Supplementary Table S2. Summary of selected studies utilizing **mental stressors**.

Authors	Sample size	Physiological signals	Psychological measures	Condition	Stressor(s)	Model and Performance	Ground truth/Labels	Dataset used	Findings
Adarsh & Gangadharan, 2024 [132]	15 (WESAD, 3 females, mean age of 27.5 years, graduate students), 25 (SWELL, 8 females, average ages of 25 years, interns and students, mostly native Dutch)	ECG only from WESAD and SWELL	PANAS, STAI, SAM, and SSSQ	L	WESAD: TSST SWELL: **knowledge worker tasks**	Accuracy: Graph CNN: 97.75% - 98.84% (WESAD), 94.48%- 95.74% (SWELL) F1-score: Graph CNN: 96.48% - 97.66% (WESAD), 94.39% - 96.37% (SWELL)	Study protocol design	WESAD, SWELL	The study used ultra-short HRV extracted from ECG signals using WESAD and SWELL for binary classification. SHAP values identified mean RR and median RR are contributors. Using pruning and quantization, the graphical CNN still maintained high accuracy of ~95% in both datasets while the runtime complexity was reduced by 60%.
Agarwal et al., 2023 [58]	40 (14 females, mean age of 21.5 years)	EEG	None	L	**SCWT, Mental arithmetic task, Mirror image test**	**SVM: 89.4% - 94.3% (SCWT vs. baseline); 90.1% - 92.4% (arithmetic task vs. baseline); 86.9% - 90.5% (mirror image test vs. baseline);** DT: 67.5% - 71.9% (SCWT vs. baseline); 65.6% - 69.1% (arithmetic task vs. baseline); 63.3% - 65.7% (mirror image test vs. baseline);	Study protocol design	SAM 40	Multivariate variational mode decomposition was used to extract time-domain features from EEG signals for stress classification (each task condition vs. baseline condition). SVM consistently showed highest accuracy among DT and kNN. The study also found a 7.5-sec of EEG window size for SCWT yielded highest accuracy, while mental arithmetic task showed consistent classification accuracy across different window size.
Al-Shargie et al., 2023 [133]	45 (undergraduate students)	EEG	PSS	L	**SCWT**	Accuracy: CNN: 61% - 80.95% F1-score: CNN: 67.10% - 81.96%	Study protocol design		The study used phase locking values to evaluate the brain connectivity during four stages (rest, control, stress, and mitigation) of the experiment. CNN was able to achieve an accuracy of 80.95% in beta frequency band

Appriou et al., 2020 [134]	24 (12 females, 18 to 54 years old) for mental workload, 32 (DEAP, 16 females, mean age of 24.9 years)	EEG	STAI (Mental workload dataset) and Self-Assessment Manikin (DEAP)	L	**N-back tasks**; music-video clips viewing for DEAP dataset	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	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

Badr et al., 2023 [63]	45 (undergraduate students)	EEG	PSS	L	**SCWT**	10-fold CV Accuracy: **SVM: 71.2% - 82.5%** DT: 40.4% - 51.7% NB: 29.3% - 30.3% DA: 27.2% - 29.5% kNN: 52.3% - 54.3% 10-fold CV Sensitivity: **SVM: 91% - 94.6%** DT: 69% - 79.4% NB: 52.9% - 57% DA: 54.9% - 60.1% kNN: 84.5% - 89.4% 10-fold CV Specificity: **SVM: 91.2% - 94%** DT: 64.7% - 75.3% NB: 50.9% - 56.9% DA: 50% - 58.4% kNN: 72.5% - 76.5%	Study protocol design		The study used EEG signals to classify four-class (rest vs. control vs. stress vs. mitigation) state and SVM achieved the best accuracy using beta frequency band.
Badr et al., 2023 [64]	25 (7 females, all senior and junior undergraduate students)	EEG	None	L	**SCWT**	10-fold CV accuracy: **SVM: 99.27% - 99.58%** kNN: 99.46% - 99.88% DA: 83.58% - 94.96% NB: 77.23% - 85.76% DT: 74.31% - 87.07%	Study protocol design		SVM stood out with the highest classification accuracy using mean power spectrum density of four frequency bands separately. Using alpha or beta bands showed higher accuracy compared to when using other frequency bands. It also demonstrated the feasibility of using dry EEG cap for stress detection.

						10-fold CV sensitivity: **SVM: 99.58% - 100%** kNN: 99.65% - 99.92% DA: 85.44% - 98.84% NB: 79.43% - 86.4% DT: 73.99% - 87.21% **10-fold CV specificity: SVM: 98.96% - 99.96%** kNN: 99.08% - 100% DA: 81.74% - 98.84% NB: 75.04% - 86.40% DT: 74.65% - 87.21%			
Bahameish et al., 2024 [65]	15 (WESAD, 3 females, mean age of 27.5 years, graduate students); 25 (SWELL, 8 females, mean age of 25 years, interns and students, mostly native Dutch)	ECG only from own three datasets	PANAS, STAI, SAM, and SSSQ	L	Own dataset: **N-back task** WESAD: TSST **SWELL: knowledge worker tasks**	Nested CV accuracy: LR: 53.6% (stress vs. neutral), 85.2% (stress vs. relax) DT: 59.5% (stress vs. neutral), **85.9% (stress vs. relax)** kNN: 57.2% (stress vs. neutral), 81.6% (stress vs. relax) NB: 54.3% (stress vs. neutral), 79.3% (stress vs. relax) RF: **61.2% (stress vs. neutral)**,	Study protocol design	WESAD, SWELL	Traditional models demonstrated higher accuracy in classifying stress vs. relax (breathing exercise) than stress vs. neutral (baseline) using authors' own dataset, with RF and DT showing best results, respectively. The test of validating ML on other datasets (WESAD and SWELL) showed the capability of ML models generalize to unseen data.

						85.5% (stress vs. relax), SVM: 56.9% (stress vs. neutral), 80.6% (stress vs. relax) Nested CV F1-score: LR: 43.1% (stress vs. neutral), 85.2% (stress vs. relax) DT: 54.4% (stress vs. neutral), 85.9% (stress vs. relax) kNN: 56.3% (stress vs. neutral), 81.6% (stress vs. relax) NB: 53.8% (stress vs. neutral), 79.3% (stress vs. relax) RF: 56.2**% (stress vs. neutral)**, **89.2% (stress vs. relax)**, SVM: 51.2% (stress vs. neutral), 84.3% (stress vs. relax)			

Barki et al., 2023 [147]	14 (6 females, ages between 22 - 35 years, body mass index between 17.6 – 33.8)	PPG	None	L	**SCWT, mental arithmetic tasks**	E2E model Accuracy: CNN: 86.01% - 96.02% F1-score: CNN: 86.4% - 95.24% ROC AUC: CNN: 97.3% - 100%	Study protocol design		Using convolution wavelet transform to convert 1D PPG to a 2D scalogram of PPG, authors achieved a 92.04% accuracy with CNN. Adding Gaussian Noise for data augmentation also improves the binary classification to 96.02%
Beh et al., 2023 [66]	32(62 (CLAS, mostly students between 20 – 27 years old, two aged between 30 – 50 years); 22 (MAUS. 2 females, mean age of 23, graduate students)	PPG only from CLAS (ECG, GSR, and ear-lobe PPG) and MAUS (ECG, fingertip and wrist PPG, GSR)	None	L	CLAS**: Math problems. SCWT, logic problems;** MAUS**: N-back**	Accuracy: SVM (linear): 64.7% -81.3% (CLAS) SVM (linear): 66.7% - 75.3% (MAUS) F1-score: SVM (linear): 66.9% - 82.1% (CLAS) SVM (linear): 67.5% - 75.4% (MAUS)	Study protocol design	CLAS, MAUS	The proposed methods for preprocessing (signal quality and ML for outlier classification) and post processing (uncertainty estimation for rejecting misclassified results) for identifying and rejecting IBI outliers from PPG signals showed slight improvement in the binary classification
Benchekroun et al., 2022 [112]	46	ECG and PPG	STAI	L	**SCWT, Mental arithmetic task**	Accuracy: RF: 71% - 84% F1-score: 70% - 83% ROC AUC: 75% - 92%	Study protocol design		Two RF models were trained on ECG or PPG signals separately and showed similar performance, but the performance decreased when the trained model was tested on the other signals (71% to 73% accuracy). ECG and PPG demonstrated similar capability as physiological markers for stress

Campanella et al., 2023 [68]	29	PPG and EDA	None	L	**Lego creation tasks, math counting, o**ral presentation	10-fold CV Accuracy: **RF: 75.4% - 76.5%** SVM: 73.8% - 74.5% LR: 72.7% - 76.4% 10-fold CV F1-score: **RF: 65% - 82%** SVM: 50% - 81% LR: 59% - 82%	Study protocol design		RF showed the highest performance for binary classification when all features were used, and the top 10 features selected using Chi-test method also showed high metrics (precision: 71%, recall, 60%, and 65% F1-scores) using RF.
Chandra et al., 2024 [56]	23 (3 females, ages between 17 - 25 years)	EEG, PPG, EDA, TEMP	None	L	MIST (mental arithmetic task)	10-fold CV accuracy: **RF: 61.51% - 99.98% (EEG only); 99.52% (E2E using EDA, HR and TEMP)** kNN: 66.32% - 99.98% (EEG only); 99.06% (E2E using EDA, HR and TEMP) 10-fold CV ROC AUC: **RF: 79.96% - 99.99% (EEG only); 99.98% (E2E using EDA, HR and TEMP)** kNN: 74.51% - 99.98% (EEG only); 99.29% (E2E using EDA, HR and TEMP)	Study protocol design		Using time-frequency domain features only of EEG signal reached high accuracy with RF and kNN compared to time-domain and frequency domain features. E2E method in utilizing EDA, HR and TEMP directly also yielded similar high accuracy as EEG time-frequency features.

Choi et al., 2012 [129]	10	HR, RESP, EDA, and EMG	Customized self-report questionnaire	L	**Dual tracking, SCWT, Memory search, Mirror tracing**, public speech	Accuracy: LR: 81%	Study protocol design		Overall binary accuracy achieved by LR is 81% using EDA features and PNS/SNS ratio which takes respiration PSD into account
Cui et al., 2023 [69]	25 (SWELL, 8 females, mean ages of 25 years, interns and students, mostly native Dutch)	ECG	NASA-TLX	L	**Knowledge worker tasks**	10-fold CV accuracy: DT: 99.60%-99.92% kNN: 99.13% LR: 98.77% - 99.44% MLP: 97.04% - 99.94% NB: 91.29% - 94.86% **RF: 99.95% - 99.98%** SVM: 89.07% - 94.96% CNN: 68.31% (E2E)	Study protocol design	SWELL	The study compared different models in classifying mental stress using HRV from SWELL dataset. Principal component analysis was shown to improve model accuracy, and E2E CNN method also showed acceptable accuracy
Delmastro et al., 2020 [72]	9 (5 females, mean ages of 83.6 years; 4 males, average 74.7 years old)	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 impact the cognitive performance in SCWT
Dhaouadi et al., 2020 [139]	15 (mostly teenagers)	ECG, EMG, and EDA, ACC	PANAS, STAI and SSSQ	L	**Gaming**	Accuracy: DNN: 65% **LSTM: 95%**	Study protocol design		LSTM performed better compared to DL for binary stress classification in gamers.

Ding et al., 2022 [167]	30 (15 females, mean age of 24.37 years, graduate students)	ECG, and EDA	STAI-S, and VAD	L	Negative visual stimuli from the International Affective Picture System, 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 the state anxiety using physiological signals and provided statistical evidence of correlation between physiological and psychological measures
Fernandez et al. 2024 [74]	30 (15 females, mean age of 20.8 years, students)	EEG, ECG, EDA	SAM		**3D puzzle, mathematical calculations, Video game**	**5-fold CV accuracy: LGBoost: 63.18% - 93.89%** CNN: 32.70% - 69.03% kNN: 70% - 82% SVM: 35.06% - 63.41% 5-fold CV F1-score: **LGBoost: 67.63% - 93.91%** CNN: 33.42% - 68.05% kNN: 72.21% - 81.73% SVM: 37.62% - 61.76%	Based on ECG, EDA, feedback from post-test interview and SAM scores		Light GBoost consistently showed high accuracy in classifying binary and three-level (relax vs. neutral vs. stress) classification using EEG signals. Using EEG signals alone (85.13%) outperformed ECG and EDA signals (75.88%), but modality fusion of three signals yielded higher accuracy (91.44%).
Giannakakis et al., 2017 [75]	23 (7 females, mean age of 45.1 years old)	Facial cues, camera-based PPG	Customized self-report questionnaire	L	Social exposure (Self-describing task, text reading), emotional recall of anxious/stressful events), viewing stressful images, **SCWT,** watching 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%)

Gjoreski et al., 2016 [116]	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	In-lab: **mental arithmetic task** Wild: real-life stressor	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		Adding the context information and activity recognition models help improve the accuracy of stress detection.
Gjoreski et al., 2017, [55]	21 (in-lab: average 28 years old); 5 (wild, average 28 years old)	PPG, EDA, Temp, and ACC	In-lab: STAI Wild: Ecological Momentary Assessment	L and W	In-lab: **mental arithmetic task** Wild: real-life stressor	LOSO accuracy: SVM: **67% - 71% (in-lab)**; 55% - 71% (F1, wild) RF: 60% - 69% (in-lab), 66% - 74% (F1, wild) Boosting: 59% - 67% (in-lab), 48% - 54% (F1, wild) Bagging: 58% - 65% (in-lab), 59% - 78% (F1, wild) kNN: 52% - 65% (in-lab), 63% - 71% (F1, wild) NB: 59% - 62% (in-lab), 57% - 60% (F1, wild) DT: 52% - 59% (in-lab) , **68% - 90% (F1, wild)** Ensemble selection: 53% - 55% (in-lab),	Study protocol design		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 is 0.47. However, similar to author's previous work, adding the context information and activity recognition models significantly help improve the accuracy of stress detection for real-life data, with a highest F-score of 0.9

						62% - 80% (F1, wild) 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)			
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	Stress-Predict Dataset	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, aged between 20 to 27 years); 3 (wild: 1 female)	ECG, PPG, and EDA	None	L and W (two day of passive monitoring)	**Memory games**, 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)

Iyer et al., 2024 [117]	20 (6 females, ages between 19 – 33 years, college students)	EDA, TEMP, ACC, HR, SpO2	None	L	Physical exercise (standing, walking, jogging), m**ental arithmetic task, SCWT**, watching horror movies	**5-fold CV accuracy: LR: 90% - 99%** RF: 89% - 99% 5-fold CV F1-score: **LR: 90% - 99%** RF: 89% - 99%	Study protocol design	Non-EEG dataset for assessment of neural status	RF and LR had comparable performance for four-class classification (relaxation vs. physical stress vs. mental stress vs. emotional stress) and 1-to-2-minute window size showed to provide best classification results.
Jiao et al., 2024 [168]	35 (16 females, mean ages of 22.46 years, college students)	Pulse wave signals	PSS, STAI-Y1	L	**SCWT**	10-fold CV accuracy: SVM: 70% - 98.57% 10-fold CV F1-score: SVM: 64.98% - 98.57%	Study protocol design		The study utilized texture features (contrast, correlation, homogeneity and energy) using a sequence coding-based gray level co-occurrence matrix to convert NN intervals to word sequences for binary stress classification, and combining traditional HRV features with texture features resulting in the highest classification accuracy of 98.57%
Kader et al., 2023 [126]	20 (college students)	EEG	PSS	L	**SCWT**	10-fold CV accuracy: NB: 61.86% - 70.74%	Study protocol design		Using NB achieved an accuracy of 70.74% when using beta band PSD of a single channel in the frontal region
Kader et al., 2024 [80]	20 (college students)	EEG and EDA	None	L	**SCWT**	**10-fold CV accuracy: kNN: 65.93% - 83.83%** DA: 42.77% - 70.32% DT: 43.06% - 68.57% SVM: 62.85% - 80.04% NB: 64.52% - 76.04% 10-fold CV sensitivity:	Study protocol design		The study utilized one-channel device that measures EEG with EDA signals simultaneously for binary stress detection. Using EEG signals achieved an accuracy of 70.74% with NB, while combining EDA features improved the accuracy to ~80% with kNN and SVM models

						kNN: 68.64% - 84.22% DA: 41.57% - 58.79% DT: 42.36% - 62.85 SVM: 72.19% - 79.05% NB: 62.39% - 77.2%			
Kalra et al., 2023 [81]	15 (5 females, ages between 21 - 25 years, college students)	PPG	None	L	**SCWT**	10-fold CV accuracy: MLP: 81.2% - 83% **DNN: 88.5% - 90.5%** LR: 76.5% SVM: 82.5%	Study protocol design		The study showed the feasibility of using PPG along to classify three-level stress (low vs. moderate vs. high). DNN achieved the best accuracy of 90.5%
Palanisamy et al., 2012 [123]	10 (all females)	EMG	None	L	**SCWT**	Accuracy: kNN: 48.67% - 90.70%	Study protocol design		Using EMG features extracted from wavelet packet transform demonstrated high accuracy (90.70%) using kNN, and min values of WPT coefficients were identified as the most effective features for stress classification
Palanisamy et al., 2012 [42]	60 (30 females, average age of 22.5 years old, including Malay, Chinese and Indian)	Temp	None	L	**SCWT**	Accuracy: Probabilistic Neural Network: 23.33% - 88.75%	Study protocol design		The probabilistic neural network reached an overall highest accuracy of 88.75% for four-level classification and it was observed that Malay subjects were more affected by the SCWT compared to Chinese and Indians. Using mean body temperature with 0.001 spread factor, the model can detect the non-stressful state with 100% accuracy, and 86.67%, 81.67% and 86.67% for identifying low, medium, and high stress levels, respectively.
Kim et al., 2021 [82]	21 (7 females, mean age of 26.6 years old,)	ECG, RESP and electrogastrogram	None	L	**SCWT, mental arithmetic test**	**5-fold CV accuracy:** **SVM: 60.69% - 66.42%** LR: 60.94% - 70.15% RF: 60.02% - 63.41% DT: 52.31% - 63.73% kNN: 54.62% - 62.47%	Study protocol design		The study demonstrated the feasibility of using electrogastrogram with ECG and RESP signals to detect mental stress. Adding EEG features like mean dominant power enhances the accuracy of stress detection by up to 8%. Significant individual differences were observed and suggested electrogastrogram's potential for personalized stress assessment

						5-fold CV ROC AUC: **SVM: 63.7% - 72.9%** LR: 68.3% - 74.2% RF: 60.3% - 68.9% DT: 52.7% - 62.9% kNN: 53.5% - 69.8%			
Kim et al., 2023 [83]	67 (28 females, mean ages of 36.5 years)	PPG	Customized self-report questionnaire	L	**Knowledge worker tasks**	LOSO accuracy: kNN: 71.9% SVM: 74.3% MLP: 74.1% DT: 75.6% RF: 78.8% **LR: 76.2% - 82.6%** LOSO F1-score: kNN: 69.2% SVM: 72.3% MLP: 71.8% DT: 73% RF: 76.6% **LR: 74.7% - 80%**	Study protocol design		The study used HRV features from PPG signals to detect mental stress. Top HRV were selected to train ML models, with LR reached an accuracy of 81.7% using 3-minute window size, which has been identified as optimal window size to detecting stress.
Konar et al., 2023 [85]	38 (8 females, mean ages of 26 years)		None	L	**Mathematical problem sets of four complexity levels**	10-fold CV accuracy: kNN: 88.45% NB: 91.37% SVM: 89.71% LDA: 71.6% QDA: 73.6% **Ensemble classifier: 93.85%** 10-fold CV precision: kNN: 84.1% NB: 92% SVM: 90% LDA: 73% QDA: 75%	Study protocol design		Compared to traditional ML algorithms, majority voting-based ensemble classifier combining kNN, SVM, and NB demonstrated higher accuracy in classifying four levels of stress (no stress vs. low vs. moderate vs. high). An increased beta power in the occipital and temporal regions were found in high stress conditions.

						Ensemble classifier: 94% 10-fold CV recall: kNN: 86.3% NB: 92% SVM: 91.2% LDA: 69% QDA: 74.6% **Ensemble classifier: 94%**			
Kurniawan et al., 2013 [86]	10	EDA	None	L	**SCWT, mental arithmetic task, mental subtraction test**	10-fold CV accuracy: kMeans: 46.1% - 53.2% Gaussian Mixture Model: 66.8% - 74.9% **SVM: 70.6% - 80.7%** DTree: 62.5% - 77.8%	Study protocol design		EDA features were used to classify binary stress classification to distinguish light mental workload from recovery period or light mental workload from high mental workload. Adding speech features significantly improve the results with late fusion using SVM (70.1% - 92.47%). A decrease in performance was also observed when using leave-one-subject-out (subject independent) compared to 10-fold CV (subject dependent).
Lingelbach et al., 2021 [87]	8 (6 females, mean ages of 36.13 years)	EDA	None	L	Simulated driving task with or without **secondary cognitive tasks**	5-fold CV accuracy: LR: 62.3% LDA: 57.7% SVM: 64.1% kNN: 65.2% RFC: 63.7% GBC: 62.7% **Automated pipeline optimization ML (AutoML) 2nd generation: 71.6%** AutoML 100th generation: 59.1% CNN-based DeepResNet (E2E): 62.5%	Study protocol design		Several conventional ML have been used for three-class mental stress classification for each participant, and kNN showed the highest accuracy (63.7%) among other. Automated pipeline optimization with tree-based pipeline optimization outperformed traditional ML, with an accuracy of 71.6%. However, End-2-End DL approach did not have significant improvement on the classification accuracy.

Mamdouh et al., 2023 [89]	5 (undergraduate students, ages between 19-22 years)	EEG	None	L	**VR video games, mental arithmetic tasks**	Accuracy: SVM: 86.75% **kNN: 87.1%** LDA: 86.88% F1-score: SVM: 78.83% - 91.55% **kNN: 87.5% - 88.46%** LDA: 85.71% - 87.12%	Study protocol design		Several ML models were trained on data with mental arithmetic stressors data using EEG and evaluated on the data with VR induced stressors to examine how well the models generalize across conditions. SVM, kNN and LDA had comparable performance, with kNN having a slightly higher performance of 87.1%
Marthinsen et al., 2023 [90]	28 (12 females, mean ages 23 years, college students)	EEG	STAI-Y1 and self-report Stress Scale	L	**Mental arithmetic task**	**10-fold CV accuracy: SVM: 87.5%** CNN: 73.61% - 84.18% 10-fold CV sensitivity: SVM: 81.25% - 82.81% CNN: 87.5% 10-fold CV specificity: SVM: 90.91% - 92.05% CNN: 63.52% - 81.76%	STAI-Y1 scores		SVM, combined with wavelet scattering for feature extraction, achieved slightly higher accuracy using EEG signals with reduced channel numbers for binary classification compared to CNN models.
Mazlan et al., 2023 [124]	16	EEG	None	L	**SCWT**	**5-fold CV accuracy: kNN: 98.53% - 99.56%, 94.51% (E2E)** MLP: 75.96% - 84.14%, 78.45% (E2E) Gaussian NB: 42.57% - 46.44%, 42.57% (E2E)	Study protocol design		Using PCA reduced EEG showed the highest classification accuracy using kNN (99.56%) for multiclass stress classification (low vs. medium vs. high stress vs. baseline). Feeding raw EEG directly into the model also showed good accuracy with kNN (94.51%).

Morshed et al., 2023 [118]	97 (53 females, 4 non-binary mean age of 24.84 years)	TEMP	None	L	Public speaking, m**ental arithmetic**	**Accuracy: RF: 68.81% - 94.88%** AdaBoost: 67.40% - 93.66% LR: 63.79% - 74.08% F1-score: **RF: 65.1% - 94.41%** AdaBoost: 63.85% - 93.1% LR: 47.85% - 66.93%	Study protocol design		The study demonstrated the feasibility of using core body temperature as biomarkers of stress. Using max and mean temperature directly showed low binary classification accuracy, while accounting for the individual differences by subtracting baseline temperature features from temperature features during each task significantly improve the accuracy, with RF reaching an accuracy of 94.88%.
Mortensen et al., 2023 [137]	25 (SWELL, 8 females, mean age of 25 years, interns and students, mostly native Dutch)	ECG	NASA-TLX, Rating scale mental effort, SAM, PSS	L	**Knowledge worker tasks**	5-fold CV accuracy: CNN: 99% 5-fold CV F1-score: CNN: 100%	Self-report scores		A 1D-CNN was applied to time- and frequency-domain HRV features from ECG signals in the SWELL dataset and reached an 99% accuracy of classifying three-state (no stress vs. interruption stress vs. time-pressure stress) in mental stress.
Mozafari et al., 2021 [130]	30 (average 23.7 years old)	PPG, EDA, and RESP	None	L	**SCWT**	LOSO accuracy: LDA: 42.55% - 57.67%	Study protocol design	Mental Arousal Level Recognition Competition	Multimodal physiological signals were used to identify 5 stress levels, and fisher's feature selection and late fusion methods improved the performance by about 7%. Domain adaptation based on transfer learning also helped reduce the inter-individual variability.
Naegelin et al., 2023 [95]	90 (44 females, mean age of 23.11 years, German speaking)	HRV	Multidimensional Mood State Questionnaire	L	**Office tasks**, TSST-G in simulated office environment	F1-score SVM: 50.1% RF: 52.9% **LightGBM: 53.1%** ROC AUC SVM: 56.2% RF: 62.7% **LightGBM: 62.7%**	Self-report perceived stress level		Light GBM performed slightly higher than other models when using HRV features alone. However, authors reported adding behavior measures like mouse and keyboard features allows better multilevel stress detection (F1 score = 62.5%)
Pinge et al., 2022 [119]	5 (2 females, aged between 25 - 35 years)	ECG and PPG	None	L	Public speaking, m**ental 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 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 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 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%
Rajendran et al., 2024 [97]	11 (all males, ages between 19 to 22 years, college students)	EEG	None	L	**Video games**	Accuracy: SVM: 59.7% - 75.8% NN: 66.7% - 77.5% DT: 65.2% - 82.3% kNN: 67% - 83.3% **Ensemble: 68.3% - 87.3%** F1-score: SVM: 77.1% - 78.45% NN: 72.9% - 77.3% DT: 63.74% - 80.5% kNN: 74.33% -	Study protocol design		Using ensemble classifiers with EEG band power and power ratio achieved a good accuracy of 87.3% to classify mental stress during the video games.

						83.14% **Ensemble: 69.13% - 87.7%**			
Rashmi et al., 2023 [144]	unknown number of subjects (ages between 21 - 25 years)	EEG	None	L	**Mathematical puzzles**	Accuracy: LSTM: 76%	Study protocol design		Using LSTM, the study reported an average accuracy of 76% to classify between the 'happy' and 'stress' state and using beta band PSD resulted in highest accuracy of 86%.
Setz et al., 2010 [99]	33 (all males, mean age of 24 years)	EDA, (ECG, breathing, body movement are measured but not used)	Customized questionnaires	L	Modified MIST with 2 conditions: **cognitive load** and social stress	**LOSO accuracy:** **LDA: 75% - 83%** SVM (linear): 73% - 81% **SVM (quadratic): 77% - 82%** SVM (polynomial): 78% - 81% SVM (rbf): 75% - 81% Nearest class center: 73% - 78%	Study protocol design		Authors performed model evaluations on features with and without and baseline calibration (subtract mean values during the baseline from features obtained during other conditions. The authors also concatenated the MIST conditions with recovery period and observed a better performance when using features without baseline calibration, with LDA reaching a highest accuracy of 82.8%. SVM with quadratic kernel also achieves 81.8% using baseline corrected features. The recovery period may contain valuable information for detecting stress from cognitive load.

Shaposhnyk et al., 2023 [101]	23 (average 29.51 years old)	ECG, EDA, and TEMP	NASA Task Load Index	L	**Cognitive task with varying difficulty**	Accuracy: DT: 58.46% RF: 64.26% NB: 61.07% kNN: 58.70% LR: 63.67% SVM: 66.46% AdaBoost: 66.91% LightGBM: 67.61% XGBoost: 66.43% Forward NN: 68.19% OCTGeneticForest: 63.61% F1-score: DT: 58.23% RF: 66.93% NB: 65.2% kNN: 60.27% LR: 66.29% SVM: 70.13% AdaBoost: 70.77% LightGBM: 71.47% XGBoost: 70.05% Forward NN: 72.60% OCTGeneticForest: 72.27%	Study protocol design	CogLoad dataset	Combining physiological and personality data is effective in estimating cognitive load. Bayesian Network models enhance decision support for workload management in high-stakes environments like rescue missions.
Silva et al., 2020 [103]	83 (medical students, 63 females, ages between 17 to 38 years)	PPG	PSS (Portuguese version)	L	**Computer-based exam**	Accuracy: LR: 72.6% - 72.7% **NN: 76.15% - 76.55%** NB: 67.2% - 68.45% SVM: 64.45% - 64.8% RF: 74.35% - 73% kNN: 75.5% - 72.3%	Study protocol design		Neural networks have slightly better performance than other classic ML models, and feature selection using independent samples t-test only marginally improve the accuracy

Sim et al., 2023 [128]	40 (14 females, mean age of 21.5 years)	EEG	None	L	**SCWT, Mental arithmetic task, Mirror image test**	Accuracy: AdaBoost: 42.25% - 75.83% **RUSBoost: 40.50% - 79.08%**		SAM40	The study compared dimension reduction techniques, such as principal component analysis, independent component analysis, and LDA on improving classification accuracy. Both boosting algorithms are improved by LDA methods.
Souchet et al., 2023 [104]	93 (46 females, mean ages of 25 years, body mass index between 22.2 – 24.6)	eye tracking, ECG, EDA	STAI-Y6	L	**SCWT in VR**	**Accuracy: GBoost 89.96%** SVM (linear): 83.63% SVM: 85.84% LR (LASSO): 83.21% LR (ElasticNet): 84.04% **F1-score: GBoost 89.8%** SVM (linear): 83.39% SVM: 85.62% LR (LASSO): 82.94% LR (ElasticNet): 83.8% **ROC AUC: GBoost 97.8%** SVM (linear): 95.15% SVM: 63.78% LR (LASSO): 94.74% LR (ElasticNet): 95.73%	STAI-Y6 and behavioral data		The study utilized physiological signals and eye-tracking data to classify two levels (low and medium) of stress using the SCWT test. SVM showed the highest accuracy. STAI-Y6 was not statistically different across conditions, while using behavior data of SCWT (e.g., reaction time) for clustering showed effective distinguishment of stress level as ground truth labeling, as indicated by STAI-Y6 scores difference between two clusters.

Subhani et al., 2017 [105]	22 (ages between 19 - 25 years)	EEG	Customized self-report questionnaire	L	**MIST (mental arithmetic task with negative feedback)**	10-fold CV accuracy: LR: 77.5% - 92.25% (binary); 73.03% - **83.43% (four-class)** SVM (linear): 81.8% - 93.88% (binary); NB: 81% - **94.58% (binary)**; 74.26% - 75.03% (four-class) 10-fold CV sensitivity: LR: 73.3% - 97.5% (binary); 97.02% - 100**% (four-class)** SVM (linear): 85.8% - 96.67% (binary); NB: 84.17% - **98.33% (binary)**; 97.02% - 99.29 (four-class) 10-fold CV Specificity: LR: 81.7% - 95.5% (binary); 30% - 100**% (four-class)** SVM (linear): 77.71% - 86.4% (binary); NB: 80.83% - **93.3% (binary)**; 12.5% - 22.5% (four-class)	Study protocol design		LR and NB showed the best performance in binary classification and multiclass classification, respectively, using EEG features for mental stress detection

Swapnil et al., 2024 [106]	36 (27 females, ages between 16 - 26 years)	EEG	None	L	**Mental arithmetic task**	10-fold CV accuracy: kNN: 98.44% DT: 86.05% RF: 98.04% SVM: 99.22% XGboost: 97.78% Categorical Boost: 98.83% NN: 99.35% **Ensemble: 99.74%** 10-fold CV F1-score: kNN: 98.44% DT: 86.05% RF: 98.04% SVM: 99.22% XGboost: 97.78% Categorical Boost: 98.83% NN: 99.35% **Ensemble: 99.74%**	Study protocol design	EEG During Mental Arithmetic Tasks Dataset	The study showed high accuracy using nonlinear features (e.g., fractal dimension, entropy) and time-frequency domain features of EEG. The majority-voting ensemble, combining all six traditional ML models, reached the best accuracy of 99.74%
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% -	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 achieved the highest accuracy for in-lab dataset. Although the detection accuracy in real-world dataset is 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.

						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)			
Troyee et al., 2024 [108]	30 (ages between 19 - 25 years)	EEG	None	L	**Mathematical word problems with three levels of difficulty (visual and auditory version)**	Accuracy: DT: 90.11% - 91.71% (binary); 68% - 74% (three-class); 63.68% - 74.46% (four-class) LDA: 80.77% - 95.07% (binary); 73% - 73.15% (three-class); 69.74% - 73.48% (four-class) kNN: 93.96% - 97.48%	Study protocol design		The study used both visual and auditory versions of mental stressors for binary and multiclass classification (three and four-class) using EEG signals alone. Visual stimuli resulted in higher classification in general compared to auditory stimuli. Authors also reported improved accuracy using their proposed processing pipeline that incorporated artifact subspace reconstruction.

						(binary); 84.35% - 86% (three-class); 79.47% - 85.27% (four-class) **SVM: 94.51% - 98.32% (binary); 88% - 89.01% (three-class); 82.63% - 89.59% (four-class)** **F1-score:** DT: 57% - 66% (binary); 68% - 74% (three-class); 63.68% - 74.46% (four-class) LDA: 49% - 57% (binary); 73% - 74% (three-class); 69.74% - 73.48% (four-class) kNN: 74% - 88% (binary); 85% - 86% (three-class); 79.28% - 85.27% (four-class) **SVM: 79% - 90% (binary); 88% - 89% (three-class); 82.44% - 89.59% (four-class)**			

Wijsman et al., 2011 [125]	18 out of 30 (ages between 18 - 53 years old, 5 females)	ECG, RESP, EDA and EMG	PSS	L	**Math calculation test, logical puzzle task, memory task**	5-fold CV accuracy: Linear Bayes Normal: 78.33% Quadratic Bayes Normal: 77.78% kNN: 76.3% **Fisher's least square: 79.26%**	Study protocol design		The accuracy of four classifiers is around 80% for binary stress classification induced by three stressors, with Fisher's least squares having the highest accuracy (78.33%). Authors claimed that other studies with higher accuracy only considered a single stressor that enabled them fine-tuned the algorithms on that single stressor.
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% -	Study protocol design	CLAS, UTD, VerBIO, WESAD	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.

						87.7% (VerBIO); 73.1% - 78.1% (WESAD); RF: 61.5% - 61.9% (CLAS); **61.3% - 73.1% (UTD)**; 64% to 71.6% (VerBIO); **68.3 % - 86.5% (WESAD)**			

ACC: acceleration; BVP: blood volume pulse; CNN: convolutional neural network; CSP: common spatial pattern; CPT: cold pressor test; CV: cross-validation; DT: decision tree; EDA: electrodermal activity; ECG: electrocardiogram; EEG: electroencephalogram; EMG: electromyography; E2E: end-to-end method; FBCSP: filter bank common spatial pattern; FBFgMDM: filter back FgMDM; FgMDM: Fisher geodesic filtering minimum distance to mean classifier; GBM: gradient boost machine; HR: heart rate; IAPS: International Affective Picture System; LASSO: least absolute shrinkage and selection operator; LDA: linear discriminant analysis; LOSO: leave-one-subject-out; LR: linear regression; LSTM: long short-term memory; MIST: Montreal Imaging Stress Task; OCTGeneticForest: genetic forest of optimal complexity trees; PPG: photoplethysmography; RESP: respiration; ; ROC AUC: receiver operating characteristic, area under the curve; RT: random forest; SCWT: Stroop color-word test; SSSQ: short stress state questionnaire; STAI: state and trait anxiety inventory; SVM: support vector machine; TEMP: body temperature; TSC: time series classification; TSST: trier social stress test. Mental stressors are highlighted in bold font.