NEURAL INFORMATION PROCESSING SYSTEMS

FEEL: Quantifying Heterogeneity in Physiological Signals for Generalizable Emotion Recognition

Pragya Singh, Ankush Gupta, Somay Jalan,
Mohan Kumar, Pushpendra Singh

IIIT-DELHI
CENTRE OF EXCELLENCE
HUMAN-CENTERED COMPUTING

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY DELHI

mélange

RIT Rochester Institute of Technology

Physiological signals has potential to capture emotional well-being in everyday life.

Progress in the field is limited by the **fragmented nature of existing datasets**, making it difficult to build **generalizable emotion recognition models**.

Introducing FEEL
the first large-scale benchmarking framework to evaluate emotion recognition models trained on physiological signals across 19 open-source datasets.

Benchmarking Steps

Data Curation

We gathered **19 datasets** with EDA and PPG signal data.

Covering lab, constraint, and real-world settings.

Spanning different devices, sampling rates, and labeling methods.

Data Preprocessing and Standardization

We applied a unified pipeline for signal cleaning, segmentation, normalization, labeling and feature extraction, to ensure all datasets follow a **consistent format** for fair comparison.

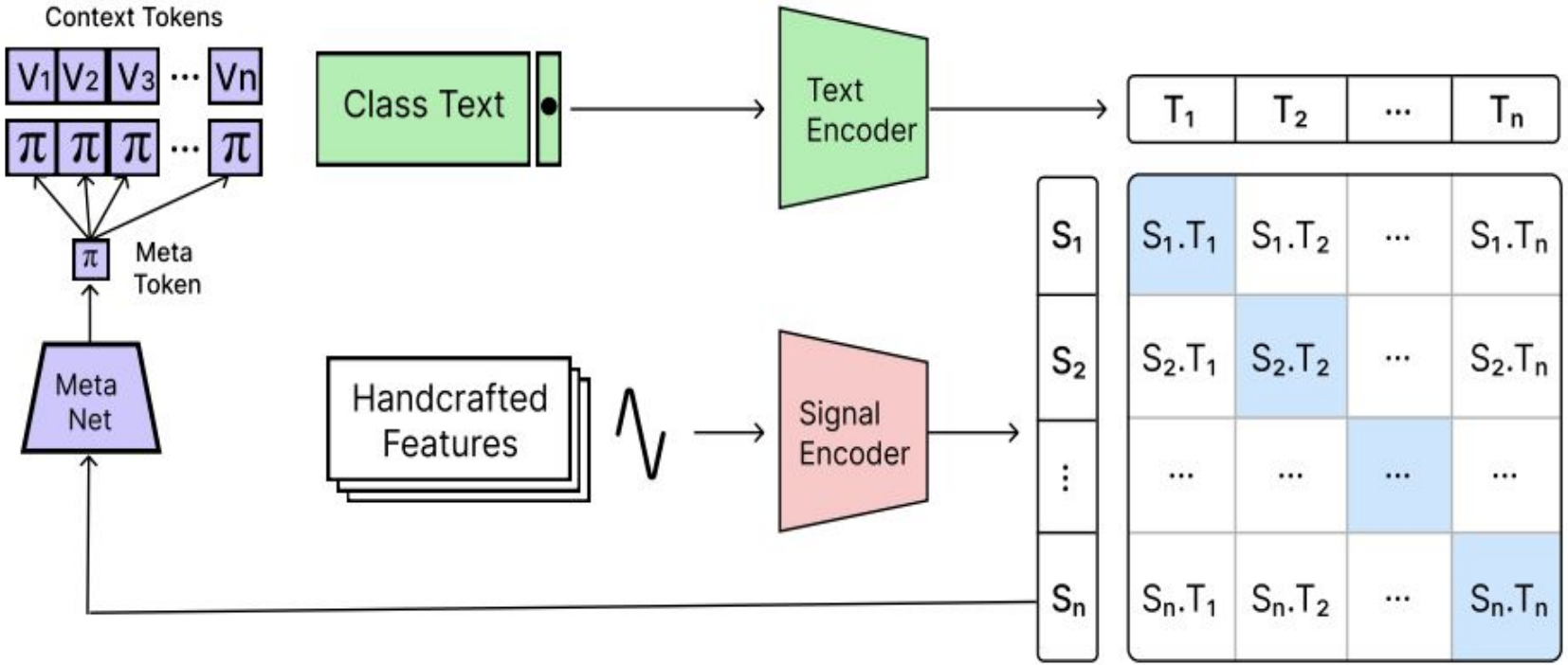
Datasets Benchmarking

We benchmarked 16 models across 19 standardized datasets to evaluate within-dataset performance and cross-dataset generalization..

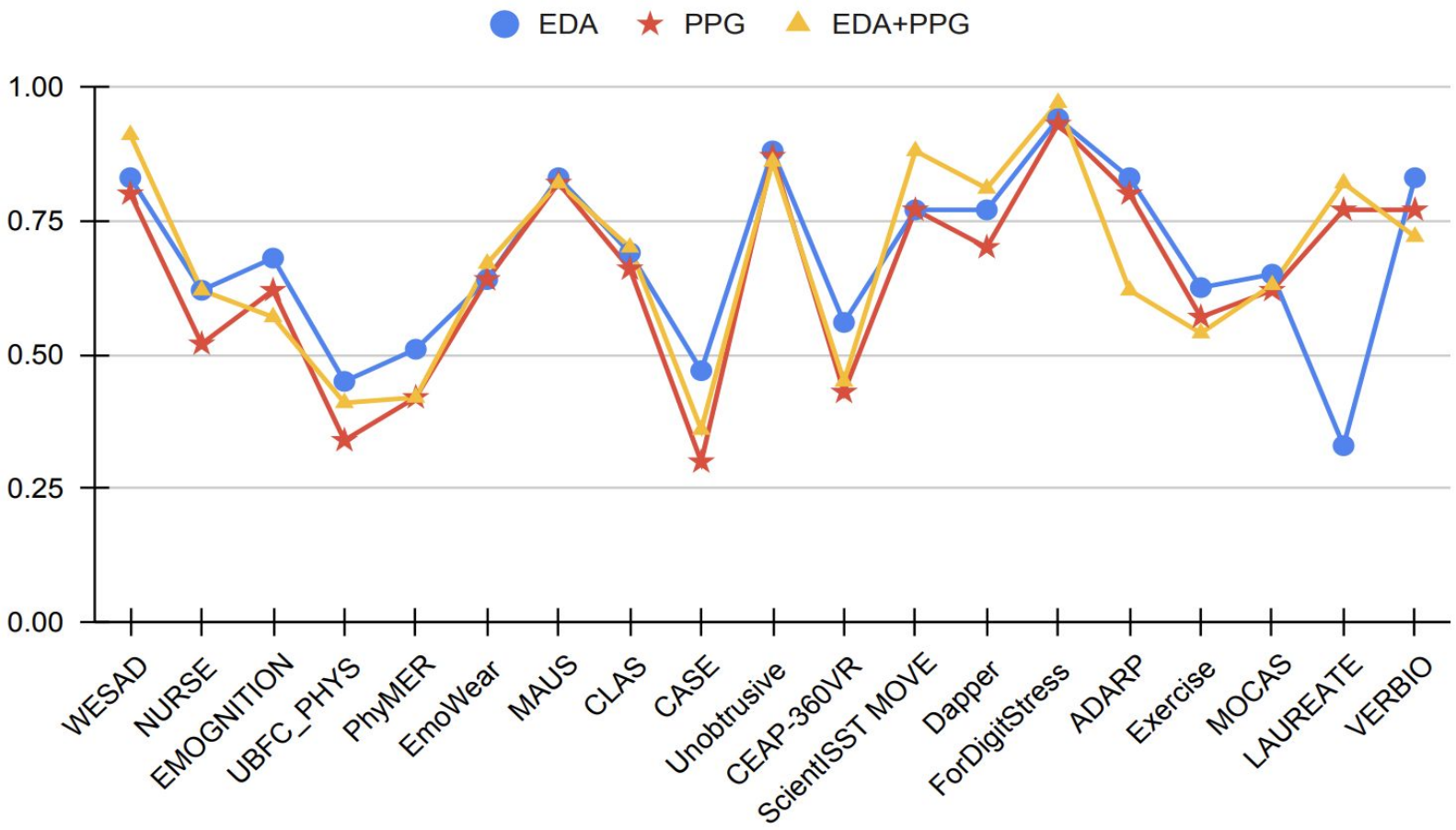
Benchmarking Overview

Paradigm	Models	Motivation
1 Traditional ML	RF, LDA	Strong, interpretable baselines; effective for small datasets
2 Deep Learning on Handcrafted Features	MLP, ResNet, LSTM+NN, Attention+NN	Combine domain knowledge with nonlinear learning for improved robustness
3 Deep Learning on Raw Signals	ResNet, LSTM+MLP, CNN + Transformer Encoder	End-to-end learning from signals without manual feature extraction
4 Pretrained Representation Learning	CLSP (Zero-shot & Fine-tuned: 5%, 25%, 50%)	Assess transferability and adaptability of pretrained and fine-tuned physiological models

Language-Signal Fine-tuning



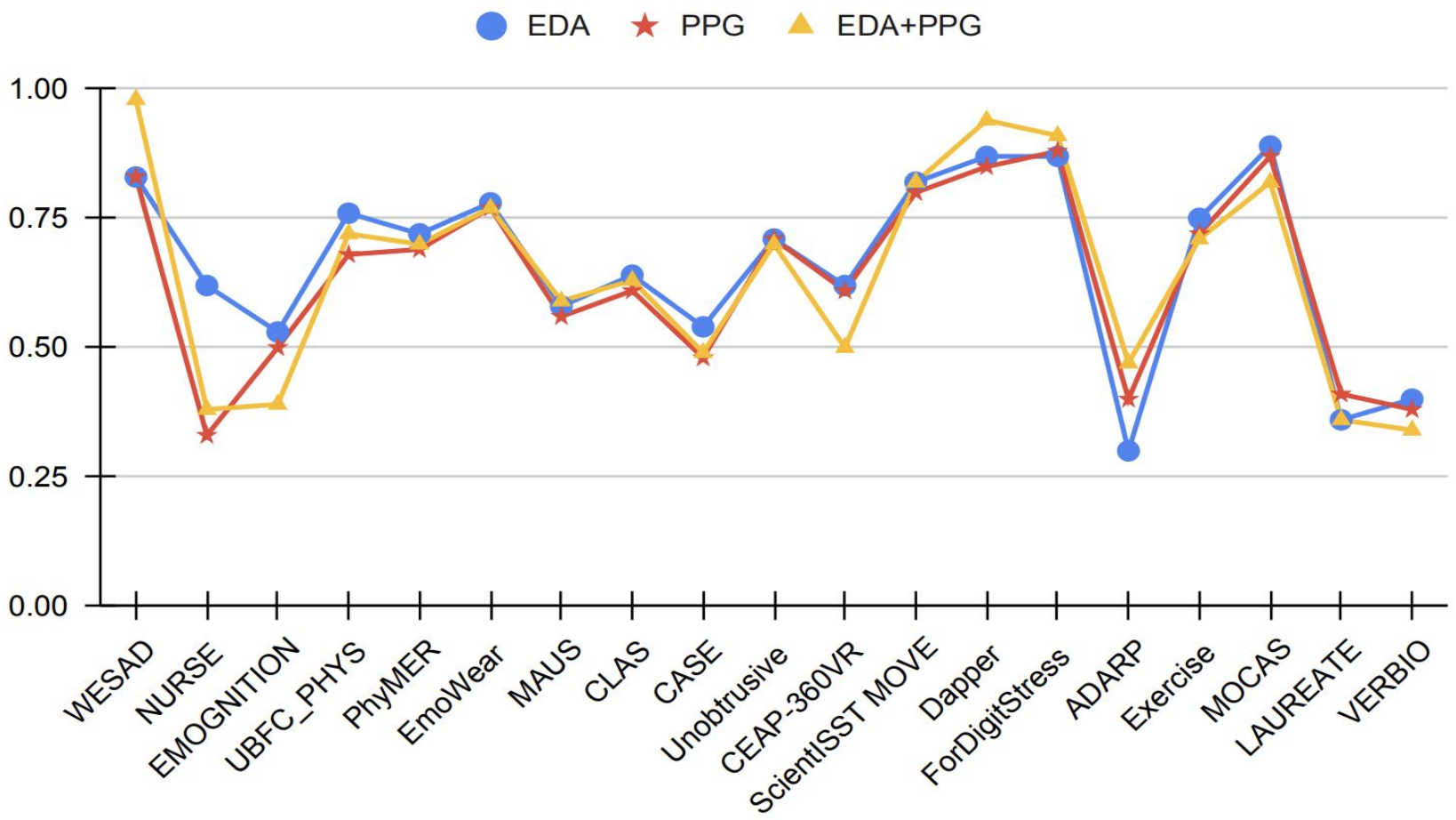
Arousal Classification



Arousal Classification

Rank	Dataset	EDA		PPG		Combined	
		F1	Best Model	F1	Best Model	F1	Best Model
1	ForDigitStress	0.94	CLSP MLP 25%	0.99	RF	0.99	RF
2	WESAD	0.83	Signal ResNet	0.80	HC-MLP	0.91	HC-MLP
3	Unobtrusive	0.88	RF	0.87	RF	0.86	CLSP MLP 50%
4	ScientISST MOVE	0.77	HC Attention MLP	0.81	RF	0.88	HC-MLP
5	ADARP	0.83	CLSP MLP 25%	0.80	CLSP CNN 50%	0.62	CLSP MLP 25%
6	MAUS	0.83	Signal ResNet	0.82	RF	0.82	RF
7	VERBIO	0.83	CLSP CNN 50%	0.77	CLSP CNN 50%	0.72	CLSP CNN 50%
8	LAUREATE	0.69	RF	0.77	CLSP MLP 50%	0.82	CLSP Zero-Shot
9	Dapper	0.77	CLSP CNN 50%	0.70	CLSP CNN 5%	0.81	CLSP MLP 5%
10	CLAS	0.69	RF	0.66	RF	0.70	RF
11	EMOGNITION	0.68	CLSP MLP 5%	0.62	CLSP MLP 50%	0.57	CLSP CNN 5%
12	MOCAS	0.65	CLSP CNN 5%	0.62	CLSP MLP 5%	0.63	CLSP MLP 25%
13	EmoWear	0.64	RF	0.64	RF	0.67	CLSP MLP 50%
14	Exercise	0.63	CLSP Zero-Shot	0.57	CLSP CNN 25%	0.54	CLSP MLP 5%
15	NURSE	0.62	CLSP CNN 5%	0.52	CLSP MLP 50%	0.62	CLSP CNN 5%
16	CEAP-360VR	0.56	CLSP CNN 5%	0.43	CLSP CNN 5%	0.45	CLSP MLP 5%
17	PhyMER	0.51	CLSP Zero-Shot	0.42	LDA	0.42	LDA
18	CASE	0.47	Signal CNN+Transformer	0.30	HC-MLP	0.40	CLSP MLP 5%
19	UBFCPHYS	0.45	CLSP MLP 5%	0.34	CLSP CNN 25%	0.41	CLSP MLP 5%

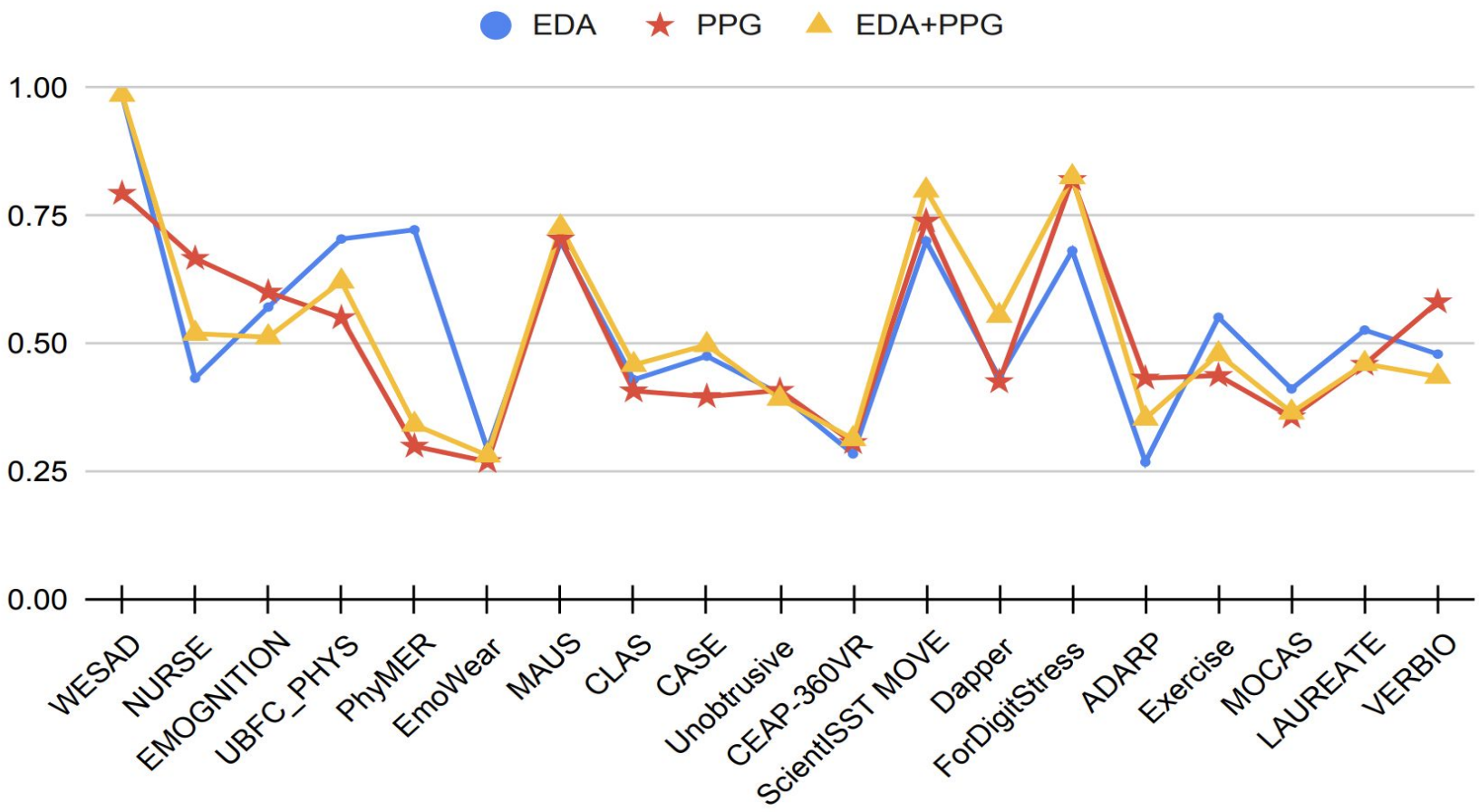
Valence Classification



Valence Classification

Rank	Dataset	EDA		PPG		Combined	
		F1	Best Model	F1	Best Model	F1	Best Model
1	WESAD	0.83	CLSP CNN 50%	0.83	CLSP CNN 50%	0.98	HC-MLP
2	Dapper	0.87	CLSP CNN 50%	0.85	CLSP CNN 50%	0.94	CLSP CNN 50%
3	ForDigitStress	0.87	CLSP CNN 5%	0.92	RF	0.92	RF
4	MOCAS	0.89	CLSP Zero-Shot	0.87	CLSP CNN 50%	0.82	CLSP CNN 25%
5	ScientISST MOVE	0.82	CLSP MLP 50%	0.80	CLSP CNN 50%	0.82	CLSP CNN 50%
6	EmoWear	0.78	CLSP CNN 50%	0.77	CLSP CNN 50%	0.77	RF
7	UBFCPHYS	0.76	RF	0.68	LDA	0.72	RF
8	Exercise	0.75	CLSP Zero-Shot	0.72	CLSP CNN 50%	0.71	CLSP MLP 50%
9	PhyMER	0.72	CLSP Zero-Shot	0.69	CLSP CNN 50%	0.70	CLSP MLP 50%
10	Unobtrusive	0.71	CLSP Zero-Shot	0.71	RF	0.70	CLSP CNN 25%
11	CLAS	0.64	CLSP Zero-Shot	0.61	CLSP CNN 25%	0.63	HC Attention MLP
12	NURSE	0.62	CLSP Zero-Shot	0.39	CLSP CNN 5%	0.38	CLSP Zero-Shot
13	CEAP-360VR	0.62	CLSP CNN 5%	0.61	CLSP CNN 5%	0.50	LDA
14	MAUS	0.58	HC-MLP	0.56	LDA	0.59	LDA
15	CASE	0.54	CLSP MLP 5%	0.48	LDA	0.49	LDA
16	EMOGNITION	0.53	CLSP Zero-Shot	0.50	CLSP MLP 5%	0.39	CLSP CNN 5%
17	LAUREATE	0.36	HC-MLP	0.41	HC-MLP	0.40	CLSP MLP 50%
18	VERBIO	0.40	HC-MLP	0.38	HC-MLP	0.34	CLSP MLP 5%
19	ADARP	0.30	CLSP Zero-Shot	0.40	CLSP Zero-Shot	0.47	HC-MLP

Four-Class Classification



Four-Class Classification

Rank	Dataset	EDA		PPG		Combined	
		F1	Best Model	F1	Best Model	F1	Best Model
1	WESAD	0.987	RF	0.794	RF	0.987	LDA
2	ForDigitStress	0.682	LDA	0.821	RF	0.826	RF
3	ScientISST MOVE	0.701	CLSP MLP 25%	0.740	CLSP CNN 50%	0.800	CLSP CNN 50%
4	MAUS	0.700	HC-MLP	0.705	RF	0.728	RF
5	PhyMER	0.723	CLSP CNN 50%	0.300	RF	0.342	RF
6	UBFCPHYS	0.705	CLSP Zero-Shot	0.551	LDA	0.622	LDA
7	MOCAS	0.701	CLSP MLP 25%	0.357	RF	0.366	RF
8	EMOGNITION	0.572	RF	0.601	CLSP CNN 50%	0.513	RF
9	Dapper	0.434	RF	0.426	RF	0.555	RF
10	Exercise	0.552	CLSP CNN 25%	0.438	HC-MLP	0.480	RF
11	LAUREATE	0.527	CLSP MLP 5%	0.460	RF	0.461	RF
12	CASE	0.476	RF	0.397	RF	0.498	RF
13	VERBIO	0.480	CLSP Zero-Shot	0.582	CLSP Zero-Shot	0.436	CLSP Zero-Shot
14	NURSE	0.433	CLSP Zero-Shot	0.667	CLSP Zero-Shot	0.520	CLSP Zero-Shot
15	CLAS	0.430	RF	0.408	HC-MLP	0.459	RF
16	Unobtrusive	0.402	RF	0.409	CLSP Zero-Shot	0.393	HC-MLP
17	ADARP	0.269	CLSP Zero-Shot	0.433	CLSP Zero-Shot	0.354	CLSP Zero-Shot
18	CEAP-360VR	0.285	CLSP MLP 25%	0.307	RF	0.314	RF
19	EmoWear	0.293	CLSP CNN 50%	0.270	HC-MLP	0.282	HC-MLP

Key Insights

- CLSP models are top performer in 88/171 evaluations.
- RF & LDA are strong, fast, interpretable baselines with 59/171 best results.
- Handcrafted features outperform raw-signal DL models.
- Overall performance was sensitive to dataset quality, including factors such as signal integrity, labeling reliability, and experimental design.
- Similar performance trends are observed across modalities

Cross-Dataset Insights

- Cross data analysis revealed that **labelling strategies** have overall **high transferability** with each other as compared to device, experiment settings and demographics.
- **Inter-gender and age transfer** showed **strong generalization for valence**, but near-random performance for arousal, suggesting demographic based physiological differences in arousal responses.
- Overall, the high cross-domain transferability across labeling strategies, settings, and devices **indicates the feasibility of combining small datasets** to train more generalizable large-scale models.

Read the paper for more details...

Connect: pragyas@iiitd.ac.in