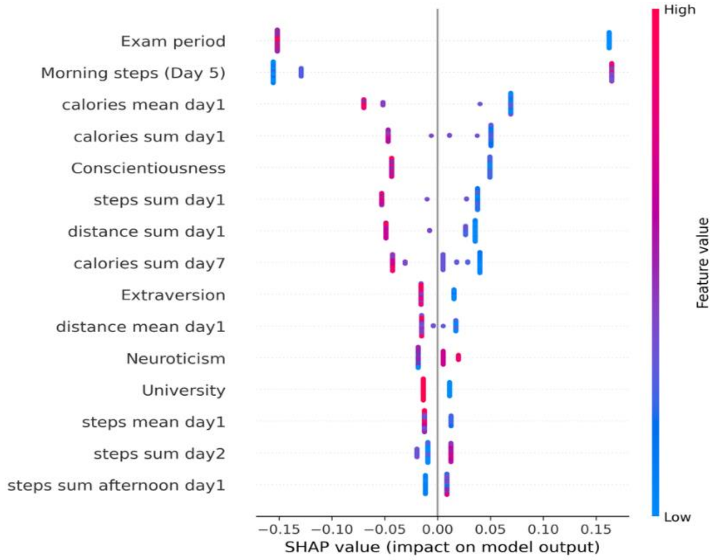
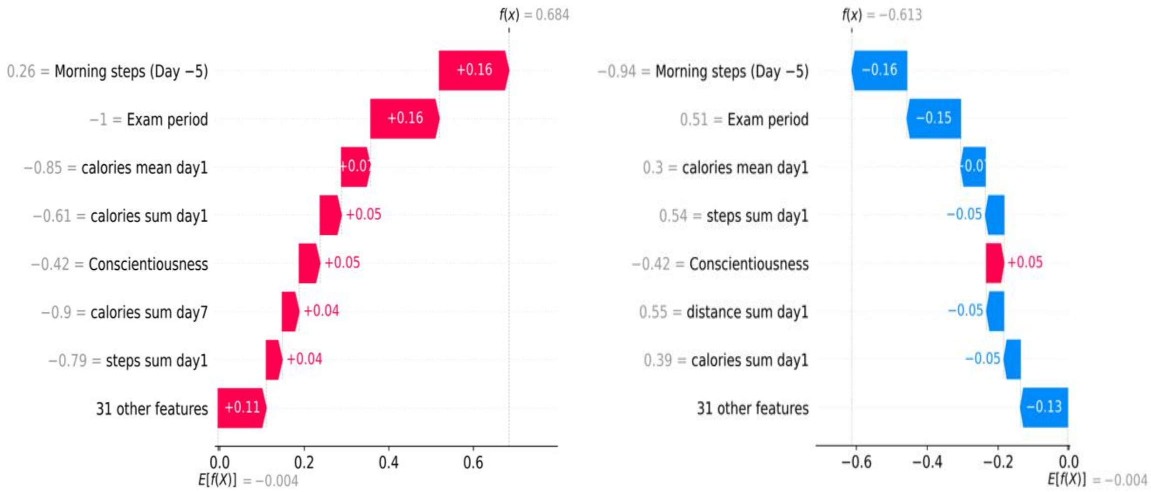


Figure 1:SHAP beeswarm plot of the 15 most influential features for stress classification on the held-out test set (n=61). Each dot is one participant. Horizontal position is the feature's SHAP value (impact on the log-odds of a "Stressed" prediction, positive values push toward Stressed). Colour encodes the participant's standardised feature value, blue (low) to red (high). Features are ordered top to bottom by mean absolute SHAP value.

3.3 Individual-level explanations

To examine whether the global feature ranking is consistent with individual-level model behaviour, we generated SHAP waterfall plots for two contrasting test cases. Figure 2 [A] present a correctly classified high-stress participant (test index 36, sum of SHAP values = +0.69), where the strongest contributors toward the *Stressed* prediction are the Day 5 morning step count and exam-period context, consistent with the global feature ranking. Figure 2 [B] shows a correctly classified low-stress participant (test index 60, sum of SHAP values = −0.61), for whom the same features contribute in the opposite direction, supporting a prediction of the *Normal* class, along with higher early-week caloric expenditure. This indicates alignment between global feature importance and local explanations, supporting the interpretability of the model across both population and individual levels.



[A] **[B]**

Figure 2:SHAP waterfall plots for selected participants showing individual-level feature contributions to model predictions across different stress classifications (stressed [A] vs normal [B]).

3.4 Feature interaction

Figure 3 is a SHAP dependence plot for the morning step count five days before assessment, with each dot coloured by the participant's Conscientiousness score. The dependence plot reveals a threshold-like rule: participants with below-average morning step counts ($z < 0$) consistently receive negative SHAP contributions (pushed toward Normal), while a small subset with unusually high step counts receives positive contributions (pushed toward Stressed), resulting in a Pearson correlation of $r = +0.74$ (Spearman = 0.86) between feature value and SHAP value. This pattern indicates that the model has learned a near-threshold rule around the morning activity level on Day 5, and that Conscientiousness modulates the size of the effect close to the threshold rather than reversing its direction.

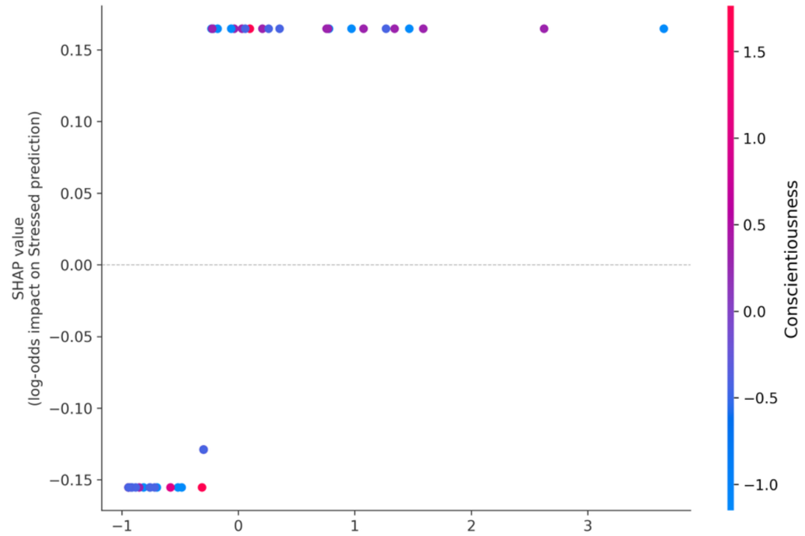


Figure 3:SHAP dependence plot for morning step count five days before assessment (test set, n=61). x-axis: standardised feature value; y-axis: the feature's SHAP value. Each dot is coloured by Conscientiousness (blue = low, red = high). The plot shows a threshold-like rule: most participants with below-average morning steps are pushed toward Normal, while those with unusually high step counts are pushed toward Stressed.

To assess the robustness of feature importance rankings, the model was retrained on 10 bootstrap resamples of the training data and SHAP feature rankings were recomputed for each iteration. Spearman rank correlations between the bootstrap-derived rankings and the original ranking yielded a mean correlation of 0.524 (SD = 0.139; range = 0.240–0.684), indicating moderate stability of the feature ordering. While the highest-ranked features, particularly examination period and morning step count, remained consistently among the most influential predictors, greater variability was observed among lower-ranked features, reflecting sensitivity to sampling variation and the modest sample size.

3.5 Feature-group contribution

To complement the individual-feature ranking, we aggregated the mean absolute SHAP values by feature domain (Fig. 4). Fitbit-derived behavioural features account for 64.7% of the total SHAP magnitude on the test set, demographic features (including exam-period context) account for 24.7% and BFI-10 personality features for 10.6%.

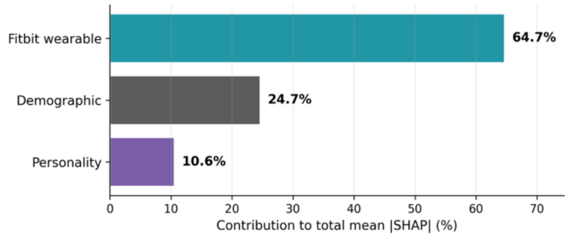


Figure 4: Percentage contribution of each feature group to the total mean absolute SHAP value of the model on the held-out test set (n=61). Fitbit-derived behavioural features account for the largest share, while personality contributes a smaller but independent

4 Discussion

The most informative behavioural feature in our analysis was the morning step count five days before the DASS-21 assessment. Passive behavioural signals from the Fitbit accounted for nearly two-thirds of the model's total explanatory power (64.7%), with demographic context and personality contributing secondary but meaningful roles; the modest personality contribution (10.6%) may partly reflect the brevity of the BFI-10 instrument rather than a true ceiling on personality's predictive value. Critically, this signal is detectable before student's self-report elevated stress; suggesting that behavioural changes captured passively by wearables may precede subjective stress awareness and provide an early-alert window for university counselling. The SHAP dependence plot (Figure 3) shows that the classifier separates the test participants into two SHAP regimes based on this single feature, indicating that the model has learned a near-threshold behavioural rule on Day 5 rather than a smooth dose-response relationship. A clinically intelligible interpretation of this finding is a disruption to habitual morning activity, whether suppression or elevation , during the exam-week window can serve as an early, objective marker of stress, in line with prior work on the bidirectional relationship between psychological stress and physical activity routines [11]. Because the signal is detectable several days before the student self-reports higher stress, the temporal lead is the most clinically actionable aspect of the result.

Conscientiousness was the only personality trait among the top five SHAP features and contributed an independent signal on top of the wearable behavioural channel. The directional analysis indicates that, in this exam-period cohort, the model linked higher Conscientiousness with a higher predicted probability of stress. At first this appears to contradict the general view of Conscientiousness as a stress-buffering trait, but it is consistent with recent evidence that the protective effect of Conscientiousness depends on the context of the stressor and is reduced or reversed in high-stakes academic settings [12]. In time-limited evaluations such as final examinations, conscientious students tend to set high standards and to closely monitor their own performance, which can lead to anticipatory stress rather than reduced stress. Notably, Neuroticism, typically the strongest personality predictor of stress in trait-based models, did not appear among the top five features, which may indicate that the Fitbit behavioural signal already captures much of the variance that Neuroticism would otherwise explain.

Beyond the specific marker identified, the contribution of this work is not a new explanation method, as SHAP with TreeExplainer for XGBoost is an established approach. Rather, we demonstrate its application to a previously unexplored combination of behavioural, demographic, and personality features in a university stress-monitoring context, with an emphasis on assessing whether model-derived explanations align with plausible stress-related patterns. Once validated in larger cohorts, such models may support identification of behavioural patterns associated with stress and inform targeted monitoring approaches. The observed contribution of Conscientiousness is consistent with personality-based stress models suggesting that traits influence stress exposure and responses [5]; however, the cross-sectional, single-cohort design prevents conclusions regarding underlying mechanisms.

This work has several limitations. First, SHAP values explain model behaviour but do not establish causal relationships between behavioural patterns and stress. Second, the test set was relatively small (n = 61, 26 participants), resulting in wide bootstrap confidence intervals for some performance metrics. SHAP values themselves carry no uncertainty quantification here, and with only 61 test samples the beeswarm distributions and dependence patterns could shift materially under resampling; their stability should be treated as an open question for replication in a larger sample rather than a settled result. Third, the cohort was drawn from a single geographical region, and personality traits were measured only once at enrolment and assumed to remain stable over time. Fourth, Fitbit-derived activity measures cannot distinguish intentional from unintentional physical activity, limiting physiological interpretation and motivating future integration of richer physiological signals such as heart-rate variability. Finally, the classification threshold was fixed at 0.5 and not optimised for deployment; future implementations should adjust the threshold according to the desired balance between sensitivity and specificity.

Prospects of Application: This SHAP-based wearable stress classifier could be deployed as a passive early-alert tool within university counselling systems, flagging students whose morning activity drops five days before assessments. The model's interpretable outputs would allow counsellors to prioritise proactive outreach, reducing escalation to clinical stress levels in high-pressure academic contexts.

5 Conclusion

This paper presented a SHAP-based interpretation of an XGBoost stress classifier trained on Fitbit data, BFI-10 personality scores and demographic information from 86 university students in Qatar and identified morning step count five days before assessment as the most informative behavioural feature alongside exam-period context, with Conscientiousness as the only personality trait among the top five predictors.

Individual-level waterfall plots and a feature dependence plot together showed that the model's reasoning at the population level also holds for individual participants. Future work will validate these findings in larger, multi-site cohorts and will examine the added value of physiological signals such as heart-rate variability and electrodermal activity.

Disclosure of Interests. The authors have no competing interests to declare.

Acknowledgements: ChatGPT (GPT-5) was used under human supervision for proofreading and language editing during manuscript preparation. The authors retain full responsibility for all content, analyses, interpretations, and final decisions. AI tools were not involved in generating scientific findings and are not authors.

6 References

1. Wang, Y., et al., *Stressors in university life and anxiety symptoms among international students: a sequential mediation model.* BMC psychiatry, 2023. **23**(1): p. 556.
2. Smets, E., et al., *Large-scale wearable data reveal digital phenotypes for daily-life stress detection.* NPJ digital medicine, 2018. **1**(1): p. 67.
3. Abd-Alrazaq, A., et al., *The performance of wearable AI in detecting stress among students: systematic review and meta-analysis.* Journal of Medical Internet Research, 2024. **26**: p. e52622.
4. Vos, G., et al., *Generalizable machine learning for stress monitoring from wearable devices: A systematic literature review.* International journal of medical informatics, 2023. **173**: p. 105026.
5. Bolger, N. and A. Zuckerman, *A framework for studying personality in the stress process.* Journal of personality and social psychology, 1995. **69**(5): p. 890.
6. Lundberg, S.M. and S.-I. Lee, *A unified approach to interpreting model predictions.* Advances in neural information processing systems, 2017. **30**.
7. Lundberg, S.M., et al., *Explainable machine-learning predictions for the prevention of hypoxaemia during surgery.* Nature biomedical engineering, 2018. **2**(10): p. 749-760.
8. Aziz, S.A.-A., Alaa; Yoosuf, Shehel; Qidwai, Uvais; Sheikh, Javaid; Househ, Mowafa; Ahmed, Arfan, *Personality-Augmented Stress Detection from Wearable Data: A Machine Learning Approach Using Circadian Features and Big Five Traits.* Manuscript under review, 2026.
9. Lovibond, S.H., *Manual for the depression anxiety stress scales.* Sydney psychology foundation, 1995.
10. Lundberg, S., *shap: A unified approach to explain the output of any machine learning model.* GitHub. URL: https://github. com/slundberg/shap [accessed 2022-02-15], 2018.
11. Stults-Kolehmainen, M.A. and R. Sinha, *The effects of stress on physical activity and exercise.* Sports medicine, 2014. **44**(1): p. 81-121.
12. Ringwald, W.R., et al., *Characterizing stress processes by linking big five personality states, traits, and day-to-day stressors.* Journal of Research in Personality, 2024. **110**: p. 104487.