

Supplementary Materials for: Multimodal Classification of Social Anxiety Using Continuous Self-Rating and Physiological Data in Virtual Reality

Marco Pardini, Sergio Frumento, Matteo Martini, Martina Alaimo, Gianluca Rho, Noemi Paparo,
Mario G. C. A. Cimino, Enzo Pasquale Scilingo, Danilo Menicucci, Manuela Chessa, Alberto Greco

I. EMBODIMENT AND VIRTUAL SCENARIO VALIDATION

To assess the level of embodiment and immersion experienced by the participants in the virtual scenario, we utilized the Igroup Presence Questionnaire (IPQ [1]). The IPQ provides a collective score (Global presence) and subscores (Spatial presence, Involvement, and Experienced realism). As shown in Supplementary Table I, all scores fell robustly above the acceptability threshold, suggesting that the level of embodiment was satisfactory.

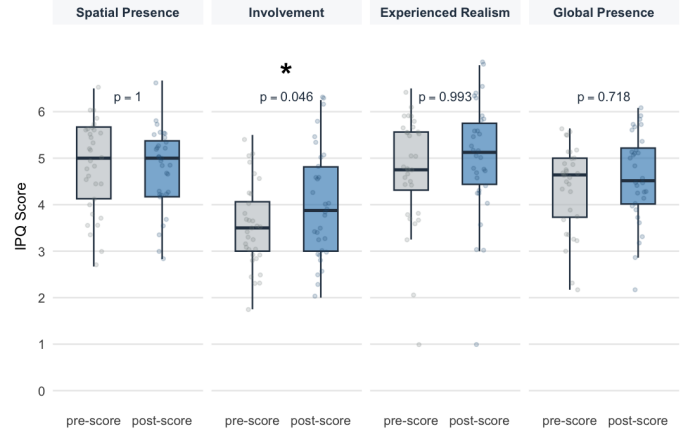
Furthermore, the comparison between the IPQ administered right after the first explorative 1-minute immersion (in a totally empty version of the room, “pre-score” in Supplementary Table I) and right after the full 10-minutes scene (with the progressive filling of interacting people, “post-score” in Supplementary Table I) revealed negligible variations attributable to the mere difference in timings (1 and 10 minutes respectively), with the only exception of the Involvement subscale. Indeed, all subscales were tested for differences using paired Wilcoxon signed-rank tests to which was applied a Bonferroni correction (Supplementary Figure 1): for the Involvement subscale only, the comparison between pre- and post-score shows that the latter increased significantly (albeit marginally: $p = 0.046$), presumably as an effect of the experimental task and of the interactions engaged with the NPCs.

Supplementary Table I
IPQ SCORES BEFORE AND AFTER THE SESSION ($N = 63$)

Subscale	pre		post		ΔM	p
	M	SD	M	SD		
Spatial presence	4.82	0.98	4.77	0.86	-0.05	1.000
Involvement	3.52	0.91	4.01	1.20	+0.49	0.046*
Exp. realism	4.74	1.20	5.00	1.25	+0.26	0.993
Global presence	4.36	0.92	4.59	0.96	+0.23	0.718

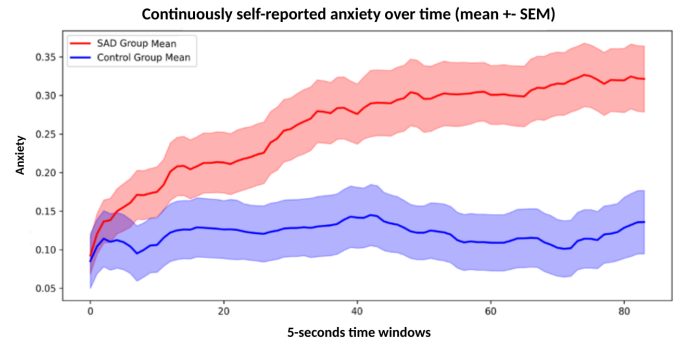
Note: M = mean; SD = Standard Deviation *Significant at $\alpha < 0.05$.

The diagnostic construct of Social Anxiety Disorder (SAD) – and the Liebowitz Social Anxiety Scale (L-SAS) – distinguishes between anxiety induced by performance situations and anxiety elicited by commonplace social interactions [2], [3]: most of the previous applications of VR to SAD focused on performance-related scenarios, whereas our study focused on a more common situation (a waiting room). To empirically demonstrate that our waiting room scenario successfully induced social anxiety, we compared the social anxiety continuously self-reported by socially-anxious and healthy volunteers.



Supplementary Figure 1. Comparison of the IPQ pre-scores and post-scores (referring to the mean values reported in Supplementary Table I). Gray and blue points represent individual data points, overlaid to show the sample dispersion and distribution density. Above each comparison, the exact p-value adjusted via Bonferroni correction is reported.

Socially-anxious participants reported higher levels of anxiety ($mean = 0.2613$, $std = 0.2250$) compared to healthy controls ($mean = 0.1204$, $std = 0.1850$): this difference was statistically significant (Mann-Whitney: $U = 705.0$, $p = 0.0018$). Comparing these two groups for what concerns further anxiety dynamics – such as the standard deviation of the joystick usage ($p = 0.0111$) and the mean trend ($p < 0.0001$) – further confirmed significantly-distinct patterns (Supplementary Figure 2).



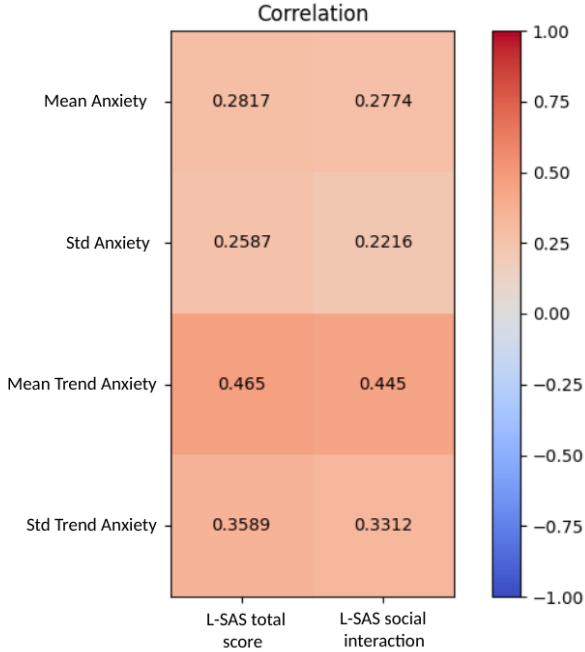
Supplementary Figure 2. Significant differences between the social anxiety continuously self-reported during the experimental session by the socially-anxious participants with respect to healthy controls.

II. CONTINUOUS SELF-RATING AND L-SAS CORRELATION

To validate the relevance of our continuous measurement approach, we correlated the features extracted from the anxiety self-ratings with the punctual L-SAS scores. As reported in Supplementary Table II and visually represented in Supplementary Figure 3, the continuous self-ratings of anxiety are positively and significantly correlated with both the global L-SAS score and the social-interaction subscale.

TABLE II
SUPPLEMENTARY TABLE II: PEARSON CORRELATION COEFFICIENTS (r)
BETWEEN ANXIETY FEATURES AND L-SAS.

Feature	L-SAS (Global)	L-SAS (Social Interaction)
mean_anxiety	0.2817	0.2774
std_anxiety	0.2587	0.2216
mean_trend_anxiety	0.4650	0.4450
std_trend_anxiety	0.3589	0.3312



Supplementary Figure 3. Relationship between the continuous self-rating of social anxiety and the L-SAS global and sub-scale scores.

Even if the punctual (i.e., L-SAS global score and social-interaction subscale) and the continuous self-report of anxiety are significantly and positively correlated (see Supplementary Figure II), the latter ones come with important added value with respect to the former ones. Indeed, behavioral ratings:

- allow a moment-by-moment characterization of social anxiety, which make them less prone than L-SAS to recall biases ([4], [5]) and to the so-called “peak-end rule” ([6], [7]) – a well-known bias consisting in an overstatement of the last and most intense moment in retrospectively evaluating a past experience (e.g., rating a movie based on its final plot-twist).);

- can be more sensitive than subjective reports to clinical improvements. For example, exposure therapies can induce healthy-like patterns in behavioral or physiological correlates which do not reflect on changes in questionnaire’s scorings (e.g., [8], [9], [10]), coherently with the observation that a confrontation with the feared stimulus “is necessary for changes in self-reported fear to occur” [11];
- allow to map temporal dynamics of anxiety and correlate them with (equally continuous) psychophysiological signals, paving the way for a more complete characterization of the psychopathological symptoms assessed [12] in a way that would not be possible through retrospective scales such as the L-SAS [5].

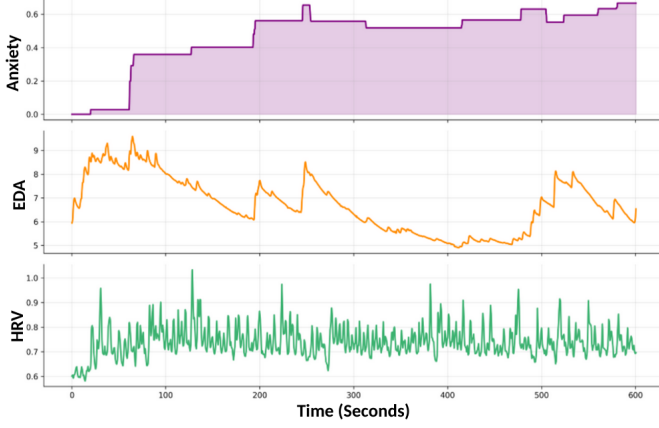
This added value also applies to a cost-benefits ratio. Indeed, the comparison of costs is undoubtably in favour of L-SAS, which needs a cheaper support (in its original version, just paper and pencil); while VR headsets are becoming more and more affordable for the average therapist, their use in the clinical practice still represent a niche. However, the benefits of behavioral ratings over self-reports counterbalance their costs: questionnaires can serve as broad tools to quickly screen large samples of whom one wants to identify the extremes, but are less sensitive to smaller – and nevertheless relevant – differences. The higher discriminative potential of virtual scenarios resembling everyday-life situations could reveal previously overlooked facets of Social Anxiety Disorder, a mental issue labeled the “neglected anxiety disorder” [13] as considered to occur only in circumscribed performance situations such as public speaking [14]. This higher sensitivity could lead to a wider awareness of SAD symptoms, improved diagnosis, increased sensitivity to refinements of therapeutic protocols, and other advantages with remarkable impacts on the sanitary systems and on the wellbeing of general population.

III. MULTIMODAL SIGNAL EXAMPLE

To better illustrate the synchronized nature of our multimodal data collection, Supplementary Figure 4 presents an example of the recorded signals from a single socially anxious subject during the VR exposure. The figure displays the continuous joystick annotations alongside the corresponding physiological measurements, specifically the electrodermal activity (EDA) and Heart Rate Variability (HRV). This visualizes how the conscious subjective appraisal of anxiety fluctuates in tandem with the involuntary autonomic arousal during the progressive filling of the virtual waiting room.

IV. BENCHMARKING OF MACHINE LEARNING ARCHITECTURES AND DATASETS

During the preliminary phase of the study, multiple candidate architectures were systematically evaluated to identify the optimal modeling approach. We conducted a fair benchmarking setup utilizing identical nested Leave-One-Subject-Out (LOSO) cross-validation, hyperparameter grid search, and forward feature selection across four distinct models: a Multi-Layer Perceptron (MLP), a 1D Convolutional Neural Network



Supplementary Figure 4. Example of the signals recorded from a socially anxious participant during the experiment (subjective anxiety, EDA and HRV).

TABLE III
PERFORMANCE METRICS OF CANDIDATE ARCHITECTURES ACROSS DIFFERENT FEATURE SETS.

Model	Feature Set	Macro F1	Accuracy
CNN	anxiety	0.745	0.761
CNN	combined	0.745489	0.761
CNN	physio	0.338	0.460
ENCODER	anxiety	0.790	0.809
ENCODER	combined	0.730	0.746
ENCODER	physio	0.790	0.793
GRU	anxiety	0.678	0.714
GRU	combined	0.720	0.746
GRU	physio	0.511	0.555
MLP	anxiety	0.363	0.571
MLP	combined	0.370	0.587
MLP	physio	0.505	0.603

(1D-CNN), a Gated Recurrent Unit (GRU), and a Transformer Encoder.

An equally important aspect of our evaluation was determining the optimal level of data preprocessing. Our analysis demonstrated the overall superiority of using the engineered Feature Time-Series dataset over the raw Processed Signal dataset. Across our evaluations, models trained on the feature dataset achieved an average macro F1-score of 0.786, whereas those trained on the preprocessed raw dataset achieved an average macro F1-score of only 0.759 (Supplementary Figure 5). This highlights that extracting targeted statistical and morphological features successfully filters noise and provides more robust temporal patterns for the classifiers.

To systematically compare the performance across the four evaluated architectures, we computed their macro F1-scores and overall accuracies. Table III illustrates these metrics evaluated across three input configurations: anxiety only, physiological only, and combined modalities (as also seen in Supplementary Figures 6 and 7).

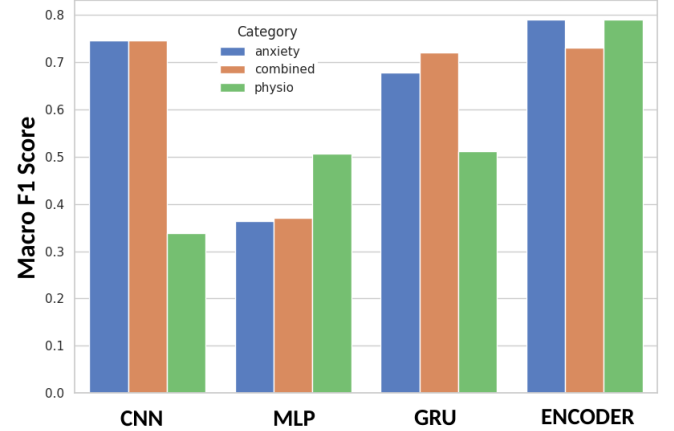
To systematically compare the performance across the four evaluated architectures, we computed their macro F1-scores and overall accuracies. Supplementary Figures 6 and 7 illustrate these metrics evaluated across three input configurations:



Supplementary Figure 5. Macro F1-score comparison between datasets consisting of either raw data (“raw”) or pre-processed data from which features were extracted (“feature”).

anxiety only, physiological only, and combined modalities.

Model performance on Feature dataset by Category (F1 score)

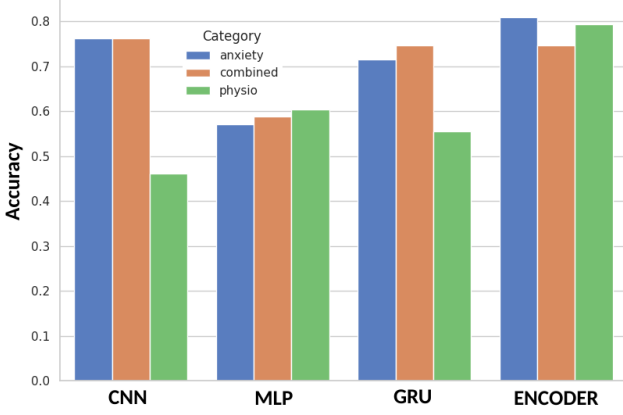


Supplementary Figure 6. Macro F1-score comparison among the four tested architectures (MLP, 1D-CNN, GRU, Transformer Encoder) across different input modalities.

V. EARLY VS. LATE FUSION INTEGRATION STRATEGIES

When integrating the continuous self-reports of anxiety with the physiological markers in the Feature Time-Series dataset, we systematically compared two multimodal integration architectures: Early Fusion and Late Fusion. In the Early Fusion setup, physiological and subjective features were directly concatenated into a single input sequence before being passed through the shared self-attention layers of the classification model. Conversely, the Late Fusion architecture utilized separate branches dedicated exclusively to physiological features and anxiety features. This allowed the model to extract high-level contextual embeddings from each modality without interference, which were subsequently concatenated and passed to a final classification head.

Model Performance on Feature Dataset by Category (Accuracy)



Supplementary Figure 7. Overall accuracy comparison among the four tested architectures across different input modalities.

As detailed in Supplementary Table IV, the Late Fusion approach significantly outperformed the Early Fusion strategy. Forcing highly distinct data types through a shared learning mechanism in the Early Fusion configuration hindered the model’s ability to efficiently identify modality-specific patterns, yielding an overall accuracy of 0.75. In contrast, the Late Fusion approach successfully harnessed cross-modal synergies, achieving an overall accuracy of 0.83 and significantly improving the classification robustness, particularly in the ability to identify non-anxious controls (Non-Anxious F1 rising from 0.67 to 0.78).

TABLE IV
SUPPLEMENTARY TABLE III: FEATURE DATASET CLASSIFICATION
COMPARING EARLY FUSION AND LATE FUSION ARCHITECTURES.

Architecture	Non-Anxious (Class 0)			Anxious (Class 1)			Accuracy
	P	R	F1	P	R	F1	
Early Fusion	0.73	0.62	0.67	0.76	0.84	0.79	0.75
Late Fusion	0.80	0.77	0.78	0.84	0.86	0.85	0.83

REFERENCES

- [1] M. Melo, G. Gonçalves, M. Bessa *et al.*, “How much presence is enough? qualitative scales for interpreting the igroup presence questionnaire score,” *IEEE Access*, vol. 11, pp. 24 675–24 685, 2023.
- [2] A. Diagnostic, “Statistical manual of mental disorders: Dsm-5 (ed.) washington,” *DC: American Psychiatric Association*, 2013.
- [3] M. R. Liebowitz, “Liebowitz social anxiety scale,” *Journal of Anxiety Disorders*, 1987.
- [4] M. D. Robinson and G. L. Clore, “Belief and feeling: evidence for an accessibility model of emotional self-report,” *Psychological bulletin*, vol. 128, no. 6, p. 934, 2002.
- [5] S. Shiffman, A. A. Stone, and M. R. Hufford, “Ecological momentary assessment,” *Annu. Rev. Clin. Psychol.*, vol. 4, no. 1, pp. 1–32, 2008.
- [6] B. Alaybek, R. S. Dalal, S. Fyffe, J. A. Aitken, Y. Zhou, X. Qu, A. Roman, and J. I. Baines, “All’s well that ends (and peaks) well? a meta-analysis of the peak-end rule and duration neglect,” *Organizational Behavior and Human Decision Processes*, vol. 170, p. 104149, 2022.
- [7] B. L. Fredrickson and D. Kahneman, “Duration neglect in retrospective evaluations of affective episodes,” *Journal of personality and social psychology*, vol. 65, no. 1, p. 45, 1993.

- [8] J. Lipka, M. Hoffmann, W. H. Miltner, and T. Straube, “Effects of cognitive-behavioral therapy on brain responses to subliminal and supraliminal threat and their functional significance in specific phobia,” *Biological psychiatry*, vol. 76, no. 11, pp. 869–877, 2014.
- [9] P. Siegel and K. A. Gallagher, “Delaying in vivo exposure to a tarantula with very brief exposure to phobic stimuli,” *Journal of behavior therapy and experimental psychiatry*, vol. 46, pp. 182–188, 2015.
- [10] K. Schmack, J. Burk, J.-D. Haynes, and P. Sterzer, “Predicting subjective affective salience from cortical responses to invisible object stimuli,” *Cerebral Cortex*, vol. 26, no. 8, pp. 3453–3460, 2016.
- [11] J. Peters, A. I. Filmer, J. B. van Doorn, V. N. Metselaar, R. M. Visser, and M. Kindt, “Re-encountering the phobic cue within days after a reconsolidation intervention is crucial to observe a lasting fear reduction in spider phobia,” *Molecular Psychiatry*, vol. 30, no. 6, pp. 2729–2738, 2025.
- [12] S. Frumento, A. Iannizzotto, A. Greco, E. P. Scilingo, A. Gemignani, D. Menicucci *et al.*, “Development of a behavioral avoidance test in virtual reality (vr-bat),” in *MetroXRaine*, 2023, pp. 949–953.
- [13] M. R. Liebowitz, J. M. Gorman, A. J. Fyer, and D. F. Klein, “Social phobia: Review of a neglected anxiety disorder,” *Archives of general psychiatry*, vol. 42, no. 7, pp. 729–736, 1985.
- [14] D. S. Mennin, D. M. Fresco, R. G. Heimberg, F. R. Schneier, S. O. Davies, and M. R. Liebowitz, “Screening for social anxiety disorder in the clinical setting: using the liebowitz social anxiety scale,” *Journal of anxiety disorders*, vol. 16, no. 6, pp. 661–673, 2002.