Supplementary Table S4. Summary of selected studies utilizing **emotional stressors**.

Authors	Sample size	Physiological signals	Psychological Measures	Condition	Stressor(s)	Model and Performance	Ground truth/Labels	Dataset used	Findings
Appriou et al., 2020 [134]	24 (12 females, 18 to 54 years old) for mental workload dataset, 32 (DEAP, 16 females, mean age of 24.9 years)	EEG	STAI (EEG mental workload dataset) and Self-Assessment Manikin (DEAP)		N-back tasks; **music-video clips viewing for DEAP**	LOSO accuracy: CSP: 52.5% - 58.26% (mental workload); 58.04% - 67.0% (DEAP) FBCSP: 55.31% - 59.13% (mental workload); 60.07% - 68.51% (DEAP) FgMDM: 47.97% - 60.04% (mental workload); 58.31% - 69.94% (DEAP) TSC: 49.14% - 60.04% (mental workload); 58.31% - 68.50% (DEAP) FBFgMDM: 48.43% - 61.01% (mental workload); 61.30% - 70.34% (DEAP) FBTSC: 48.75% - **61.09% (mental workload)**; 60.18% - 68.72% (DEAP) CNN: 40.15% - 48.05% (mental workload); 63.74% - **72.73% (DEAP)**	Study protocol design as ground truth for mental workload dataset; Self-assessment scores for DEAP dataset (5 as cutoff values)	EEG mental workload dataset by Muhl et al; DEAP	CNN demonstrated the best performance in both subject-independent and subject-specific classifications. However, the authors do not recommend using CNN when dataset is relatively small
Ding et al., 2022 [167]	30 (15 females, mean age is 24.37 years, graduate students)	ECG, and EDA	STAI-S, and VAD	L	**Negative visual stimuli from the International Affective Picture System (IAPS)**, and stop-signal task	Correlation coefficient between predicted and actual STAI scores using LASSO: 0.4748 - 0.5528	STAI scores		This study used regression models to predict state anxiety using physiological signals and provided statistical evidence of correlation between physiological and psychological measures
Erkus et al., 2020 [73]	9 (mean age of 27.11 years old)	Pupil diameter	PANAS, Beck's Depression Index, STAI.	L	Watching video clips with positive, negative, and neutral arousing	10-fold CV accuracy: LDA: 44.4% - **67.9% (FB)**; 37.2% (E2E) kNN: 46.9% - 66.7% (FB); 43.2% (E2E) NB: 35.8% - 63% (FB); 37.4% (E2E) SVM: 39.5.4% - 65.4% (FB);	Study protocol design		LDA and Coarse tree reached the highest accuracy for three-class arousal classification as FB classifier using range

						48.8% (E2E) Coarse tree: 37.3% - **67.9% (FB)**; 34.5% (E2E)			features of pupil diameter. Using raw pupil diameter data, SVM also achieved high accuracy. kNN was also considered as an overall best FB classifier when all features were used
Gazi et al., 2021 [115]	55 (ages between 18 to 40 years, arachnophobia individuals)	ECG, EDA, and RESP	Non-standard	L	Spider exposure in virtual environment for arachnophobia individuals	Accuracy: RF: 60% - 88% F1-score: RF: 59% - 88%	Videos clips contain the highest normalized HR means across all subjects were labeled as 'stress', and the rest of the clips were labled as 'rest' for each subject	PhysioNet	Incorporating respiratory variability helps improve the anxiety detection accuracy as complementary information for EDA and ECG features. Combining EDA and RSP features gave the highest classification accuracy (88%)
Giannakakis et al., 2017 [75]	23 (7 females, average ages of 45.1 years old)	Facial cues, camera-based PPG	Customized self-report questionnaire	L	Social exposure (Self-describing task, text reading), **emotional recall (recall anxious/stressful event), Viewing stressful images from IAPS**, SCWT, w**atching stressful videos**	10-fold CV accuracy: kNN: 85.54% - 88.70% Generalized likelihood ratio: 72.39% - 86.69% NB: 71.63% - 86.09% **AdaBoost: 81.03% - 91.68%** SVM: 65.82% - 83.15%	Study protocol design		Facial features and heart rate were used to train ML models for binary classification for four types of stressors, and kNN showed consistent and highest accuracy of stress identification (88.32% - 88.70%) except for Social Exposure, which AdaBoost obtained the best performance instead (91.68%)

Hag et al., 2023 [76]	25 of 32 (DEAP, 16 females, mean age of 24.9 years)	EEG	Self-Assessment Manikin	L	**Music-video clips viewing**	10-fold CV accuracy: Regularized LDA: 79% **SVM: 81% - 85.68%** kNN: 75%	Self-Assessment Manikin scores		The study proposed a new channel selection technique using correlation coefficient of Hjorth parameter to rank channels, and SVM reached an 81.56% accuracy with only 8 channels to classify calm vs. stressful states, comparable to 85.68% classification accuracy when using full channels set.
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	None	L and W (two day of passive monitoring)	Memory game, mosquito sound, **Images test using IAPS**, 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)
Henry et al., 2023 [78]	30 (CASE, 15 females, mean age of 27.15 years) 15 (WESAD,	CASE: ECG and BVP via PPG only WESAD: ECG	CASE: SAM WESAD: PANAS, STAI, SAM, and SSSQ	L	CASE: **watch video clips that contain different**	Accuracy: **SVM: 56.0% to 76.3% (CASE); 58.6% to 81.1% (WESAD)** RF: 58% - 75.2% (CASE);	CASE: customized anxiety index from self-report	CASE, WESAD	SVM performed slightly better than other models in both within and

	average 27.5 years old, 12 males, graduate students)	and BVP via PPG only			**emotions** WESAD: TSST	56.8% - 79.5% (WESAD) XGBoost: 58.2% - 73.5% (CASE); 56.1% - 79.6% MLP: 56.6% - 70.4% (CASE); 60% - 80.1% (WESAD) **ResNet: 58.0% - 70.7% (CASE); 59.6% - 80.5% (WESAD)** Feature-tokenizer transformer: 54.3% - 67.7% (CASE); 58.2% - 79.5% Feature-tokenizer transformer pretrained: 45.5% - 53.3% (CASE); 53.8% - 59.6% (WESAD) F1-score: **SVM: 51.7% to 77% (CASE); 31.3% to 81.8% (WESAD)** RF: 53.3% - 75.9% (CASE); 25.3% - 80% (WESAD) XGBoost: 53.7% - 73.6% (CASE); 23.6% - 80% MLP: 54.7% - 71.3% (CASE); 35.4% - 81.3% (WESAD) **ResNet: 58.2% - 73% (CASE); 34.8% - 81.6% (WESAD)** Feature-tokenizer transformer: 51% - 63.2% (CASE); 30.5% - 81.1% Feature-tokenizer transformer pretrained: 45.3% - 53.3% (CASE); 39.6% - 68%	arousal and valance scores WESAD: Study protocol design of baseline, amusement, stress		between experiments. Using ECG also yielded higher classification accuracy than PPG due to better signal quality and less noises or artifacts. Inclusion of various emotional states during the training stages also improves the generalizability of ML on unseen data.
Kim et al., 2024 [84]	26	ECG, RESP, Electrogastrogram	None	L	**Images test using IAPS**, 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% Multiple CNNs: 88.5%	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 with efficient channel attention (ECA): 90.2% **Multiple CNNs with shuffled ECA: 59% - 91.4%**			
Mevlevioglu et al., 2024 [91]	29 (11 females, ages between 18 and 55 years, university population)	EDA, HR, EEG	Customized self-report questionnaire	L	**Emotional virtual reality Stroop task**	Accuracy: kNN: 62% SVM: 63% **CNN: 75% (LOSO)** NN: 74% (LOSO)	Study protocol design		CNN achieved an accuracy of 75% in three-class classification (no vs. mild vs. high anxiety), and a 74.47% accuracy using window size of 5s also demonstrated the potential of real-time application
Uddin, 2023 [141]	25 of 32 (DEAP, 16 females, mean age of 24.9 years)	EEG	Self-Assessment Manikin	L	**Music-video clips viewing**	10-fold accuracy: Autoencoder NN: 77.09% (males) - 80.93% (females)	Self-Assessment Manikin scores		The study proposed a two-layered autoencoder based NN to detect emotional stress with a hybrid bag of features. It reached an average accuracy of 79.01%, while slight differences in model performance were reported when training based on genders.

Xing et al., 2020 [110]	47 (subjects with diagnosed social anxiety disorders, 34 females average age 25.7 years, 66% Caucasian) and 46 (demographically matched healthy subjects, 34 females, average age 25.8 year, 52.2% Caucasian)	fMRI	Clinical Interview for DSM-5, LSAS, Ham-A, and Ham-D	L	**Emotional face matching task**	10-fold CV accuracy: SVM (linear): 69.98% 10-fold CV ROC AUC: SVM (linear): 72% 10-fold CV sensitivity: SVM (linear): 71% 10-fold CV specificity: SVM (linear): 69%	Study protocol design		This study showed promising resulting in classifying brain activity in responses to threat or happy faces. Feature selection suggested the significance of including whole-brain activity to capture large-scale changes in neural activity

ACC: acceleration; BVP: blood volume pulse; CPT: cold pressor test; CSP: common spatial pattern; CV: cross-validation; EDA: electrodermal activity; ECG: electrocardiogram; EMG: electromyography; FBCSP: filter bank common spatial pattern; FBFgMDM: filter back FgMDM; FgMDM: Fisher geodesic filtering minimum distance to mean classifier; LDA: linear discriminant analysis; LOSO: leave-one-subject-out; LR: linear regression; LSTM: long short-term memory; RESP: respiration; ROC: receiver operating characteristic area under the curve; RT: random forest; SVM: support vector machine; TEMP: body temperature; TSST: trier social stress test. Emotional stressors are highlighted in bold font.