Supplementary Table S5. Summary of selected studies utilizing **driving stressors**.

Authors	Sample size	Physiological signals	Psychological measures	Condition	Stressor(s)	Model and Performance	Ground truth/Labels	Dataset used	Findings
Amin et al., 2023 [146]	17 out of 27 (SRAD), 14 (AffectiveROAD)	ECG, HR, EDA, EMG, and RESP from SRAD; BR, EDA, BVP, ECG, TEMP, ACC, posture, and activity from AffectiveROAD	Customized questionnaire	Semi-wild	**Driving task**	E2E methods Accuracy: 1D CNN: **92.7% (SRAD binary)**; 86.53% - 95.6% (AffectiveROAD binary); 79.5% (SRAD three-class); 76.5% - 85.66% (AffeictiveROAD three-class) 1D CNN-LSTM: 91.8% (SRAD binary); **88.83% - 96.59% (AffectiveROAD binary); 85.6% (SRAD three-class); 76.22% - 87.95% (AffectiveROAD three-class)** **F1-score:** 1D CNN: **89.42% - 94.44% (SRAD binary)**; 86.22% - 96.2% (AffectiveROAD binary); 64.75% - 90.74% (SRAD three-class); 19.35% - 94.06% (AffeictiveROAD three-class) 1D CNN-LSTM: 88.89% - 93.52% (SRAD binary); **88.68% - 96.85% (AffectiveROAD binary); 71.53% - 92.62% (SRAD three-class); 46.91% - 95.15% (AffectiveROAD three-class)**	SRAD: study protocol design AffectiveROAD: self0-report stress level and video recordings	SRAD, AffectiveROAD	Deep learning approaches without explicitly feature extraction with 1D CNN and 1D CNN-LSTM showed promising results in distinguishing binary and three-level mental stress in drivers using 2 publicly available datasets. Data fusion of modalities two wearables achieved the first ranks as evaluated by fuzzy evaluation based on distance from average solution
Akbas, 2011 [170]	10 (healthy drivers)	ECG, EMG, RESP, EDA, EDA	Customized questionnaire	Semi-wild	**Driving task**	N/A	Study protocol design	SRAD	The study applied statistic test on physiological signals and found that increased heart rate, muscle activity and skin conductance were correlated with elevated stress level. However, no ML was applied

Arya et al., 2023 [62]	17 out of 27 (healthy drivers)	ECG, EDA only from SRAD	Customized questionnaire	Semi-wild	**Driving task**	K-fold CV Accuracy (k not specified by authors): CNN: 94.2% (binary); 85.60% (three-class) **LSTM: 95.67% (binary); 88.70% (three-class)** CNN-LSTM: 94.75% (binary); 87.56% (three-class) SVM: 83.92% (binary); 64.93% (three-class) Gradient Boosting: 88.89% (binary); 70.03% (three-class) LR: 76.32% (binary); 52.45% (three-class) MLP: 80.05% (binary); 53.95% (three-class) RF: 93.77% (binary); 82.42% (three-class) DT: 91.67% (binary); 79.53% (three-class) kNN: 87.83% (binary); 76.12% (three-class) K-fold CV F1-score (k not specified by authors): CNN: 93% (binary); 87% (three-class) **LSTM: 96% (binary); 90% (three-class)** CNN-LSTM: 95% (binary); 88% (three-class) SVM: 79% (binary); 68% (three-class) Gradient Boosting: 84% (binary); 68% (three-class) LR: 68% (binary); 50% (three-class) MLP: 82% (binary); 54% (three-class) RF: 89% (binary); 83% (three-class) DT: 86% (binary); 82% (three-class) kNN: 80% (binary); 77% (three-class)	Self-report stress level for binary classificatio; Study protocol design for multiclass classification	SRAD	RF showed the best performance in binary and 3-level classifications as classic models, and LSTM also outperformed other deep learning models in both classifications. Overall higher performance of deep learning models compared to classic models suggest the potential of advantages of using DL in stress detection

Cruz et al., 2020 [70]	27 (healthy drivers)	ECG	Customized questionnaire	Semi-wild	**Driving task**	Accuracy: Cubic SVM (Gaussian kernel): 92.6% **Tree optimized SVM: 88.9% - 96.3%**	Study protocol design	SRAD	The combination of decision tree and SVM showed the highest performance on detecting stress during the driving test using ECG signals only
Dalmeida et al., 2021 [71]	27 (SRAD: healthy drivers) 4 (own dataset)	ECG	Customized questionnaire	Semi-wild	**Driving task**	ROC AUC NB: 60% (best features , 60% (HRV features) kNN: 80% (best SRAD features), 74% (HRV features only) SVM: 81% (best SRAD features), 74% (HRV features only) **MLP: 83% (best SRAD features), 75% (HRV features only)** RF: 85% (best SRAD features), 77% (HRV features only) Gradient Boost: 85% (best SRAD features), 73% (HRV features only) Prediction probabilities of 71% and 79% were obtained when using HRV features from Apple watch in authors' own dataset to classify a stress state and a relaxing state, respectively F1-score: NB: 61% (best features , 63% (HRV features) kNN: 74% (best SRAD features), 71% (HRV features only) SVM: 77% (best SRAD features), 74% (HRV features only) **MLP: 77% (best SRAD features), 72% (HRV features only)** RF: 78% (best SRAD features), 72% (HRV features only) Gradient Boost: 79% (best SRAD features), 70% (HRV features only)	The median GSR values were used as a cut-off for labeling the stress state	SRAD	MLP showed the highest sensitivity score in the stress classification task using HRV features. The models trained on the SRAD dataset HRV features and then tested on a different dataset, though the performance metrics were not reported in the paper

Lingelbach et al., 2021 [87]	8 (mean age 36.13 years, two males)	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% RF: 63.7% GB: 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 has 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 than traditional ML, with an accuracy of 71.6%. However, E2E approach did not have significant improvement on the classification accuracy.
Liu et al., 2023 [88]	27 (healthy drivers)	ECG only from SRAD	Customized questionnaire	Semi-wild	**Driving task**	ROC AUC: SGD: 0.781 SVM: 0.866 **RF: 91.5%** **Tree-based pipeline optimization with extra trees classifier: 93.4%**	The windows with the foot GSR values greater than first 80% is labeled as the stress state, and the window with foot GSR values smaller than first 20% are labeled as no stress state	SRAD	RF showed the highest AUC values (91.5%), and TPOT further improved the AUC to 93.4% for binary mental stress classification in drivers using HRV features only, with the best window length of 29 seconds and a window overlap size of 7 seconds.
Meteier et al., 2021, 2022, and 2023 [92–94]	90 young participants (average 24.15 years old, 49 females, and 1 identify as other, mostly students)	EDA, ECG, and RESP	NASA Task Load Index	L	**Simulated driving task with or without secondary verbal cognitive tasks**	10-fold CV accuracy: RF: 73% - 89% MLP: 94% **SVM: 90% - 95%**	Study protocol design		Physiological signals, particularly respiration, effectively detect mental workload during automated driving, supporting real-time monitoring to enhance driver safety under automated conditions, suggesting the viability of implementing workload classification in real-world settings with sensor fusion.
Siam et al., 2023 [102]	17 out of 27 (healthy drivers)	ECG, EDA, EMG, Respiration	Customized questionnaire	Semi-wild	**Driving task**	5-fold CV accuracy: kNN: 91.2% LR: 91.02% MLP: 92.8% DT: 94.01% SVM: 95.2% **RF: 98.2%** 5-fold CV F1-score: kNN: 91% LR: 91%	Study protocol design	SRAD	RF outperformed other ML models (98.2%) for three-class classification of mental stress using only 10 features, implicating the feasibility of applying stress classification in real-world scenarios with short recording time (1 minute).

						MLP: 93% DT: 92% SVM: 95% **RF: 98%**			
Zontone et al., 2023 [138]	18 (Own dataset: 18 subjects) 9 (SRAD)	EDA and ECG	Customized questionnaire	L (own dataset) Semi-wild (SRAD)	**Driving task**	Accuracy: CNN: 84.35% - 91.78% (own dataset); 96.64% - 99.24% (SRAD) Sensitivity: CNN: 85.41% - 89.32% (own dataset); 94.59% - 98.92% (SRAD) Specificity: CNN: 82.71% - 94.30% (own dataset); 98.07% - 99.52% (SRAD)	Study protocol design	SRAD	A 2D CNN was applied to the scalogram plots generated from EDA and ECG features for binary classification, yielding a 91.78% accuracy on data collected by authors, and higher accuracy of 99.24% was reached on SRAD dataset.

ACC: acceleration; BR: breathing rate; BVP: blood volume pulse; CNN: convolutional neural network; CV: cross-validation; DT: decision tree; EDA: electrodermal activity; ECG: electrocardiogram; EMG: electromyography; E2E: end-to-end method;; LDA: linear discriminant analysis; LOSO: leave-one-subject-out; LR: linear regression; LSTM: long short-term memory; RESP: respiration; ROC AUC: receiver operating characteristic area under the curve; RT: random forest; SVM: support vector machine; TEMP: body temperature; TSST: trier social stress test; MLP: multilayer perceptron; GSR: galvanic skin response. Driving stressors are highlighted in bold font.