Supplementary Table S3. Summary of selected studies utilizing **physical stressors**.

Authors	Sample size	Physiological signals	Psychological measures	Condition	Stressor(s)	Model and Performance	Ground truth/ Labels	Dataset Used	Findings
Ahmad et al., 2023 [57]	9 (own dataset, aged between 20 – 40 years); 15 (WESAD, average 27.5 years old, 12 males, graduate students)	ECG	Own dataset: Customized questionnaires; WESAD: PANAS, STAI, SAM, and SSSQ	L	Own dataset: roller coaster simulation WESAD: **TSST**	LOSO Accuracy: Bagging classifier: 53.9% (own dataset), 64.6% (WESAD) kNN: 45.1% - 57.1% (own dataset); 57.2% - 66.2% (WESAD) RF: 46.2% (own dataset); 61.8% (WESAD) XGBoost: 52.1% (own dataset); 65.8% (WESAD) SVM: 46.2% (own dataset); 60.2% (WESAD) 1D CNN: 62.8% (E2E, own dataset); 64.1% - 66.6% (feature fusion, own dataset); 51.4% (E2E, WESAD); 69% - 72.7% (feature fusion, own dataset) **ResNet: 63.4% (E2E, own dataset)**; **72% (E2E, WESAD)** LOSO F1-score: Bagging classifier: 27.1% (own dataset); 51% (WESAD) kNN: 27.4% - 33.9% (own dataset); 46.9% - 50.1% (WESAD) RF: 30.4% (own dataset); 50.8% (WESDA) XGBoost: 32% (own dataset); 49.5% (WESAD) SVM: 23.3% (own dataset); 46.6% (WESAD) 1D CNN: 61.6%	Own dataset: Self-report scores WESAD: study protocol design	WESAD	The study used ECG to detect stress in simulated VR environment and used WESAD as benchmark. CNN showed better performance compared to traditional ML models and feature fusions resulted in higher accuracy (72.7%) in three-class classification using their own dataset and 87.7% in WESAD binary classification. E2E method of using ECG signals on their own dataset also resulted in high accuracy of 66.6% compared to other traditional ML models.

						(E2E, own dataset); 63.3% - 65.6% (feature fusion, own dataset); 48.7% (E2E, WESAD); 69.1% - 73.1% (feature fusion, own dataset) **ResNet: 62.3% (E2E, own dataset)**; **72.5% (E2E, WESAD)**			
AlShorman et al., 2022 [61]	14 (age between 18 - 23 years, 14 males, undergraduate students)	EEG	Depression Anxiety Stress Scales	L	cold pressor test	K-fold CV Accuracy and F1-score: SVM (RBF): 80.5% **SVM (Linear): 87.9% - 90.2%** SVM (Polynomial): 79.9% SVM (Sigmoid): 80.5% NB: 80.5% - 80.1% LOSO CV Accuracy and F1-score: SVM (RBF): 80.5% **SVM (Linear): 90.2%** SVM (Polynomial): 79.9% SVM (Sigmoid): 80.5% NB: 80.1%	Study protocol design		SVM linear reached the highest performance with a 90% accuracy on binary mental stress classification using frontal lobe EEG activity. Note that authors used CPT but claimed to detect mental stress.
Delmastro et al., 2020 [72]	9 (5 females, mean age 83.6 years; 4 males, mean age 74.7 years)	ECG and EDA	Apathy Evaluation Scale, Hamilton Depression Scale, Cohn Perceived Stress	L	Physical exercise using a cycle-ergometer, SCWT	10-fold CV accuracy: RF: 85.4% - 87.0% **AdaBoost: 85.3% - 88.2%** 10-fold CV AUPRC: RF: 94% - 98% AdaBoost: 92% - 97%	Study protocol design		RF and AB showed comparable performance in distinguishing cognitive stress induced by SCWT and rest conditions. The physical activity induces a low stress level and positively impacts the cognitive performance in SCWT
Halder et al., 2024 [152]	35 (aged between 18 – 75 years)	EDA, ACC, TEMP, HR	None	L	SCWT, **TSST**, Hyperventilation provocation test	E2E model Accuracy: LSTM: 91.78% F1-score: LSTM: 88.06% - 93.73%	Study protocol design		LSTM on binary stress classification using the Empatica E4 wristband achieved accuracy of 91.13% using EDA, ACC, TEMP, and HR

Han et al., 2020 [77]	17 (in-lab data: 4 females, ages between 20 - 27 years); 3 (wild: 1 female)	ECG, PPG, and EDA	None	L and W (two day of passive monitoring)	Memory game, mosquito sound, images test using International Affective Picture System, **plank exercise**, **CPT**, TSST, SCWT	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 as ground truth for in-lab data; Self-report scores for wild data		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)
Kim et al., 2024 [84]	26	ECG, RESP, Electrogastrogram	None	L	Image test using International Affective Picture System, **CPT**	5-fold CV accuracy: LR: 64.8% SVM: 66.8% RF: 78.6% Single CNN: 60.8% - 83.2% Multiple CNNs: 88.8% Multiple CNNs with efficient channel attention (ECA): 90.6% **Multiple CNNs with shuffled ECA: 62.6% - 91.6%** 5-fold CV F1-score: LR: 61.7% SVM: 60.6% RF: 77.1% Single CNN: 58.3% 82.4%	Saliva cortisol levels		The study used multiple CNN with a shuffled ECA to detect binary stress with feature fusions from different physiological signals, reaching an accuracy of 91.6%

						Multiple CNNs: 88.5% Multiple CNNs with efficient channel attention (ECA): 90.2% **Multiple CNNs with shuffled ECA: 59% - 91.4%**			
Pinge et al., 2022 [119]	5 (2 females, age between 25 - 35 years)	ECG and PPG	None	L	Public speaking, mental arithmetic, **CPT, walking exercise**	Accuracy: RF: 81% - 85% F1-score: RF: 80% - 85%	Study protocol design		Three hear rate monitoring devices were compared for their performance in measuring physiological changes in the presence of multiple stressors. Slight differences in HR readings were observed across devices, but the accuracy of using signals from each device are comparable.
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 (two day of passive monitoring)	Public speaking, mental arithmetic, **CPT**	10-fold CV accuracy: J48 Decision tree: 82.33% - 87.67% J48 AdaBoost: 88% - 90.17% SVM: 88.17% - 89.17%	Study protocol design		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 improves 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 state 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%
Ribeiro et al., 2024 [169]	16 (4 females, ages between 16 - 91 years)	PPG, EDA	Customized self-report questionnaire	L	Thermal stress (cold and heat stimuli)	Accuracy: Fuzzy logic:89% - 100% Sensitivity: Fuzzy logic: 75% - 94% Specificity: Fuzzy logic: 90% - 100%	Self-report scores		The study used PPG and EDA signals to classify five levels of stressful states using fuzzy logic, which resulted in high accuracy between 90% to 100%.
Sandulescu et al., 2015 [22]	5 (ages between 18 - 39 years, college students, including both females and males) for TSST, 6 participants for thermal stress experiment	HR, TEMP, body movement, speech	STAI	L	TSST, **Thermal stress (**physical activity at 40 Celsius degrees)	Accuracy: SVM (Gaussian kernel): 85.7% for TSST Precision: SVM (Gaussian kernel): 94.47% for TSST Recall: SVM (Gaussian kernel): 84.54% for TSST	Study protocol design		SVM achieved an accuracy of 86% for detecting mental stress. It is unclear of thermal stress detection result though.
Toshnazarov et al., 2024 [107]	28 (15 females, mean ages of 22/8 years, 16 university students, 12 from the general population)	PPG, ACC	EMA	L and W	**Public speaking**, mental arithmetic, CPT; Daily life stressor: two days 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-	Study protocol design (in-lab); PSS scores (wild)		The study developed pipelines for detecting stress in both in-lab and in-the-wild condition using PPG signals. LR achieves 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 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.

						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)			
Zhu et al., 2023 [111]	62 (CLAS, , mostly students between 20 – 27 years old, two aged between 30 – 50 years), 20 (UTD, 14 males), 18 (VerBIO, 9 females), 15 (WESAD: average 27.5 years old, 12 males, graduate students)	EDA only from multiple datasets	PANAS, STAI, SAM, and SSSQ	L	CLAS: Math problems, SCWT, watching videos and pictures UTD: walking, counting, SCWT, watching videos VerBIO: **public speech** WESAD: **TSST**	LOSO Accuracy: kNN: 57.6% - 59.5% (CLAS), 56.9% - 58.6% (UTD); 62% - 69.1% (VerBIO); 66.1% - 69.1% (WESAD); SVM: **66.4% - 68.5% (CLAS)**; 62% - 66.3% (UTD); **75.4% - 92.9% (VerBIO)**; 71.4% - 75.2% (WESAD) NB: 63.2% - 63.5% (CLAS), 58.6% - 64.8% (UTD); 53.8% - 67.9% (VerBIO); 67.8% - 71.2% (WESAD); LR: 63.9% - 66.6% (CLAS), 67.6% - 68% (UTD); 73.2% - 87.7% (VerBIO); 73.1% - 78.1% (WESAD); RF: 61.5% - 61.9% (CLAS); **61.3% - 73.1% (UTD)**; 64% to 71.6%	Study protocol design		Using EDA features alone showed high classification accuracy in multiple datasets with different models. Gender also plays a role in model performance that models trained on females have higher accuracy than males.

						(VerBIO); **68.3% - 86.5% (WESAD)**			
ACC: acceleration; AUPRC: area under the precision-recall curve; 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; NB: Naïve Bayes; RESP: respiration; RT: random forest; SCWT: Stroop color-word test; SVM: support vector machine; TEMP: body temperature; TSST: trier social stress test. Physical stressors are highlighted in bold font.									