Supplementary Table S6. Summary of selected studies utilizing **daily life stressors**.

Authors	Sample size	Physiological signals	Psychological measures	Condition	Stressor(s)	Model and Performance	Ground truth/Labels	Dataset used	Findings
Al-Alim et al., 2024 [60]	238 out of 1002	ECG, TEMP, EDA, ACC	Customized questionnaires	W	**Daily life stressor: five days monitoring over two years**	5-fold CV accuracy: kNN: 95.70% - 96.37% (binary); 97.73% - 97.87% (three-class) SVM: 74.97% - 96.22% (binary); 91.30% - 96.48% (three-class) DT: 93.01% - 95.88% (binary); 76.50% - 95.57% (three-class) **RF: 93.28% - 98.29% (binary); 97.98% - 98.67% (three-class)** XGBoost: 91.44% - 98.29% (binary); 84.83% - 98.98% (three-class) 5-fold CV F1-score: kNN: 95.5% - 95.7% (binary); 97.64% - 97.73% (three-class) SVM: 74.94% - 95.33% (binary); 91.27% - 95.76% (three-class) DT: 91.48% - 95.88% (binary); 59.44% - 95.55% (three-class) **RF: 93.27% - 97.89% (binary); 97.22% - 98.67% (three-class)** XGBoost: 91.43% - 97.89% (binary); 76.59% - 98.98% (three-class)	Self-report scores	SWEET	SMOTE technic was used to address the imbalanced samples in the SWEET dataset. RF performed the best in both binary and three-class (no vs. low vs. high stress) classification with an accuracy of 98.29% on imbalanced dataset. Using SMOTE technique to address the imbalanced samples, kNN achieved 95.70% in binary classification and XGBoost has the best accuracy of 98.98% in three-class classification.

Başaran et al., 2023 [151]	14 (5 females, ages between 20 and 25 years, university students)	BVP, EDA, TEMP, ACC	PSS	W	**Daily life stressor: (1-week monitoring)**	E2E Accuracy: LSTM: 72.73% - 90.38% (supervised) **CNN-LSTM: 74.89% - 91.35% (supervised)** MLP: 68% - 76% (semi-supervised) RF: 68% - 77% (semi-supervised) LR: 61% - 63% (semi-supervised) k-means: 73% (unsupervised) BIRCH: 70% (unsupervised) DBSCAN: 56% (unsupervised) F1-score: LSTM: 72.75% - 82.6% (supervised) **CNN-LSTM: 72.52% - 76.32% (supervised)**	PSS scores		The study proposed a semi-supervised learning to address the challenge in intensive labeling in massive data collected in the wild condition for stress classification. The semi-supervised MLP achieved an accuracy of 78%, comparable to supervised models and better than unsupervised clustering models. Imbalance data was observed as such that 73% of collected data was reported as non-stress.
Gjoreski et al., 2016 [115]	21 (in-lab: mean age of 28 years); 5 (wild, mean age of 28 years)	PPG, EDA, Temp, and ACC	In-lab: STAI Wild: Ecological Momentary Assessment	L and W	Mental arithmetic task; **Daily life stressor: 55 days of passive monitoring**	LOSO accuracy: RF: 72% (three-class, in-lab), 83% (binary, in-lab), 76% - 92% (binary, wild) LOSO F1-score: RF: 60% - 85% (three-class, in-lab), 80% - 85% (binary, in-lab), 39% - 96% (binary, wild)	Study protocol design (in-lab); EMA scores (wild)		Adding context information and activity recognition models help improve the accuracy of stress detection.
Gjoreski et al., 2017, [55]	21 (in-lab: mean age of 28 years); 5 (wild, mean age of 28 years)	PPG, EDA, Temp, and ACC	In-lab: STAI Wild: Ecological Momentary Assessment	L and W	Mental arithmetic task; **Daily life stressor: 55 days of passive monitoring**	LOSO accuracy: SVM: **67% - 71% (in-lab)**; RF: 60% - 69% (in-lab) Boosting: 59% - 67% (in-lab) Bagging: 58% - 65% (in-lab) kNN: 52% - 65% (in-lab) NB: 59% - 62% (in-lab) DT: 52% - 59% (in-	Study protocol design (in-lab); EMA scores (wild)		SVM showed the best classification results for in-lab data. ML models based on in-lab data were tested on data collected in the wild, and mean F1 scores were 0.47. However, similar to the author's previous work, adding context information and activity recognition models improved the accuracy of stress detection for real-life data, with a highest F-score of 95%

						lab) Ensemble selection: 53% - 55% (in-lab) LOSO F1-score: SVM: 55% - 71% (wild) RF: 66% - 74% (wild) Boosting: 48% - 54% (wild) Bagging: 59% - 78% (wild) kNN: 63% - 71% (wild) NB: 57% - 60% (wild) DT: **68% - 90% (wild)** Ensemble selection: 62% - 80% (wild)			
Han et al., 2020 [77]	17 (in-lab data: 4 females, ages between 20 to 27 years); 3 (wild: 1 female)	ECG, PPG, and EDA	Customized self-report questionnaire	L and W	Memory game, mosquito sound, images test using International Affective Picture System, plank exercise, CPT, TSST, SCWT; **Daily life stressor: two days of passive monitoring**	10-fold CV accuracy: **kNN 1 neighbor: 66.86% - 94.55% (in-lab), 63.645 - 100% (wild)** kNN 3 neighbors: 68.32% - 92.37% (in-lab), 54.55% - 100% (wild) kNN 5 neighbors: 71.33% - 89.37% (in-lab), 54.55% - 83.33% (wild) kNN 7 neighbors: 73.71% - 87.19% (in-lab), 63.64% - 66.67%(wild) kNN 9 neighbors: 75.28% - 84.47% (in-lab), 64.64% - 72.73% (wild), SVM: 79.84% - 84.74% (in-lab), 36.36% - 83.33% (wild) NB: 70.54% - 83.1% (in-lab), 45.45% - 72.73% (wild)	Study protocol design (in-lab); Self-report scores (wild)		kNN with 1 neighbor showed the highest accuracy for in-lab data and even reached 100% accuracy when physical activities were excluded from wild data due to motion artifact. The generalized model has the highest accuracy of 94.55% using 10-fold CV, and an accuracy of 85.71% when using leave-one-participant CV (generalize to individuals)

Pasha et al., 2024 [54]	15 (all females, ages between 30 - 55 years, nurses)	EDA, TEMP, PPG	Customized self-report questionnaire	W	**Daily life stressor: 1-week monitoring**	Accuray: Bidirectional LSTM: 96% **Ensemble: 97%** **F1-score:** Bidirectional LSTM: 91% - 98% **Ensemble: 93 - 99%**	Self-report scores	Multimodal sensor dataset of nurse	Ensemble classifiers (DT, RF, XGBoost, MLP, LR) performed slightly better than bidirectional LSTM for three-class (no vs. low vs. high) classification, though imbalance in the dataset impacted the performance.
Plarre et al., 2011 [96]	21 (in-lab data: college students, average 20.6 years old, half were females and 96% of them were Caucasians); 17 out of 21 (wild condition data)	ECG, TEMP, ACC, EDA, and RESP	Customized self-report questionnaire	L and W	Public speaking, mental arithmetic, CPT**; Daily life stressor: two days of passive monitoring**	10-fold CV accuracy: J48 Decision tree: 82.33% - 87.67% **J48 AdaBoost: 88% - 90.17%** SVM: 88.17% - 89.17%	Study protocol design (stressors vs. rest conditions) and self-report scores		The physiological model uses ECG and respiration features collected from the in-lab session, and AdaBoost showed the highest accuracy for binary classification. When trained on identifying individual stressors, the performance ranges between 82% to 99%. Normalization using baseline conditions also marginally improved the classification accuracy. The perceived stress model predicts the perception of stress level, with J48 decision tree having an accuracy of 84%. Only 33.55% of wild data were preserved due to the presence of noise and missing data, and both models showed close estimation of stress compared to the ground truth. Namely, the physiological model estimate that participants felt stressful 35.14% of time, and 26.62% predicted by the perceived stress model, with the ground truth from the self-report to be 28.08%
Schmidt et al., 2019 [23]	11 (4 females, mean age 26)	BVP, ACC, EDA, TEMP	STAI, SAM	W	**Daily life stressor: 16 days of passive monitoring**	Leave-target-questionaires-out F1-score: **DT: 31.3% - 37.5%** ET: 33.3% - 36.8% RF: 31.7% - 37.4% Single-task CNN: 35.1% - 39.2% (E2E) **Multi-task CNN: 36.3% - 41.5% (E2E)** **LOSO F1-score:** **DT: 46.7% - 46.8%** ET: 46.5% RF: 46.2%	STAI scores		E2E CNN models perform slightly better than FB models in three-class anxiety detection using data collected in the wild, but the low F1 scores still highlight challenges for anxiety detection in uncontrolled conditions due to data and label quality, imbalanced classes, and awareness of affective states

Toshnazarov et al., 2024 [107]	28 (15 females, mean ages of 22/8 years, 16 university students, 12 from the general population)	BVP, ACC	Ecological momentary assessment	L and W	Public speaking, mental arithmetic, CPT; **Daily life stressor: two weeks of passive monitoring**	LOSO Accuracy: AdabBoost: 76.4% (in-lab); 64.7% - 70.1% (F1, wild) GBoost: 77.6% (in-lab); 65.3% - 70.7% (F1, wild) LR: 81.6% (in-lab); 64.9% - 71% (F1, wild) **MLP: 79.6% (in-lab); 61.8% - 71.2% (F1, wild)** RF: 78% (in-lab); 63.0% - 66% (F1, wild) SVM: 78% (in-lab); 64.7% - 70.9% (F1, wild) XGBoost: 79.2% (in-lab); 63.3% - 67.9% (F1, wild) LOSO F1-score: AdabBoost: 79.4% (in-lab); 64.7% - 70.1% (wild) GBoost: 79.8% (in-lab); 65.3% - 70.7% (wild) LR: 83.5% (in-lab); 64.9% - 71% (wild) **MLP: 82% (in-lab); 61.8% - 71.2% (wild)** RF: 80.2% (in-lab); 63.0% - 66% (wild) SVM: 80.4% (in-lab); 64.7% - 70.9% (wild) XGBoost: 81.3% (in-lab); 63.3% - 67.9% (wild)	Study protocol design (in-lab); EMA scores (wild)		The study developed pipelines for detecting stress in both in-lab and in-the-wild condition using PPG signals. LR obtained the highest accuracy for in-lab dataset. Although the detection accuracy in real-world datasets are generally lower than in-lab dataset, adding the contextual features (e.g., phone logs, activity, locations) provided the highest performance of F1-score using LR. MLP also achieved high F1-scores when using PPG and considering previous stress state information for detecting stress in the wild.
ACC: acceleration; BVP: blood volume pulse; CPT: cold pressor test; CV: cross-validation EDA: electrodermal activity; ECG: electrocardiogram; EMG: electromyography; LDA: linear discriminant analysis; LOSO: leave-one-subject-out; LR: linear regression; LSTM: long short-term memory; RESP: respiration; ROC AUC: receiver operating characteristic area under the curve; RT: random forest; SVM: support vector machine; TEMP: body temperature; TSST: trier social stress test. Daily life stressors are highlighted in bold font									