

Supplementary Information

Supervised Information Gain-Based Feature Selection for Multimodal Physiological Signals in Stress Prediction

Shefali Arora et al.

Appendix A1. Additional Experimental Details without LOSO evaluation

Table S1 reports non-generalizable results achieved using Random Forest classifier for different feature selection methods applied on the STFT features extracted using WESAD dataset without LOSO protocol, using a train-test split of 80:20. It is observed that Information Gain(IG) achieved the highest performance, reaching an accuracy of 99.66% with Top-50 features. Pearson correlation also demonstrated strong discriminative power, yielding 98.52% accuracy at Top-50 features. In contrast, the F-test performed poorly with very few features i.e 72.44% for Top-10 features but improved significantly as more features were included, achieving an accuracy of 97.61% at Top-50. The Chi-Square method provided balanced performance across all feature sizes, with accuracy improving from 87.70% (Top-10) to 95.33% (Top-50). The Multiset Canonical Correlation Analysis(MCCA) technique combined with IG gives an accuracy of 90% for Top-50 features. For all CCA blocks, the number of canonical components was fixed at five. The pairwise projected components (from ACC-BVP, BVP-TEMP, TEMP-EDA, and EDA-ACC) were concatenated to form a unified fused feature matrix. The solver used for SVD-based and no penalty was applied. Before applying CCA, each modality was standardized using z-score normalization to ensure equal scaling in the canonical space. Overall, IG emerged as the most effective feature selection strategies for stress prediction using STFT-based physiological features. Further, Table S2 compares the results obtained using proposed approach with popular feature selection methods combined with Random Forest classifier for top 50 selected features. IG proves to be the most efficient technique in finding discriminative features with respect to class labels.

Table S1: Performance comparison of different correlation-based feature selection methods on the WESAD STFT feature set using a Random Forest classifier. Metrics are reported for Top- K selected features.

Method	Top- K	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Information Gain	10	88.50	88.73	88.50	88.54
	20	87.36	87.57	87.36	87.40
	50	99.66	99.66	99.66	99.66
ANOVA F-test	10	72.44	70.32	72.44	70.59
	20	92.60	92.61	92.60	92.54
	50	97.61	97.64	97.61	97.58
Chi-Square (χ^2)	10	87.70	87.74	87.70	87.71
	20	90.77	90.90	90.77	90.81
	50	95.33	95.40	95.33	95.30
Pearson Correlation	10	93.39	93.47	93.39	93.42
	20	94.99	95.03	94.99	94.99
	50	98.52	98.52	98.52	98.51
CCA+IG	10	85.39	85.47	85.39	85.42
	20	87.99	88.03	87.99	87.99
	50	90	90	90	90

Table S2: Performance Comparison of Popular Feature Selection Methods without LOSO (Selected Features = 50)

Feature Selection Method	Accuracy	F1	Precision	Recall
Information Gain	0.99	0.98	0.99	0.98
Chi-Square Test	0.96	0.95	0.96	0.95
ANOVA F-test	0.98	0.97	0.98	0.97
Pearson Correlation	0.98	0.97	0.98	0.96
Variance Threshold	0.92	0.90	0.93	0.87
LASSO (L1 Regularization)	0.92	0.89	0.92	0.87
Recursive Feature Elimination (RFE)	0.98	0.97	0.98	0.97