Supplementary Table S1. Summary of selected studies utilizing **social stressors**.

Authors	Sample size	Physiological signals	Psychological measures	Condition	Stressor(s)	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, mean age 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%.
Ahmad et al., 2023 [57]	9 (own dataset, aged between 20 – 40 years); 15 (WESAD, 3 females, mean age of 27.5 years, 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:	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.

						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% (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)**			
Akella et al., 2021 [59]	80 (aged between 18 - 45 years, 30 nurses, 50 non-health professionals)	EEG	None	L	**TSST**	5-fold CV accuracy**:** **SVM: 24% - 91%** AdaBoost: 50% - 55% LDA: 47% - 75% Ridge: 45% - 63% RF: 49% - 87% kNN: 45% - 89% Deep Belief Network: 13% - 23%	Study protocol design		SVM outperformed other classifiers, and the latent representation of EEG biomarkers yielded higher classification accuracy than using traditional EEG biomarkers. The use of autoencoder neural network for latent features extraction by projecting the data into the latent spaces helped reduces the noise from the input data to improve accuracy. Using raw EEG signals for 4-class classification yielded 30% accuracy.
Benita et al., 2024 [135]	15 (3 females, mean age of 27.5 years, graduate students)	PPG	PANAS, STAI, SAM, and SSSQ	L	**TSST**	Accuracy: CNN: 49.83% - 98.55% F1 score: CNN: 58.57%	Study protocol design	WESAD	CNN achieved an accuracy of 98.26% of binary stress classification using HRV features from PPG signals, selected by RF, in the WESAD dataset.

Bobade et al., 2020 [67]	15 (3 females, mean age of 27.5 years, graduate students)	BVP, ACC, EDA, RESP, ECG, EMG, and TEMP	PANAS, STAI, SAM, and SSSQ	L	**TSST**	LOSO accuracy: DT: 87.59% (binary); 68.16% (three-class) RF: 89.53% (binary); 75.95% (three-class) AdaBoost: 91.06% (binary); 78.19% (three-class) LDA: 90.15% (binary); 74.83% (three-class) kNN: 87.92% (binary); 74.71% (three-class) SVM: 93.20% (binary); 81.65% (three-class) **NN: 95.21% (binary); 84.32% (three-class)** LOSO F1-score: DT: 84.92% (binary); 53.73% (three-class) RF: 88.32% (binary); 63.09% (three-class) AdaBoost: 89.88% (binary); 67.55% (three-class) LDA: 87.60% (binary); 64.66% (three-class) kNN: 84.63% (binary); 66.76% (three-class) SVM: 92.31% (binary); 73.57% (three-class) **NN: 94.24% (binary); 78.71% (three-class)**	Study protocol design	WESAD	ANN showed the best performance in both binary and three-class classification. It is also observed that three-class classifications have overall worse results compared to binary problems.
Campanella et al., 2023 [68]	29	PPG and EDA	None	L	Lego creation tasks, Math counting, O**ral**	10-fold CV Accuracy: **RF: 75.4% - 76.5%** SVM: 73.8% - 74.5% LR: 72.7% - 76.4% 10-fold CV F1-score:	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.

					presentation	**RF: 65% - 82%** SVM: 50% - 81% LR: 59% - 82%			
Chatterjee et al., 2022 [136]	15 (3 females, mean age of 27.5 years, graduate students)	ECG, EDA, EMG, EDA, TEMP, and ACC	PANAS, STAI, SAM, and SSSQ	L	**TSST**	LOSO accuracy: CNN: 94.2% (binary), 90.3% (three-class) LOSO F1-score: CNN: 90% (binary and three-class)	STAI scores	WESAD	Light weight CNN showed promising performance in binary and three-class classification using frequency-domain features extracted using WESAD datasets.
Chauhan et al., 2023 [113]	15 (3 females, mean age of 27.5 years, graduate students)	BVP, ACC, EDA, RESP, ECG, EMG, and TEMP	PANAS, STAI, SAM, and SSSQ	L	**TSST**	Accuracy: DT: 65.6% - 83.9% RF: 68.6% - 80.7% AdaBoost: 67.2% - 79.7% Gaussian NB: 46.9% **NN: 71.3% - 93.6%** F1-score: DT: 66.9% - 82.1% RF: 67.1% - 79.7% AdaBoost: 68.9% - 78.5% Gaussian NB: 48.4% **NN: 88.9% - 92.4%**	Study protocol design	WESAD	NN performed better compared to traditional ML models, and it was found that EDA is the most contributing modality while ECG is least important. Using ACC, EDA, and BT alone reached 93.3% of accuracy using ANN.
Choi et al., 2012 [129]	10	HR, RESP, EDA, and EMG	Customized self-report questionnaire	L	Dual tracking, SCWT, memory search, Mirror tracing, and **public speech**	Accuracy: LR: 81%	Study protocol design		Overall binary accuracy achieved by LR is 81% using EDA features and parasympathetic nervous system/sympathetic nervous system ratio which takes respiration power spectrum density into account.
Dahal et al., 2023 [114]	15 (WESAD, 3 females, mean age of 27.5 years, graduate students); 25 (SWELL, 8 females, mean 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	15-fold CV accuracy: RF: 99% - 100% (WESAD, SWELL); 97% - 100% (Combined datasets) 15-fold CV F1-score: RF: 99% - 100% (WESAD, SWELL); 97% - 100% (Combined datasets)	Study protocol design	WESAD, SWELL	A RF was trained with a global framework (trained on all instances and individually tested) yielded high accuracy (99.5%) with top eight HRV features from WESAD and SWELL datasets. The accuracy ranged from 31.43% to 94.9% using leave-one-out validation due to huge individual variance.

Dziezyc et al., 2020 [11]	15 (3 females, mean age of 27.5 years, graduate students)	BVP, ACC, EDA, RESP, ECG, EMG, and TEMP	PANAS, STAI, SAM, and SSSQ	L	**TSST**	E2E methods 5-fold CV accuracy: **FCN: 79%** Resnet: 74% Time-CNN: 75% MCDCNN: 74% Stresnet: 69% MLP-LSTM: 73% Inception: 71% Encoder: 71% MLP: 72% CNN-LSTM: 70% 5-fold CV F1-score: **FCN: 73%** Resnet: 69% Time-CNN: 66% MCDCNN: 62% Stresnet: 62% MLP-LSTM: 61% Inception: 58% Encoder: 57% MLP: 57% CNN-LSTM: 56% 5-fold CV ROC AUC: **FCN: 91%** Resnet: 89% Time-CNN: 86% MCDCNN: 84% Stresnet: 82% MLP-LSTM: 82% Inception: 81% Encoder: 83% MLP: 78% CNN-LSTM: 79%	Study protocol design	WESAD	E2E FCN reaches the highest accuracy (79%) with a ROC AUC score of 91% on raw physiological signals. However, the E2E models performed poorly on other emotion datasets possibly due to difficulty in recognizing subtle affective changes in emotion datasets versus significant stress induced by TSST in WESAD datasets.
Fan et al., 2023 [148]	15 (3 females, mean age of 27.5 years, graduate students)	ECG	PANAS, STAI, SAM, and SSSQ	L	**TSST**	E2E methods 10-fold CV accuracy: Deep CNN: 96.1% - 97.2% **CNN-Convolutional Block Attention Module: 96.5% - 97.5%** 10-fold CV F1-score: Deep CNN: 95.3% - 97.8% **CNN-Convolutional Block Attention**	Study protocol design	WESAD	Deep CNN reached a high accuracy of 97.5% in three-class classification using WESAD dataset, and the implementation of attention mechanism for adding weight information to ECG features (extracted by CNN) improves the classification accuracy. It was also reported that increasing the kernel size resulted in better performance.

						Module: 96.7% - 97.1%			
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%)
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 a smart wristband achieved accuracy of 91.13% using EDA, ACC, TEMP, and HR
Henry et al., 2023 [78]	30 (CASE, 15 females, mean age of 27.15 years); 15 (WESAD, 3 females, mean age of 27.5 years, graduate students)	CASE: ECG and BVP WESAD: ECG and BVP	CASE: SAM WESAD: PANAS, STAI, SAM, and SSSQ	L	CASE: watch video clips that contain different emotions; WESAD: **TSST**	Accuracy: **SVM: 56.0% to 76.3% (CASE); 58.6% to 81.1% (WESAD)** RF: 58% - 75.2% (CASE); 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% -	CASE: customized anxiety index from self-report arousal and valance scores WESAD: Study protocol design of baseline, amusement, stress	CASE, WESAD	SVM performed slightly better than other models in both within and between experiments. Using ECG also yielded higher classification accuracy than PPG due to better signal quality and less noise or artifacts. Inclusion of various emotional states during the training stages also improves the generalizability of ML on unseen data.

						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%			
Huynh et al., 2021 [149]	15 (3 females, mean age of 27.5 years, graduate students)	BVP, ACC, EDA, RESP, ECG, EMG, and TEMP	PANAS, STAI, SAM, and SSSQ	L	**TSST**	E2E methods LOSO accuracy: MLP: 55.17% - 83.19% (binary); 52.16% - 78.11% (three-class) FCN: 67.88% - 84.15% (binary); 46.53% - 79.04% (three-class) ResNet: 62.54 - 83.14% (binary); 45.25% - 79.48%	Study protocol design	WESAD	Manual designed E2E DNN showed slightly better performance compared to FB models, and StressNAS further enhanced model performance in both binary and three class classification. Using all sensor modality also yields best performance.

						(three-class) **StressNAS: 71.46 % - 93.14% (binary); 55.81% - 83.43% (three-class)**			
Jahanjoo et al., 2024 [79]	15 (3 females, mean age of 27.5 years, graduate students)	PPG	PANAS, STAI, SAM, and SSSQ	L	**TSST**	LOSO accuracy: AdaBoost: 87.8% - 91.64% kNN: 76.08% - 88.51% LDA: 86.86% - 95.04% RF: 83.33% - 91.83% DT: 78.49% - 93.41% **SVM: 92.69% - 95.55%** LOSO F1-score: AdaBoost: 68.91% - 86.99% kNN: 76.08% - 88.51% LDA: 79.82% - 89.6% RF: 66.88% - 86.99% DT: 63.60% - 88.27% **SVM: 89.68% - 91.42%** LOSO ROC AUC: AdaBoost: 82.21% - 92.01% kNN: 83.22% - 93.04% LDA: 86.16% - 92.42% RF: 78.66% - 89.71% DT: 76.78% - 92.77% **SVM: 91.37% - 93.25%**	Study protocol design	WESAD	The study used the proposed 3-step (bandpass, moving average, Kalman) pipeline to denoise and extract HRV from BVP PPG signals, and SVM achieved the best accuracy of 95.55%
Jain & Kumar et al., 2024 [127]	15 (3 females, mean age of 27.5 years, graduate students)	EDA, ECG	PANAS, STAI, SAM, and SSSQ	L	**TSST**	LOSO accuracy: DT: 71% - 89.9% **Gradient Boosting DT: 77.2% - 97.3%**	STAI scores	WESAD	The study sued EDA and ECG signals from the WESAD dataset for three-class (low vs. moderate, vs high anxiety) classification, and higher accuracy of 97.3% was achieved using gradient boosting DT when EDA and ECG were combined

									with contextual features (e.g., caffeine consumption, sleep time, smoking habits etc.).
Masrur et al., 2024 [131]	15 (3 females, mean age of 27.5 years, graduate students); 14 (Apple Watch dataset, university students)	ECG from WESAD and Apple watch dataset	PANAS, STAI, SAM, and SSSQ	L	**TSST**	Accuracy: Ensemble: 68.86% - 94.17% **DNN: 70.17% - 94.66%** Precision: Ensemble: 66.94% - 93.4% **DNN: 65.06% - 88.91%** Recall: Ensemble: 64.95% - 83.91% **DNN: 71.24% - 90.31%**	Study protocol design	WESAD, Real-time Apple Watch ECG dataset	DNN achieved the highest accuracy of 94.66% using ECG signals alone from the WESDA dataset. Ensemble models (RD, NB, LightGBM) also achieved comparable performance. Both models remain high accuracy when train on one data set and validate on the other dataset.
Morshed et al., 2024 [118]	97 (53 females, 4 non-binary mean age of 24.84 years)	TEMP	None	L	**Public speaking**, mental 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%.
Mozos et al., 2017 [21]	18 (aged between 18 39 years, college students)	EDA, PPG, HRV	STAI	L	**TSST**	Accuracy: **AdaBoost: 79% - 94%** SVM (RBF): 92% - 93% SVM (linear): 80% - 85% kNN: 62% - 87% Precision: **AdaBoost: 79% - 94%** SVM (RBF): 93%	Study protocol design		AdaBoost and SVM with RBF kernel showed better results in the binary classification. Combining physiological signals and body activity patterns improve the classification accuracy.

						SVM (linear): 84% kNN: 87% Recall: **AdaBoost: 86% - 96%** SVM (RBF): 96% SVM (linear): 94% kNN: 94%			
Naegelin et al., 2023 [95]	90 (44 females, mean age of 23.11 years, German speaking)	ECG	Multidimensional Mood State Questionnaire and a customized self-report questionnaire	L	Office tasks, **TSST-Group** 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 scores		Using HRV features alone, LightGBM slightly performed better in detecting three perceived stress levels (low vs. medium vs. high) compared to SVM and RF. Combining behavior measures like mouse and keyboard features with HRV improved the classification accuracy to 61.2%, while using mouse and keyboard features only resulted in a higher accuracy of 62.5%. SHAP values also indicate that mouse movement and typing behaviors are indicative for work-related mental stress
Onim & Thapliyal, 2024 [143]	40 (28 females, aged between 60-80 years)	EDA, BVP, TEMP	None	L	**TSST**	Accuracy: RNN: 94% **LSTM: 96%** F1-score: RNN: 83% - 97% **LSTM: 89% - 97%**	Saliva cortisol level		RNN and LSTM showed similar accuracy in binary classification using physiological signals from Empatica E4 wristband, while LSTM demonstrated higher recall and F1-scores. Authors used cortisol as ground truth, and it was argued the cortisol could be used as an unbiased marker for labeling.
Pinge et al., 2022 [119]	5 (2 females, age between 25 to 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);	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% (in-lab) **J48 AdaBoost: 88%**	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

	17 out of 21 (wild condition data)					**- 90.17% (in-lab)** SVM: 88.17% - 89.17% (in-lab) Correlation between self-report and predicted stress level: 0.71 (wild)			individual stressors, the performance ranges between 82% to 99%. Normalization using baseline conditions also marginally improved the classification accuracy. Only 33.55% of wild data were preserved due to the presence of noise and missing data, and a correlation of 0.71 was obtained between ground truth and predicted stress level. The physiological model estimated 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%/
Praveenkumar et al., 2022 [140]	15 (3 females, mean age of 27.5 years, graduate students)	ECG, ACC, RESP, TEMP	PANAS, STAI, SAM, and SSSQ	L	**TSST**	Accuracy: DNN 76% **LSTM: 82%** Precision: DNN: 32% - 58% **LSTM: 66% - 77%** Recall: DNN: 63% - 72% **LSTM: 67% - 76%**	Study protocol design	WESAD	LSTM and DNN were used to classify stress using WESAD dataset with deep local features developed from a CNN-based model, and LSTM was observed to have slightly higher performance.
Quadir et al., 2023 [120]	15 (3 females, mean age of 27.5 years, graduate students)	ECG, EDA, TEMP	PANAS, STAI, SAM, and SSSQ	L	**TSST**	Accuracy: RF: 98% F1-score: RF: 95%	Study protocol design	WESAD	RF was applied on WESAD dataset to detect stress using ECG, GSR, and TEMP features only, and GSR features were found to contribute to the performance the most. The authors also proposed an Internet of Things framework using similar modality collected via Arduino for potential data collection in real-life condition in patients

Ragav et al., 2019 [150]	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)	WESAD: BVP, ACC, EDA, RESP, ECG, EMG, and TEMP SWELL: HRV and EDA	PANAS, STAI, SAM, and SSSQ	L	WESAD: **TSST** SWELL: knowledge worker tasks	E2E methods Accuracy: Eearly-Stopping Multiple Instance (EMI)-LSTM: 93.9% (WESAD); 86.79% (SWELL) EMI-GRU: 85.54% (WESAD); 84.74% (SWELL) **EMI-FastGRNN: 99.86% (WESAD); 87.87% (SWELL)**	Study protocol design	WESAD, SWELL	Early Stopping - Multiple Instance Learning was adopted to evaluate their efficiency on stress detection using publicly available datasets. They utilized raw physiological signals and had better classification performance than benchmark models, and shorter inference time, with EMI-FastGRNN reached the highest accuracy.
Rashid et al., 2023 [121]	15 (3 females, mean age of 27.5 years, graduate students)	BVP, ACC, EDA, RESP, ECG, EMG, and TEMP	PANAS, STAI, SAM, and SSSQ	L	**TSST**	LOSO Accuracy: DT: 54.77% - 84.27% (binary); 45.37% - 62.32% (three-class) RF: 69.12% - 89.01% (binary); 54.5% - 76.62% (three-class) AB: 69.16% - 88.96% (binary); 55.92% - 75.78% (three-class) LDA: 39.07% - 84.19% (binary); 38.2% - 69.63% (three-class) kNN: 53.44% - 69.04% (binary); 45.01% - 86.56% (three-class) **SELFCARE: 92.17% - 94.12% (binary), 84.59% - 86.34% (three-class)** LOSO F1-score: DT: 52.55% - 80.8% (binary); 39.86% - 56.23% (three-class) RF: 53.69% - 86.37% (binary); 47.94% - 62.88% (three-class) AB: 64.19% - 86.13% (binary); 45.72% - 64.72% (three-class)	Study protocol design	WESAD	RF showed the best performance in both binary (89.01%) and three-class (76.62%) classification using the wrist device modality BVP, EDA, and TEMP. Using chest device modality yielded slightly lower classification accuracy. A generalized selective sensor fusion method, SELF-CARE, was proposed for processing features and reducing noises, outperforming the traditional ML models. ACC from writs device and EMG from the chest devices are representative of stress, as selected by SELF-CARE

						LDA: 32.61% - 85.66% (binary); 23.22% - 60.23% (three-class) kNN: 61.97% - 83.74% (binary); 37.45% - 65.73% (three-class) **SELFCARE: 90.72% - 92.93% (binary); 71.97% - 79.31% (three-class)**			
Sah et al., 2022 [166]	15 (3 females, mean age of 27.5 years, graduate students)	EDA	PANAS, STAI, SAM, and SSSQ	L	**TSST**	10-fold CV accuracy: Stressalyzer (CNN based): 90% (binary); 80% (three-class) 10-fold CV F1-score: Stressalyzer (CNN based): 86% - 90% (binary); 75% - 79% (three-class)	Study protocol design	WESAD	The CNN based stressalyzer achieved 92% in binary and 82% in three class classification using EDA features only from WESAD dataset. Leave-one-subject-out validation showed large interindividual variance, and accuracy was improved by ~30% after retraining with personalization using an online learning paradigm.
Sandulescu et al., 2015 [98]	5 (18 - 39 years old, college students, including both females and males)	BVP, EDA, HRV	STAI	L	**TSST**	Accuracy: SVM (RBF): 73.26% - 83.08% Precision: SVM (RBF): 73.61% - 83.87%	Study protocol design		SVM was trained and tested on each participant separately, and two participants' data reaches accuracy over 82%.
Sandulescu et al., 2015 [22]	5 (18 - 39 years old, 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.

Schmidt et al., 2018 [122]	15 (3 females, mean age of 27.5 years, graduate students)	BVP, ACC, EDA, RESP, ECG, EMG, and TEMP	PANAS, STAI, SAM, and SSSQ	L	**TSST**	LOSO accuracy: DT: 64.8% - 85.16% (binary); 41% - 63.56% (three-class) RF: 63.44% - 88.33% (binary); 45% - 76.17% (three-class) **AB:** 62.88% - 89.77% (binary); 49.39% - **80.34% (three-class)** **LDA:** 60.02% - **93.12% (binary)**; 47.73% - 78.19% (three-class) kNN: 57.81% - 83.16% (binary); 40.03% - 59.44% (three-class) LOSO F1-score: DT: 55.36% - 83.03% (binary); 41.46% - 59.08% (three-class) RF: 59.02% - 90.44 (binary); 38.10% - 66.33% (three-class) AB: 50.84% - 89.57% (binary); 37.68% - 71.10% (three-class) LDA: 41% - 90.74% (binary); 30.96% - 72.49% (three-class) kNN: 47.79% - 79.44% (binary); 35.18% - 52.94% (three-class)	Study protocol design	WESAD	Chest-based physiological modalities showed highest accuracy in both three-class and binary classification using AB (80.34%) and LDA (93.12%) respectively. Overall lower performance in three-class classification were observed potentially due to subtle difference in physiological responses in baseline and amusement conditions and self-perceived affective state.
Setz et al., 2010 [99]	33 (all males, mean age of 24 years, mean body mass index of 23.63)	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% - 82.8%** SVM (linear): 73% - 81% **SVM (quadratic): 77% - 81.3%** SVM (polynomial): 78% - 81% SVM (RBF): 75% - 81%	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

						Nearest class center: 73% - 78%			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.
Shanbhog et al., 2024 [100]	15 (3 females, mean age of 27.5 years, graduate students)	EDA	PANAS, STAI, SAM, and SSSQ	L	**TSST**	Accuracy: RF: 93.33% SVM: 86.67% kNN: 80% **DT: 100%** LR: 46.7% NB: 86.7% F1-score: **RF: 100%** **SVM: 100%** kNN: 55.6% DT: 100% LR: 55.6% NB: 55.6%	Labels from k-mean clusters	WESAD	This study aims to classify levels of stress in subjects when they are exposed to TSST using EDA from the chest device in WESAD dataset. K-means clustering was applied to six time-domain features of EDA to cluster 15 subjects into three levels of stress: low, moderate and high. Then traditional ML algorithms were trained to classify three levels of stress, with DT reaching 100% accuracy. Feature importance scores showed that it means normalized EDA as the most contributing feature, followed by mean peak amplitude, max peak amplitude and min peak amplitude.
Tanwar et al., 2023 [153]	15 (3 females, mean age of 27.5 years, graduate students)	ACC, EDA, RESP, ECG, EMG, and TEMP	PANAS, STAI, SAM, and SSSQ	L	**TSST**	E2E methods Accuracy: XGBoost: 53.90% - 95.95% **Light GBoost: 53.89% - 98.87%** Category Boost: 53.89% - 98.84%	Study protocol design	WESAD	Fusion of different physiological modalities of chest device from WESAD showed improved accuracy in three-level classification (baseline vs. stress vs. amusement) using E2E models. The combination of ACC, ECG, EMG, EDA and Temperature yielded high accuracy using XGBoost, while LGBoost and CatBoost showed the best results using acceleration, ECG, EDA, RESP and TEMP. SHAP values indicated EDA as the most important indicator of stress, while ECG being the least important modality.
Tigranyan & Martirosyan, 2024 [145]	15 (3 females, mean age of 27.5 years, graduate students)	ECG	PANAS, STAI, SAM, and SSSQ	L	**TSST**	Accuracy: ResNet: 57.6% - 99% F1-score: ResNet: 34.1% - 97.7% ROC AUC:	Study protocol design	WESAD	The study proposed a deep learning framework for inter-subject stress detection using WESAD dataset. The framework includes a self-supervised contrastive pretrained network and a data pseudo-labeling approach to enhance the classification results,

						ResNet: 73.2% - 99.9%			with the classification accuracy of 93.1%.
Toshnazarov et al., 2024 [107]	28 (15 females, mean ages of 22/8 years, 16 university students, 12 from the general population)	PPG, ACC	Ecological momentary assessment	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-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); 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 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.

Vaz et al., 2023 [24]	15 (3 females, mean age of 27.5 years, graduate students)	ECG, EDA, and EMG	PANAS, STAI, SAM, and SSSQ	L	**TSST**	5-fold CV F1 score: LR: ~50% - 65% LDA: ~58% - 61% DT: ~55% - 65% SVM: ~50% - 60% ADB: ~75% - 85% RF: ~65% - 83% **XGBoost: ~70% - 83%** 5-fold CV ROC AUC: LR: ~78% - 85% LDA: ~62% - 72% DT: ~71% - 79% SVM: ~70% - 82% ADB: ~91% - 92% **RF: ~98% - 99%** XGBoost: ~91% - 98%	STAI scores	WESAD	RF and XGBoost showed the best performance in binary classification in terms of F1 scores and AUCC scores. Feature selection techniques found ECG derived features to be the most relevant features, and SMOTE techniques used to address imbalanced classes improved the classifier performances.
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: 3 females, mean age of 27.5 years, graduate students)	EDA only from multiple datasets	PANAS, STAI, SAM, and SSSQ	L	CLAS: Matth 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%	Study protocol design	CLAS, UTD, VerBIO, WESAD	Using EDA features alone showed high classification accuracy in multiple datasets with different models. Gender also played a role in model performance that models trained on females have higher accuracy than males.

						to 71.6% (VerBIO); **68.3% - 86.5% (WESAD)**			

ACC: acceleration; BVP: blood volume pulse; CNN: convolutional neural network; CPT: cold pressor test; CV: cross-validation; DT: decision tree; ECG: electrocardiogram; EDA: electrodermal activity; EEG: electroencephalogram; EMG: electromyography; E2E: end-to-end method; GBM: gradient boosting machines; GRNN: general regression neural network; GRU: gated recurrent unit; HR: heart rate; LDA: linear discriminant analysis; LOSO: leave-one-subject-out; LR: linear regression; LSTM: long short-term memory; PANAS: Positive Affect and Negative Affect Schedule; RESP: respiration; RT: random forest; SAM: self-assessment manikin; ROC AUC: receiver operating characteristic, area under the curve; SCWT: Stroop color-word test; SSSQ: short stress state questionnaire; STAI: state and trait anxiety inventory; SVM: support vector machine; TEMP: body temperature; TSST: trier social stress test. Social stressors are highlighted in bold font.