FEEL: Modeling Physiological Diversity for Robust Emotion Recognition

Emotion recognition from physiological signals shows strong potential for mental health support and emotion-aware technologies. Yet progress is limited by the lack of standardized, large-scale evaluations across diverse datasets, restricting comparability and generalization. Advancing the field requires treating data as a *shared resource* and harmonizing signal representations and labeling strategies. This would enable large-scale training, domain adaptation, and fair, reproducible benchmarks. Without such coordination, research remains fragmented and findings fail to generalize. Notably, no systematic benchmark yet exists for evaluating models on widely used physiological signals such as EDA and PPG from wearable devices. To address this gap, we present FEEL, the **first benchmarking effort** that systematically evaluates emotion recognition using electrodermal activity (EDA) and photoplethysmography (PPG) data from **19 publicly available datasets**. Our study assesses **16 models**, covering traditional machine learning methods, deep neural networks with features or with raw signal segments, and self-supervised language-based pretraining approaches, organized into four main modeling paradigms. Evaluations were conducted both within individual datasets and across datasets, examining how well models generalize under variations in experimental setups (lab, constrained, and real-life settings), device types (lab–grade, custom wearables, and commercial devices), and annotation strategies (expert-annotation, participant self-reports, and stimulus-based labelling). Our **findings** indicate that fine-tuned contrastive signal-language pretraining models achieve the strongest overall F1 scores (73/114) for arousal and valence classification tasks. Nonetheless, simpler approaches such as Random Forests, LDA, and MLP remain competitive (36/114). Models that incorporate features (109/114) outperform those relying solely on raw signal inputs, highlighting the importance of domain knowledge. Cross-group analyses further show that models trained on real-world data transfer effectively to laboratory (F1 = 0.79) and constrained settings (F1 = 0.78). Likewise, models trained on expert-annotated datasets perform well on stimulus-based (F1 = 0.72) and self-reported labels (F1 = 0.76). Device heterogeneity also plays a role: models trained on lab-grade devices generalized successfully to both custom wearables (F1 = 0.81) and commercial sensors such as Empatica E4 (F1 = 0.73). In sum, FEEL establishes a unified benchmarking framework for physiological emotion recognition and offers practical insights for building robust, generalizable emotion-aware systems.

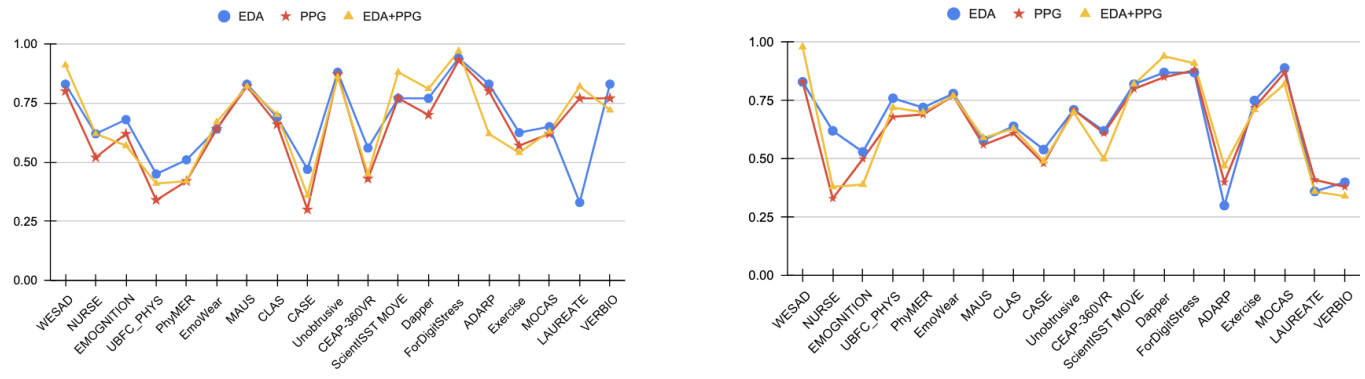

Figure 1: Comparative performance (F1 score) of the best-performing models per dataset across three physiological modalities (EDA, PPG, EDA+PPG) for emotion recognition.