

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

Marco Pardini¹, Sergio Frumento¹, Matteo Martini², Martina Alaimo³, Gianluca Rho¹, Noemi Paparo¹, Mario G. C. A. Cimino¹, Enzo Pasquale Scilingo¹, Danilo Menicucci³, Manuela Chessa², Alberto Greco¹

¹Department of Information Engineering, University of Pisa, Pisa, Italy

²Department of Informatics, Bioengineering, Robotics, and Systems Engineering, University of Genoa, Genoa, Italy

³Department of Surgical, Medical, Molecular and Critical area pathology, University of Pisa, Pisa, Italy

Assessment of Social Anxiety Disorder (SAD) is limited by paradigms that often target performance-related fears and its reliance on static self-reports that fail to capture the dynamic, in-the-moment nature of anxiety. This study introduces a novel Virtual Reality (VR) framework to assess SAD that integrates continuous self-ratings of anxiety with objective physiological data. Sixty-three participants were immersed in an ecologically valid VR waiting room scenario. They continuously rated their anxiety levels in real-time using a joystick, providing a direct subjective-behavioral measure, while their electrocardiogram and electrodermal activity were simultaneously recorded. We systematically compared three machine learning pipelines — a static Support Vector Machine, an end-to-end Transformer Encoder, and a Transformer Encoder on engineered features — to classify participants into the high and the low anxiety groups previously determined through the Liebowitz Social Anxiety Scale (L-SAS). To effectively harness psychophysiological and subjective data and overcome integration challenges, we developed a Late Fusion Transformer framework. This optimized architecture yielded the most robust and balanced classification (F1-score = 0.853 for the anxious class; 82.5% overall accuracy). The results show a synergistic benefit from combining modalities: while feature selection identified continuous self-rated anxiety as the most powerful individual predictor, the late integration of physiological markers provided a significant complementary contribution that enhanced the model's overall accuracy, particularly in identifying non-anxious controls. In conclusion, our framework provides a dynamic assessment of SAD that captures the moment-to-moment experience of anxiety. By integrating a continuous self-rating of social anxiety with objective physiological data, this approach can identify clinically relevant incongruencies between a person's retrospective self-perception and their in-the-moment reactivity, paving the way for a new era of objective, personalized, and temporally sensitive psychometrics.

Index Terms—Social Anxiety Disorder, Virtual Reality, Deep learning, Continuous Self-rating, Psychophysiological Signals,

I. INTRODUCTION

Social Anxiety Disorder (SAD) is a psychopathology characterized by an excessive fear of social situations, particularly those involving potential scrutiny by others [1]. Its prevalence has increased in recent years [2], yet its assessment still relies mainly on subjective measures such as clinical interviews and self-reported questionnaires. Instruments like the L-SAS [3] are widely used to quantify symptoms but cannot fully capture the complexity and variability of SAD. While the subjective nature of these tools provides insight into patients' perceptions, it also introduces bias, including intentional misreporting [4] and overestimation [5]. Moreover, questionnaires could fail to grasp the temporal fluctuations of anxiety (typically oversimplified into a punctual retrospective judgment) as well as its wide range of symptoms, varying for severity and for situational triggers. With regard to this last point, two SAD subtypes have been described [6]: one mainly related to performance situations (e.g., public speaking, job interviews) and another triggered by commonplace social

interactions (e.g., small talk in a waiting room). Standard assessments often provide only a single static score that does not specify which social context triggers the most anxiety.

To address these limitations, researchers have used Virtual Reality (VR) to create controlled yet realistic scenarios for SAD assessment. VR offers a standardized and safe way to expose individuals to feared situations, enabling the collection of objective behavioral and physiological measures integrating self-reports [7]. It can also reduce barriers for socially anxious individuals reluctant to participate in real-life evaluations involving direct scrutiny (such as clinical interviews). Immersing participants in virtual interactions allows clinicians to observe anxiety responses in real time within a controlled but ecologically valid setting.

However, most VR environments for SAD assessment have focused on the *performance* subtype, often simulating public-speaking scenarios [8], [9], [10]. Examples include delivering a speech to a virtual audience or interacting with morphing avatars [11]. While effective for performance anxiety, these paradigms overlook more common social fears. Public speaking is relatively rare and can often be avoided, whereas everyday interactions (e.g., sitting among strangers) are harder to evade and may better represent the daily burden of SAD. There is a need for VR assessments targeting the more prevalent interactional subtype [10].

Another limitation of existing VR-based assessments concerns the measurement of anxiety fluctuations. Many studies still rely on discrete pre- and post-exposure ratings [8] or infer

The research leading to these results has received partial funding from the Italian Ministry of Education and Research (MIUR) in the framework of projects ForeLab Project (Departments of Excellence), and has been supported by projects PRIN2022 BRAVE (funded by Italian Minister of University MUR with project code n. 2022PTSX4L) and PNRR—M4C2-Investimento 1.3, Partenariato Esteso PE00000013 — 'FAIR - Future Artificial Intelligence Research' - Spoke 1 'Human-centered AI', funded by the European Commission under the NextGeneration EU programme. No conflicts of interest are reported for this study.

Corresponding author: Alberto Greco (email: alberto.greco@unipi.it).

anxiety from scenario phases rather than direct reports [12]. Such methods miss the temporal dynamics of anxiety during exposure: retrospective ratings are influenced by memory and comparison processes [13], potentially leading to discrepancies with the actual moment-by-moment experience. This phenomenon, known as a response shift [14], reflects changes in the internal standards by which experiences are judged.

Continuous self-report of anxiety during VR exposure can capture these dynamics, revealing temporal patterns obscured by aggregate measures. Such data can be aligned with simultaneously recorded physiological signals, enabling fine-grained analyses of the co-evolution of subjective and autonomic responses under social stress.

A next-generation SAD assessment tool should meet three requirements. First, it should include scenarios beyond public speaking, focusing on routine social encounters that are harder to avoid, while minimizing confounding cognitive demands (e.g., speech preparation [15]). Second, it should capture the temporal dynamics of anxiety through continuous self-report, summarized into meaningful features that can be compared with physiological measures to assess coherence between subjective and physiological responses. Tracking variability over time may yield more sensitive assessments [16]. Third, outputs should be validated against established instruments such as questionnaires (e.g., L-SAS [3]): despite their limitations, these tools currently represent the most standardized and reproducible way to assess SAD symptoms, recommended in clinical guidelines [17]. Agreement with L-SAS classifications should be substantial but not necessarily perfect, as discrepancies may reveal important divergences between subjective reports and observed behaviors.

It is also interesting to disentangle the contribution of different data modalities. Comparing performance when using only self-rated anxiety, only physiological signals, or both provides information about the tool’s scalability: if physiological data do not substantially improve accuracy, a simpler setup with only a VR headset and controller may suffice.

In this study, we present a VR-based assessment paradigm meeting these criteria. We developed a waiting-room simulation representing an everyday social situation [18] in which participants continuously reported their anxiety level via a joystick-controlled visual slider while electrodermal activity (EDA) and electrocardiogram (ECG) signals were recorded. This design yields synchronized streams of continuously self-rated social anxiety and of its physiological correlates. We analyzed these multimodal signals using both traditional machine learning with support vector machines (SVM) and a deep learning approach based on a Transformer Encoder architecture applied to the time-series data. A rigorous leave-one-subject-out (LOSO) cross-validation was used to classify participants as socially anxious or non-anxious based on VR session data, with comparisons to their L-SAS-derived labels. Feature selection analyses identified the most discriminative data components.

Previous studies have explored automated classification of social anxiety using behavioral, physiological, or questionnaire-based data [9], [19], [12], [20], [21], but often rely on single modalities, post-hoc labeling, or low-ecological

validity tasks. Some use only offline questionnaire data [21], others audio from structured interviews or public speaking tasks [22], [20], and some neuroimaging data such as EEG or fMRI collected in artificial settings [23], [24]. Smartphone and wearable devices have been used to capture digital phenotypes in daily life [9], but without the controlled context needed to examine responses to specific social stressors. Few studies have implemented continuous physiological monitoring in immersive environments [19], [12], and these typically focus on performance-based scenarios or infer anxiety from task phases rather than directly capturing it through continuous self-report.

These approaches lack a direct link between real-time subjective experience and anxiety measurement, either because labels are retrospective or because the VR scenario lacks interactivity and naturalism. Our paradigm addresses these gaps through an ecologically valid VR scenario integrating continuous self-ratings of anxiety with their physiological correlates, allowing a temporally aligned multimodal classification. The modeling approach combines interpretable machine learning with feature selection and end-to-end deep learning, providing complementary perspectives on the data.

By meeting these criteria, our study introduces a framework designed to complement — not replace — traditional diagnostic tools, offering objective and temporally resolved insights into anxiety dynamics during naturalistic social interactions. The following sections describe the experimental protocol, VR scenario, preprocessing and analysis of continuous self-rating of anxiety and its correlates, machine learning pipelines, and classification results, highlighting implications for improved SAD assessment and personalized intervention strategies.

II. METHODS

A. Subject recruitment procedure and experimental protocol

After obtaining the approval from the local bioethical committee (protocol 0017548/2024 released on 02/12/2024), recruitment was carried online and through printed posters placed in University buildings. Participants were invited to fill a preliminary battery of questionnaires to check the presence of inclusion criteria — with the L-SAS [3] — and the absence of potentially-confounding psychopathological factors — with the Symptoms Check-List 90 revised (SCL-90-R [25]). As a result of this preliminary screening, a total of 63 volunteers aged between 21 and 35 years (39 of female gender, 23 of male gender and one non-binary person) were invited to participate in the experiment: based on the L-SAS score being below or above the conventional cutoff of 55 [3], participants were preliminary labeled as either “socially-anxious” or “controls”, resulting in 37 socially-anxious participants and 26 control participants. Indeed, even if a higher-than-55 L-SAS score cannot be considered as a diagnosis of SAD (as well as no self-reported questionnaires has diagnostic value), this assessment tool represents the most spread choice in the research practice [7], [8], [10] and is recommended as a reproducible and standardized way to assess SAD symptoms [17].

The experimental protocol was carried on as follows. Each participant was welcomed by an experimenter blind to their

L-SAS score, and was invited to read and sign the informed consent. Volunteers were then made to wear i) a device — Shimmer3 GSR+ unit — for the recording of electrodermal activity (EDA), ii) a device — Shimmer3 ECG unit — for the recording of electrocardiographic activity (ECG), and iii) a VR headset — Meta Quest 3 — wired via an USB-C cable to a Windows workstation running the Unity 3D engine.

After a careful check of signals' quality, participants were immersed in an introductory virtual scenario where they selected an avatar (male or female) that best matched their gender, and practiced with the interface for continuously reporting their social anxiety. This interface (see Fig. 1) consisted of green curved bars, appearing at the sides of the visual field at a customized distance (tunable in height using the joystick of the VR controller). Each volunteer was instructed to perform a simple task: use the VR controller to continuously rate their experienced anxiety (visually represented in real time by the green bars). After these initial steps, a 3-minute baseline period was observed. During this phase, a scenario consisting of an empty gray space with a "Relax" sign floating in mid-air was displayed. Afterwards, the main experimental scenario — a waiting room — was displayed for a total of 10 minutes. Initially, the room was kept empty for 1 minute. Then, it gradually filled with non-player characters (NPCs). The NPCs entered the room distanced 25 seconds circa one from the other, randomly choosing equidistant spots from those already occupied [26], and engaged in activities such as checking their smartphones, stretching, or yawning. After 7 minutes, the room was completely full. At that point, the NPCs began social interactions, including small talk about everyday-life topics and non-verbal behaviours (i.e., making eye contact with the participant), for 2 minutes. This process continued until the room was completely full, at which point the NPCs began social interactions, including small talks about everyday-life topics and non-verbal behaviours (i.e., making eye contact with the participant). After 10 minutes in the waiting room, the scenario faded away, marking the end of the experiment and the dismissal of the participant.

B. Virtual Reality scenario

The virtual environment (Fig. 1) was designed with the goal of simulating a neutral waiting room, minimizing possible associations with specific real-world settings. The setting featured three gray walls and a large window overlooking an urban environment. The window was slightly opaque, but still provided a view on a street with sidewalks, trees and parked cars. Some typical urban sounds coming from outside, like traffic and construction noise, were included. The room was furnished with 18 chairs arranged in two horse-shoe shapes, two tables with magazines in the middle of them, a couple of vending machines and some potted plants. The subject was always seated in the central chair of the side of the right group, ensuring an unobstructed view of the entire room.

NPCs were created using the Ready Player Me tool, ensuring a realistic amount of diversity in their traits as well as a balance between genders. They entered the room one at the time from a corridor placed on the opposite side of the subject,

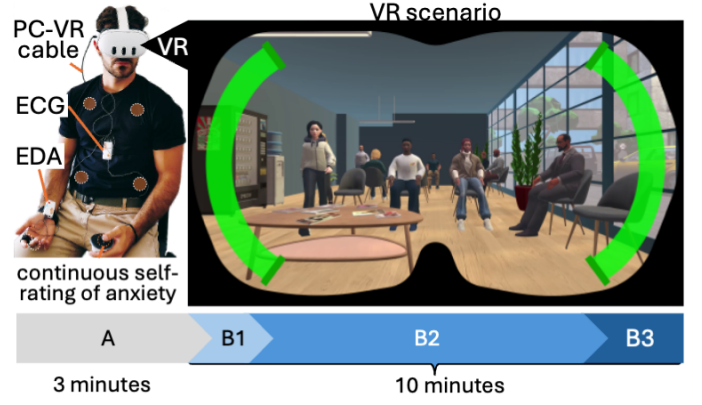


Fig. 1. graphical schematization of the experimental procedure, setting (black circles indicate electrodes' position), and virtual scenario (first person view). The timeline represents the procedure's steps and its timing: a baseline period (A) lasting 3 minutes was followed by 10-minutes experimental task composed by a 1-minute immersion in the empty room (B1), 7 minutes of NPCs spawning (B2) and 2 minutes of interacting NPCs (B3). PC = personal computer; VR = virtual reality; ECG = electrocardiography; EDA = electrodermal activity

and a chair is assigned to them. The chair assignment was done through a Bayesian model that simulates a natural selection dynamically updating chair-picking probabilities according to the current seating arrangement [26]. When a chair is assigned to an NPC, it moves towards it and sits down, starting to alternate neutral sitting poses to other individual actions, such as stretching, yawning, or checking their smartphones.

To enable a continuous and unobtrusive monitoring of the subjects' anxiety level during the VR exposure, we implemented a custom system, controllable through the joystick of the VR headset controllers. This system utilizes two green bars positioned at the edges of the subject's field of view, shaped to align to the lens profile of the HMD (Figure 1). By manipulating the stick, subjects could adjust the height of the bars to represent their current anxiety level on a scale from 0 to 1, marked by end-stop notches, with a fully filled bar indicating maximum anxiety and an empty one indicating none. The sliders were designed to be minimally intrusive, but at the same time easily locatable and unlikely disregardable. When not manipulated, their opacity is slightly reduced, to ensure a clearer view of the room.

C. Signal processing and feature extraction

1) Electrocardiogram and heart rate variability (HRV) signal processing

The ECG signal was processed using the Kubios HRV software [27], which implements a validated pipeline for the extraction of HRV indices. The software automatically detects R-peaks using a QRS detection algorithm based on adaptive thresholds, and calculates the series of RR intervals. Artifact correction was applied using the automatic algorithm implemented in the software, which combines thresholding and median filtering. To extract a uniformly sampled HRV, RR intervals were interpolated using cubic spline interpolation at a sampling frequency of 5 Hz. The HRV time-series was segmented into consecutive non-overlapping win-

dows of 5 seconds for time-domain analysis and 30 seconds for frequency-domain analysis. The following HRV features were computed within each window: mean of RR intervals (meanHRV), standard deviation of RR intervals (stdHRV), root mean square of successive RR differences (RMSSD), normalized spectral power in the low-frequency band (0.04–0.15 Hz, LFnu), normalized power in the high-frequency band (0.15–0.4 Hz, HFnu), and their ratio LF/HF).

2) Electrodermal Activity (EDA) signal processing

The EDA signal was decomposed into its tonic and phasic components using the *cvxEDA* algorithm, which is based on a convex optimization framework with Bayesian priors [28]. The tonic component (or skin conductance level, *SCL*) reflects slow changes in baseline conductance associated with general arousal levels, while the phasic component (or skin conductance response, *SCR*) captures rapid event-related conductance fluctuations [29]. Each SCR is the result of a neural burst in the sudomotor nerve activity (SMNA), which elicits the sweating from the glands under the skin surface of the hand. Prior to decomposition, EDA signals were normalized using a z-score transformation to facilitate convergence of the *cvxEDA* model. From the decomposed signals, we extracted the following features in 5-second non-overlapping windows: the maximum value of the phasic signal (SCR_{peak}), the number of SMNA bursts (SCR_{n}), the sum of SMNA peak amplitudes (Ampsum), the mean and standard deviation of the phasic component (SCR_{mean} , SCR_{std}), and the mean and standard deviation of the tonic component (SCL_{mean} , SCL_{std}). To further characterize sympathetic activity from EDA, we considered two frequency-domain features. First, the spectral power of the EDA signal in the 0.045–0.25 Hz band (EDAsymp) was computed using 30-second windows, following the method described in [30]. This feature provides a summary of sympathetic dynamics over relatively long temporal segments. In addition, we computed the TVSymp index, a time-varying sympathetic EDA descriptor based on the 0.08–0.24 Hz band [31]. Unlike EDAsymp , TVSymp captures the temporal evolution of sympathetic activity. This representation was used both to derive a subject-level summary index and to obtain temporally resolved inputs for the downstream classification pipelines.

3) Processing of the continuous report of anxiety

Participants continuously reported their perceived anxiety level during the Virtual Reality exposure by manipulating the joystick on one of the headset’s controllers. The resulting continuous signal was recorded and segmented into consecutive non-overlapping windows of 5 seconds. For each window, we computed four features: the mean and standard deviation of the raw signal (MeanAnxiety , StdAnxiety), and the mean and standard deviation of its first temporal derivative (MeanAnxietyRate , StdAnxietyRate). These features capture both the absolute level and the dynamic variability of the subjective anxiety report over time.

D. Dataset Construction for Temporal and Static Modeling

E. Dataset Construction for Temporal and Static Modeling

To explore different classification strategies, we constructed three datasets from the pre-processed physiological and self-reported signals.

- **Processed Signal Dataset** — This dataset includes 5 time-series signals: tonic and phasic components of electrodermal activity (*SCL* and *SCR*), the time-varying sympathetic index (*TVSymp*), the interpolated HRV signal, and the joystick-derived anxiety signal. All signals were resampled at 5 Hz and temporally aligned. To standardize the duration across subjects and minimize differences attributable to changes in the VR scenario, the first minute (prior to avatar spawning) and the final two minutes of the 10-minute session (when NPCs stopped spawning) were discarded. This resulted in a 420-second analysis window, yielding a final temporal length of 2100 time points.
- **Feature Time-Series Dataset** — This dataset contains the time-series of features computed over non-overlapping 5-second windows (84 windows per subject) from the 5 time series included in the *Processed Signal Dataset*. Accordingly, a set of 15 statistical and morphological features from the resampled physiological signals and from the continuously self-reported social anxiety were included.
- **Aggregated Feature Dataset** — This dataset was obtained by averaging the *Feature Time-Series Dataset* over time for each subject. Prior to this aggregation, a baseline correction was applied by subtracting the median value computed during the initial rest phase from each window feature. This resulted in a single feature vector per individual. In addition, global features requiring the 30-seconds time-window length for computation — such as frequency-domain indices (LFnu , HFnu , LF/HF , EDAsymp) — were included in this static representation.

To systematically evaluate the contribution of different data modalities, we conducted separate classification experiments for each of the three datasets using three distinct input configurations: (1) **Continuous anxiety self-rating only**, (2) **Physiological data only**, and (3) a **Combined** set of all available data. This resulted in a total of nine experimental runs. Table I summarizes the precise dimensionality for each of these conditions.

TABLE I
SUMMARY OF DATASET DIMENSIONS BY SIGNAL SOURCE.

Dataset Name	Dimensions (Subject \times Feature \times Time)		
	Self-ratings	Physiological	Combined
Processed Signal	$63 \times 1 \times 2100$	$63 \times 4 \times 2100$	$63 \times 5 \times 2100$
Feature Time-Series	$63 \times 4 \times 84$	$63 \times 11 \times 84$	$63 \times 15 \times 84$
Aggregated Feature	$63 \times 4 \times 1$	$63 \times 15 \times 1$	$63 \times 19 \times 1$

These three datasets were designed to assess different modeling paradigms. The *Processed Signal* and *Feature Time-Series* datasets preserve the temporal dynamics of the signals and were used to train a Transformer Encoder for time-series

classification. In contrast, the *Aggregated Feature* dataset collapses the temporal dimension, enabling the use of a more interpretable static model based on Support Vector Machine with Recursive Feature Elimination (SVM-RFE). This design allows us to compare the performance and interpretability of deep temporal models versus sparse static classifiers.

F. Classification Pipelines

To evaluate model generalization in a realistic and rigorous manner, all classification pipelines were assessed using a Leave-One-Subject-Out (LOSO) cross-validation strategy. In each fold, all data from one subject were entirely excluded from training and used exclusively for testing, thereby emulating the deployment scenario in which predictions must be made for a previously unseen individual. This approach prevents information leakage across subjects and ensures that model evaluation is not biased by subject-specific temporal patterns inadvertently included in both training and test sets. Moreover, the LOSO approach maximizes the size of the training set ($N - 1$ subjects), which is critical with a limited number of subjects. Therefore, by consistently holding out entire subjects, the LOSO protocol preserves the independence between individuals and supports fair estimation of cross-subject generalizability, which is particularly critical in analyzing the time-series of the continuous anxiety self-ratings and of its physiological correlates.

1) Pipeline 1: SVM on Aggregated Features

The classification pipeline based on Support Vector Machine with Recursive Feature Elimination (SVM-RFE) [32] was applied on the *Aggregated Feature Dataset*. This approach simultaneously performs embedded feature selection and hyperparameter optimization, aiming to identify both the most discriminative feature subset and the best classifier configuration for each training fold. As an embedded method, SVM-RFE performs feature selection jointly with model training, leveraging the structure of the classifier itself to guide the elimination process. In our implementation, we used a non-linear SVM with a Radial Basis Function (RBF) kernel, which allows capturing complex, non-linear relationships between features — an important property given the expected non-linearity in physiological patterns across subjects. Importantly, this analysis was performed on the static feature dataset obtained by aggregating temporal features over time, resulting in one feature vector per subject. This makes the SVM-RFE particularly suitable for identifying compact sets of informative summary features that generalize across subjects.

For each outer LOSO iteration, the data of a single subject were held out for final testing, while the remaining $N - 1$ subjects formed the training set. Within the training set, an inner LOSO loop was used to perform hyperparameter tuning and recursive feature selection. Specifically, we evaluated multiple combinations of the SVM hyperparameters — namely, the regularization parameter C and the RBF kernel width γ — over a predefined grid. For each hyperparameter pair, features were ranked using the SVM-RFE-CBR algorithm, which incorporates a correlation-bias reduction strategy during feature elimination [32]. Then, a sequence of inner validation

cycles was performed using increasing numbers of top-ranked features (from 1 to m), yielding a subject-wise performance matrix across feature set sizes. The optimal number of features and corresponding hyperparameters (C, γ) pair were selected based on the average F1-score (or accuracy) across the inner folds. After identifying the optimal configuration, the final model was trained on the full outer training set using the selected features and hyperparameters, and evaluated on the held-out subject. This procedure was repeated for all subjects. Importantly, this procedure means that each outer-fold prediction is generated by a model trained only on the feature subset selected within the corresponding training data. Accordingly, the obtained SVM results reflect reduced-feature models obtained in a leakage-free nested validation framework which prevents overfitting. Furthermore, by aggregating the rankings across folds, we obtained a global estimate of feature relevance.

2) Pipeline 2 & 3: Transformer Encoder model

Both the pipelines used on the *Process Signal Dataset* (Pipeline 2) and the *Feature Time-Series Dataset* (Pipeline 3) utilized a Transformer Encoder model within a nested cross-validation framework for feature selection and hyperparameter tuning. The choice of the Transformer Encoder architecture was driven by preliminary benchmarking analyses, in which multiple candidate architectures (including 1D-CNNs, MLPs, and recurrent models) were evaluated in terms of both classification performance and computational efficiency. For completeness, a detailed comparison of the tested architectures is reported in the Supplementary Materials. For unimodal evaluations (using only self-reported anxiety or only physiological data), the adopted Transformer Encoder architecture includes:

- **Input Projection & Positional Encoding:** The multi-variate time-series input (downsampled via an initial 1D-CNN with kernel and stride 10 for Pipeline 2 only) was first passed through a linear projection layer to map the features into a consistent embedding space (d_{model}). To ensure the model retained the chronological order of the physiological and behavioral signals, a positional encoding layer was subsequently applied.
- **Transformer Encoder Blocks:** The core temporal processing was handled by a Transformer Encoder consisting of 1 or 2 stacked layers. Each layer utilized a multi-head self-attention mechanism with 4 parallel heads, followed by a feed-forward network. This allowed the architecture to weigh and learn complex, long-range dependencies across the entire time series. A dropout rate of 0.3 was applied throughout to mitigate overfitting.
- **Pooling & Classifier Head:** To aggregate the temporal data, a Global Average Pooling layer was applied over the sequence dimension, compressing the context-aware sequence into a single d_{model} -dimensional summary vector per subject. This vector was passed into a dense layer with 32 units and a Rectified Linear Unit (ReLU) activation, followed by another Dropout layer ($p=0.3$). A final single-neuron output layer with a Sigmoid activation produced the binary classification probability.

For multimodal classification (the Combined set) in Pipelines 2 and 3, a **Late Fusion** architecture was designed. Because continuous self-rated anxiety and physiological signals represent fundamentally distinct types of data with different underlying dynamics, directly concatenating them before a shared self-attention mechanism risks obscuring the discriminative patterns inherent to each modality. To prevent this, our Late Fusion setup utilizes two independent Transformer Encoders: one trained exclusively on physiological data and the other trained exclusively on the continuously self-reported anxiety. Once trained, these specialized branches were frozen and used to extract high-level contextual embeddings (the output of the Global Average Pooling layers). These separate embeddings were then concatenated and fed into a new Multi-Layer Perceptron (MLP) classification head to learn the cross-modal synergies and output the final prediction. For completeness, the results of a standard “early fusion” approach (direct input concatenation prior to the self-attention layers) are provided in the Supplementary Materials. The models were trained to minimize Binary Cross-Entropy loss using an Adam optimizer. A visual representation of the architecture is presented in Figure 2.

Similarly to the SVM-RFE, the validation strategy consisted of the following:

- **Outer Loop (Evaluation):** The primary evaluation was a 63-fold LOSO-CV. In each fold, one subject was designated as the test set, with the remaining 62 subjects used for training and validation.
- **Inner Loop (Grid Search and Feature Selection):** For each outer fold, a two-stage inner loop was executed on the $N - 1$ training subjects to jointly optimize model hyperparameters and select the most discriminative feature subset.

- 1) A **grid search** was first performed to explore combinations of key model hyperparameters. The specific search space was tailored to each pipeline and it is detailed in Table II.
- 2) To evaluate each hyperparameter combination, a full **Forward Feature Selection (FFS)** procedure was conducted.
 - The FFS begins with an empty set and iteratively adds the feature that provides the greatest performance improvement.
 - To score a candidate feature set (the current set plus one new feature), a complete **3-fold cross-validation** was conducted on the $N - 1$ subjects. The Transformer Encoder was trained in parallel across the 3 folds with an early stopping mechanism (**patience of 20 epochs**).
 - The performance of the candidate feature set was determined by the F1-score (for the anxious class) on the inner validation folds.
 - This process repeated until no new feature could improve the F1-score.

Consequently, the reported model results correspond to reduced-input models as well, since the final network in each outer fold was trained only on

the subset of channels/features selected within the nested training procedure.

- 3) The hyperparameter combination that enabled its FFS run to achieve the highest overall F1-score was selected as optimal for the outer fold.
- 4) This exhaustive inner loop determined four crucial parameters for the outer fold: (1) the **optimal hyperparameter set**, (2) the **optimal feature subset** (found by the FFS using those hyperparameters), (3) the **optimal number of training epochs (e^*)**, and (4) the **optimal classification threshold**, both derived from the best-performing step of the winning FFS run.

- **Final Model Training and Testing:** After the inner loop identified the optimal set of hyperparameters and features, a single, final Transformer Encoder model was trained from scratch on the *entire* $N - 1$ training set, using only the selected feature subset and optimal hyperparameters, and training for exactly e^* epochs. This model was then used to classify the one held-out test subject, applying the optimal threshold to the output probability.

This entire nested process was repeated for all 63 subjects.

TABLE II
HYPERPARAMETER GRID SEARCH SPACE PER PIPELINE.

<i>Pipeline</i>	<i>Grid-Searched Hyperparameters</i>
Pipeline 1 (SVM-RFE)	C [0.1, 1, 10], γ -RBF kernel [2^{-6} , 2^{-4} , 2^{-2}]
Pipeline 2 & 3 (Processed Signals & Features time-series)	<i>Learning Rate</i> [10^{-3} , $5 \cdot 10^{-4}$], <i>d_model</i> [32, 64], <i>n° of Layers</i> [1, 2]

III. RESULTS

We evaluated and compared the performance of the three classification pipelines described above. Here we present the results in an order from the simplest model (SVM on static features) to the most complex (Transformer Encoder on temporal features). For each pipeline, we report overall accuracy as well as precision (P), recall (R), and F1-score for each class (Non-Anxious = class 0, Anxious = class 1).

A. SVM Classification on the Aggregated Features Dataset

Using the Aggregated Feature Dataset, the SVM-RFE achieved a respectable performance, setting a baseline for our task. Table III shows the detailed metrics. When using all features (physiological + self-ratings), the SVM reached an overall accuracy of 0.76, with an F1-score of 0.80 for the anxious class. Interestingly, using only the features derived from the continuous self-rating of social anxiety yielded an identical accuracy of 0.76 and the same anxious-class F1 of 0.80. In contrast, using only physiological features resulted in a substantially lower accuracy of 0.62, and the model struggled particularly with identifying non-anxious controls (precision 0.56, recall 0.38). In that case, the SVM was biased towards predicting the anxious class. This shows that, in a static model, the features derived from anxiety self-ratings are the dominant predictors, with physiological features offering no additional benefit.

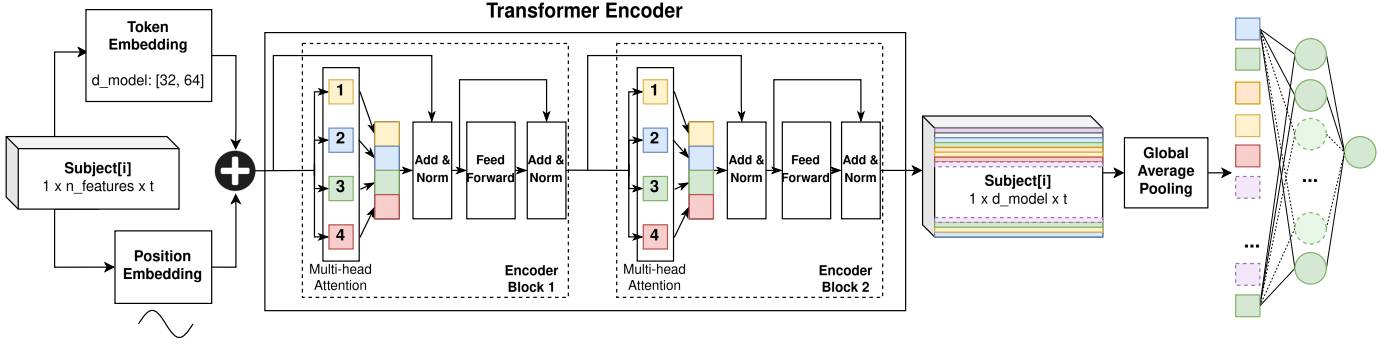


Fig. 2. The architecture of the proposed classification model. An input sequence is first processed through a linear projection layer and a positional encoding module. The sequence is then passed through a Transformer encoder consisting of two identical layers with multi-head self-attention. The resulting sequence features are aggregated using global average pooling. This pooled representation is finally passed to a two-layer multilayer perceptron, incorporating a ReLU activation function and Dropout, with a final Sigmoid function applied to produce a probability score.

TABLE III
DETAILED PERFORMANCE OF SINGLE-POINT SVM CLASSIFICATION BY FEATURE SET.

Features	Non-Anxious (Class 0)			Anxious (Class 1)			Accuracy
	P	R	F1	P	R	F1	
Physio-only	0.56	0.38	0.45	0.64	0.78	0.70	0.62
Anxiety-only	0.72	0.69	0.70	0.79	0.81	0.80	0.76
Both	0.72	0.69	0.70	0.79	0.81	0.80	0.76

B. Transformer Encoder Classification on the Processed Signal Dataset

As a second step, we assessed the performance of an end-to-end deep learning approach (Pipeline 2), applying the Transformer Encoder directly to the *Processed Signal Dataset*. As shown in Table IV, the unimodal network using only the raw anxiety signal achieved an overall accuracy of 0.76 and an F1-score of 0.82 for the anxious class. Conversely, the physio-only model struggled significantly, yielding an accuracy of 0.49 and a near-zero F1-score of 0.11 for the non-anxious control class. When integrating both raw modalities via the Late Fusion strategy, the performance actually degraded compared to the anxiety-only baseline: overall accuracy dropped to 0.67, yielding an F1-score of 0.75 for the anxious class and 0.49 for the controls (precision = 0.67, recall = 0.38). These results indicate that the Transformer Encoder struggles to automatically extract highly discriminative features from raw, unengineered physiological time-series, consequently passing these noisy physiological embeddings into the Late Fusion classification head.

C. Transformer Encoder Classification on the Feature Time-Series Dataset

The third modeling approach (Pipeline 3) applied the Transformer Encoder to the *Feature Time-Series Dataset*. When analyzing the modalities in isolation, the standard architecture demonstrated strong predictive capabilities: the model utilizing only engineered anxiety features achieved an accuracy of 81% (F1-score = 0.85 for the anxious class), while the physio-only

TABLE IV
RAW SIGNALS CLASSIFICATION WITH TRANSFORMER ENCODER (PIPELINE 2).

Features	Non-Anxious (Class 0)			Anxious (Class 1)			Accuracy
	P	R	F1	P	R	F1	
Physio-only	0.20	0.08	0.11	0.55	0.78	0.64	0.49
Anxiety-only	0.82	0.54	0.65	0.74	0.92	0.82	0.76
Both	0.67	0.38	0.49	0.67	0.86	0.75	0.67

model reached a 79% accuracy (F1-score = 0.82). Building upon these unimodal foundations, the Late Fusion architecture successfully harnessed the cross-modal synergy between the subjective reports and the physiological markers. By extracting high-level contextual embeddings from the specialized, independent branches before combining them, the Late Fusion approach yielded the best overall results in our study. The model achieved a peak overall accuracy of 83%, maintaining a high F1-score of 0.85 for the anxious class while significantly boosting the control class F1-score to 0.78 (up from 0.73 in the anxiety-only model), as detailed in Table V and Figure 3.

TABLE V
FEATURE DATASET CLASSIFICATION WITH TRANSFORMER ENCODER (PIPELINE 3)

Feature	Non-Anxious (Class 0)			Anxious (Class 1)			Accuracy
	P	R	F1	P	R	F1	
Physio-only	0.72	0.81	0.76	0.85	0.78	0.82	0.79
Anxiety-only	0.89	0.62	0.73	0.78	0.95	0.85	0.81
Both	0.80	0.77	0.78	0.84	0.86	0.85	0.83

D. Analysis of Selected Features

To understand which data modalities and specific features were most influential for classification, we analyzed the feature selection results from each of the three pipelines. The importance of each feature or feature group was quantified by

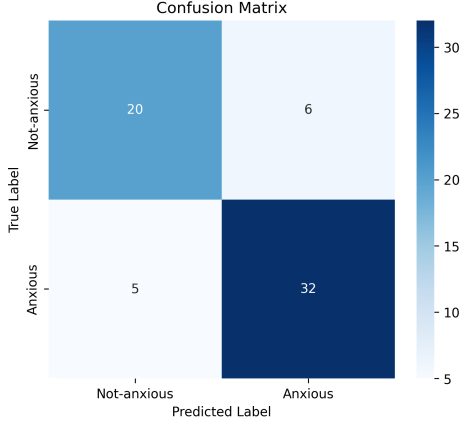


Fig. 3. Confusion Matrix representing the results from Pipeline 3 (Late Fusion Transformer Encoder over Feature Dataset on all features).

its selection frequency across the LOSO folds. Furthermore, to provide a more granular view, we computed feature importance scores based on the rank of selection within each fold (a feature selected first receives a higher score than one selected second), which were then aggregated across all subjects.

1) Pipeline 1: SVM on Aggregated Features

The analysis of the static model using aggregated features reveals that the SVM-RFE method consistently selected a highly sparse set of predictors. As detailed in Table VI, only three features out of the nineteen available were ever chosen by the model across the 63 cross-validation folds. The behavioral feature `MeanAnxietyRate` and the physiological feature `EDAsymp` were selected in 100% of the folds, establishing them as the most robust predictors. The `StdAnxietyRate` feature was also frequently included, being selected in 79.4% of folds. All other features, including all HRV metrics and the remaining EDA-derived features, were never selected. The feature importance scores, shown in Figure 4, visually confirm this result, with `MeanAnxietyRate` and `EDAsymp` being the overwhelmingly dominant predictors.

TABLE VI
PIPELINE 1 (SVM ON AGGREGATED FEATURES): % OF FEATURE SELECTION ACROSS 63 LOSO FOLDS (RANKED BY SELECTION COUNT)

Feature Name	Feature Type	Selection Count	Selection (%)
MeanAnxietyRate	Behavioral	63	100.0%
EDAsymp	Physiological	63	100.0%
StdAnxietyRate	Behavioral	50	79.4%
All other 16 features*	Mixed	0	0.0%

*Unselected features include: `meanHRV`, `stdHRV`, `RMSSD`, `LFnu`, `HFnu`, `LF/HF`, `SCR_peak`, `SCR_n`, `Ampsum`, `SCR_mean`, `SCR_std`, `SCL_mean`, `SCL_std`, `tvsymp`, `MeanAnxiety`, and `StdAnxiety`.

2) Pipeline 2: Transformer Encoder on Processed Signal Dataset

Because of the Late Fusion architecture, FFS was conducted independently on the anxiety-only and physio-only input branches. As presented in Table VII, the raw `Anxiety` signal was selected in 100% of the cross-validation folds since

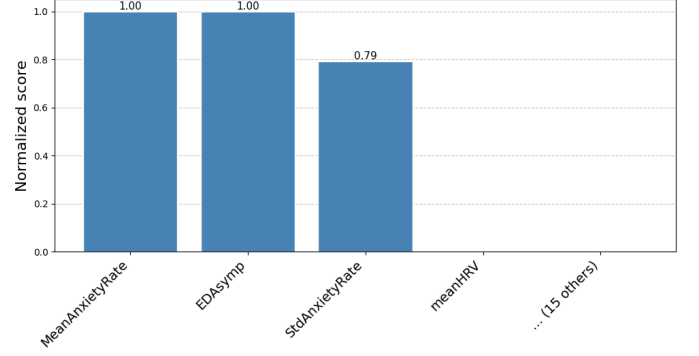


Fig. 4. Aggregated feature importance scores for Pipeline 1 (SVM on Aggregated features). Scores are derived from the selection rank in each of the 63 LOSO folds, reflecting both selection frequency and priority.

it is the sole behavioral time-series in the *Processed Signal Dataset*. Conversely, FFS on the independent physiological branch revealed a fragmented reliance on the unengineered signals: the model primarily leveraged `SMNA` (selected in 80.95% of folds), followed by inconsistent contributions from `tvsymp` (41.27%), `HRV` (41.27%), and `SCL` (31.75%), while `SCR` was entirely ignored. As established in the performance analysis, integrating these optimized physiological and behavioral embeddings via the final classification head ultimately degraded the model's accuracy compared to the anxiety-only baseline. These scattered selection frequencies and the resulting performance drop reinforce the conclusion that the Transformer Encoder struggles to extract generalizing, synergistic patterns directly from raw physiological time-series, causing the physiological branch to act as noise when not preceded by explicit feature engineering.

TABLE VII
PIPELINE 2 (TRANSFORMER ENCODER ON PROCESSED SIGNALS): % OF INPUT SIGNAL SELECTION ACROSS 63 LOSO FOLDS USING FFS

Signal Name	Signal Type	Selection Count	Selection Pct. (%)
Anxiety	Behavioral	60	100.0%
SMNA	Physiological	34	80.95%
tvsymp	Physiological	15	41.27%
HRV	Physiological	9	41.27%
SCL	Physiological	4	31.75%
SCR	Physiological	0	0.00%

3) Pipeline 3: Transformer Encoder on Feature Time-Series Dataset

In the best-performing pipeline, which utilized the *Feature Time-Series Dataset*, FFS was conducted on the time-series of individual engineered features. Consistent with the Late Fusion approach described in the previous pipeline, feature selection and frequency analysis were performed independently for the physio-only and anxiety-only branches. As detailed in Table VIII and Figure 6, when evaluating physiological features alone, the model heavily prioritized the time-varying sympathetic index (`tvsymp`, selected in 93.6% of folds) and the frequency of `SCR_n` (92.1%). These were followed by moderate contributions from heart rate variability metrics,

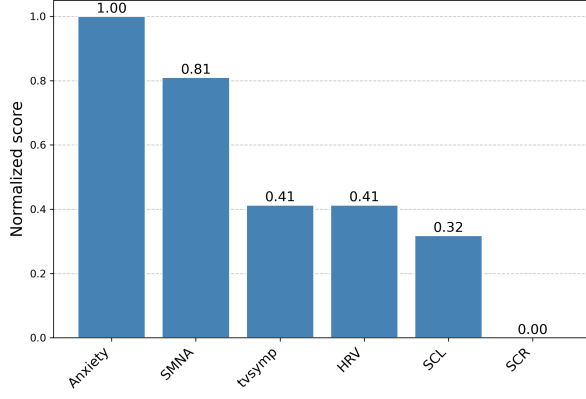


Fig. 5. Input signal importance scores for Pipeline 2 (Transformer Encoder on Processed Signal Dataset). The scores reflect the aggregated rank and frequency of selection for each signal.

specifically RMSSD (28.6%) and stdHRV (27.0%). Table IX and Figure 7 show the selection frequencies for the isolated behavioral features. Consistent with previous pipelines, the continuous self-report was dominant: MeanAnxiety was selected in 100% of the folds, supplemented by variations in the joystick movement such as StdAnxiety (25.4%) and StdAnxietyRate (22.2%).

TABLE VIII
PIPELINE 3 (TRANSFORMER ENCODER ON FEATURE TIME-SERIES):
NUMBER (N°) AND PERCENTAGE (%) OF PHYSIOLOGICAL FEATURE
SELECTION ACROSS 63 LOSO FOLDS BY FFS

Feature	n°	%	Feature	n°	%
tvsymp	59	93.65%	SCR_std	9	14.29%
SCR_n	58	92.06%	SCR_mean	9	14.29%
RMSSD	18	28.57%	SCL_mean	6	9.52%
stdHRV	17	26.98%	SCR_peak	4	6.35%
SCL_std	12	19.05%	Ampsum	0	0.0%
meanHRV	10	15.87%			

TABLE IX
PIPELINE 3: NUMBER (N°) AND PERCENTAGE (%) OF ANXIETY FEATURE
SELECTION (63 LOSO FOLDS)

Feature	n.	%	Feature	n°	%
MeanAnxiety	63	100%	StdAnxietyRate	14	22.2%
StdAnxiety	16	25.4%	MeanAnxietyRate	11	17.5%

IV. DISCUSSION

This study developed and validated a classification system to distinguish socially-anxious individuals from controls by integrating continuous self-reported anxiety with psychophysiological data from a Virtual Reality (VR) scenario simulating an everyday-life social situation. We compared three modeling pipelines, showing that a hybrid approach combining targeted feature engineering with deep learning on time-series data yielded the most robust classification.

The superiority of the final pipeline, which applied a Late Fusion Transformer Encoder to the feature time-series dataset,

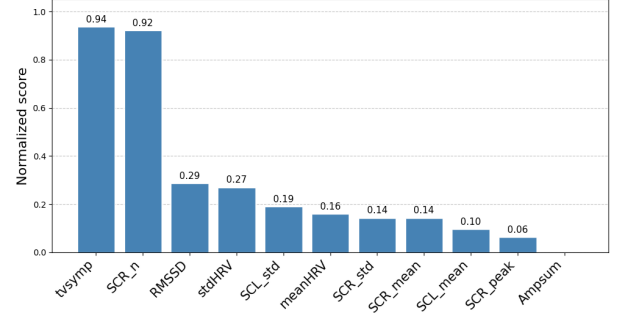


Fig. 6. Feature importance scores for Pipeline 3 (Transformer Encoder on physio-only Feature Time-Series Dataset). Scores reflect selection frequency and rank from the FFS procedure.

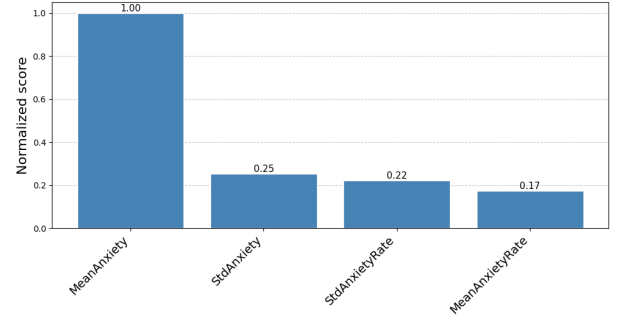


Fig. 7. Feature importance scores for Pipeline 3 (Transformer Encoder on behavioral only Feature Time-Series Dataset). Scores reflect selection frequency and rank from the FFS procedure.

derives from its ability to effectively merge these paradigms without diluting their individual predictive power. This optimized architecture achieved the highest and most balanced performance in our study (F1-score: anxious = 0.85, non-anxious = 0.78). By first condensing the preprocessed signals into meaningful temporal features and subsequently processing each modality through independent, specialized self-attention branches, the model generated highly informative contextual embeddings. The late fusion of these embeddings ultimately enabled the classification head to successfully capture and leverage synergistic cross-modal patterns, maximizing the combined predictive potential of both the subjective-behavioral and physiological data.

Crucially, however, our evaluations revealed that engineered physiological features are highly predictive in their own right. When processed independently through the Transformer Encoder, the physiological modality achieved a robust overall accuracy of 79% (F1-score = 0.82), driven primarily by the time-varying sympathetic index *tvsymp* and SCR peak frequency *SCR_n*. The peak performance of our Late Fusion framework arose from integrating these two strong, standalone predictors. This indicates that while continuous self-report acts as the strongest single modality, physiological data is not merely a secondary supplement; it provides a highly accurate, parallel measure of autonomic arousal. Furthermore, our results strongly support a hybrid approach where feature engineering precedes deep learning. Transformer Encoders, while highly capable of modeling complex temporal dependencies, can still

be hindered by the noise of raw physiological signals. Feature extraction effectively acted as a denoising step, transforming raw autonomic fluctuations into robust sequences that enabled the self-attention mechanism to isolate predictive patterns. Psychologically, these findings highlight a dual-process assessment: the continuous self-report of anxiety reflects a conscious cognitive appraisal of the virtual exposure, while features like tv_{symp} capture involuntary autonomic regulation. The success of the Late Fusion approach proves that merging these parallel windows into the user’s state enriches the model, compensating perfectly for moments when the conscious self-report might be delayed or imprecise.

This methodological rigor provides a solid foundation for evaluating our model’s generalization capabilities. While direct numerical comparisons across studies utilizing distinct paradigms, samples, VR scenarios, and experimental procedures would be inappropriate, contextualizing our findings within the broader literature suggests that our results (overall accuracy = 0.83, F1-score = 0.85) are satisfactory. For general context, the studies more comparable to the present one (i.e., using digital tools to assess social or generalized anxiety) have reported (slightly) poorer performances: a study combining physiological and acoustic data reported an F1-score of 0.76 [9] and another using CNNs on biosensor data achieved 0.75 accuracy [12]. Reports of near-perfect accuracy [20] involve a validation scheme that does not strictly separate subjects, leading to data leakage where windows from the same subject appear in both the training and the test set. Compared to VR-based SAD paradigms [7], [8], [10], our work introduced novel elements in paradigm design, analytical approach, and data interpretability. The first novelty concerns the paradigm: a waiting-room scenario that progressively fills with interacting avatars elicited increasing social anxiety, targeting the SAD subtype triggered by everyday-life social interactions: these are more frequent and unavoidable than performance situations (e.g., giving a public speech), but less studied [19], [20], [10]. This offers greater ecological validity than artificial laboratory tasks involving EEG/fMRI during cognitive challenges [23], [24]. The second novelty is the continuous anxiety report, combining subjective and behavioral components in a single signal, capable to capture anxiety fluctuations while avoiding recall biases. This overcomes the limitations coming from discrete measures [8] and from inferences based on the experimenter’s assumptions about how much anxiety is elicited by each experimental phase [12]. The third novelty is methodological: unlike single-modality studies [21], [22], we integrated behavioral and physiological data and systematically compared SVM, end-to-end Transformer Encoder, and hybrid Transformer Encoder approaches.

A critical analysis of the results requires careful consideration of cases in which the model’s classification diverged from the L-SAS-based label. These discrepancies are not necessarily errors; rather, they provide insight into the limitations of retrospective self-report measures [14] and into the added value of our multimodal, moment-by-moment approach. The L-SAS, while widely used and validated, as all questionnaires is inevitably subject to recall bias, limited interoceptive accuracy, and social desirability effects [5], [4]. In the case of

controls classified as anxious by our system (false positives), most profiles suggest the presence of subclinical social anxiety, either actively managed by the participant or underestimated by the original L-SAS cut-offs. Of the six participants in this category, four fell within the intermediate L-SAS range (30–55) later proposed by Liebowitz as indicative of social anxiety symptomatology [6]: their continuously self-reported anxiety was higher than that of the rest of the control group, suggesting that the system detected a latent vulnerability overlooked by the questionnaire’s dichotomous classification. Of note, the gender distribution among false positives (50% females, 50% males) and false negatives (100% females) could suggest that males have underestimated their social anxiety (e.g., to adhere to masculine stereotypes), however the current numbers are too small to allow robust interpretations. While the subsample of participants falling in this subclinical range of L-SAS score is too small to allow a reliable analysis, its deeper characterization in future studies could represent a noteworthy application of our classification model: in particular, refining the role of psychophysiological measures in this specific subsample could reveal symptoms’ underestimations due to alexythymia, reluctance in admitting our own weaknesses, or the perception of stigma in reporting the full intensity of symptoms. A disalignment between behavioral and physiological predictors could underlie interoceptive inaccuracies whose role in psychopathologies is highly debated (see [33] *for an authoritative review on this topic*). *Further research is needed to address these promising applications of our VR tool.*

The critical interpretation of these results also requires specifying that our study (coherently with the previous literature) is based on a “ground truth” (the L-SAS score), which is a self-report measure, not a formal clinical diagnosis. Choosing it as a ground-truth was a choice dictated by i) its validated robustness in measuring SAD symptoms [34], ii) its widespread in the scientific literature about SAD [7], [8], and iii) its reproducibility and standardizability [17]. However, in the overall interpretation of results, L-SAS classification should not be intended as a ground truth but rather as a widely validated tool assessing the self-perception of SAD symptoms (*which, coherently, is correlated with the continuous self-report of social anxiety in the present study; see Supplementary Figure 3*). Consequently, our model has effectively learned to predict a questionnaire score rather than the clinical disorder itself. Moreover, our sample ($N = 63$) consisted mainly of young university students, limiting generalizability. While VR increases ecological validity, it remains a simulation, and our single-scenario design may not capture anxiety patterns in other contexts.

Beyond classification, this work has therapeutic potential. Linking continuous anxiety reports to in-scenario events enables identification of specific anxiety triggers, supporting personalized exposure protocols. Combining subjective-behavioral and physiological data yields a comprehensive profile of anxiety across experience, behavior, and physiology. The system’s scalability is notable: models using only behavioral data from commercial VR controllers performed well, suggesting a cost-effective “lite” version for broader adoption.

Future work will apply this paradigm in longitudinal studies

to track changes in social anxiety and test diverse VR scenarios for contextual anxiety profiling. The architecture is well-suited for closed-loop biofeedback, where real-time anxiety signals could dynamically adjust VR stressors, or for implicit systems predicting anxiety solely from physiological signals, enabling adaptation without even requiring continuous self-report.

V. CONCLUSION

In conclusion, this study introduces a novel paradigm for the assessment of Social Anxiety Disorder. By integrating a continuous, in-the-moment subjective report with objective physiological data within an ecologically valid Virtual Reality environment, our approach offers a deeper, more dynamic, and temporally-resolved understanding of the individual anxiety experience. Our classification model can be seen as a tool to identify instances of incoherence between a participant's retrospective self-report (L-SAS) and their in-the-moment psychophysiological and subjective-behavioral responses. Finally, coherently with the principles of open science, all the software used to run the experiment — as well as the code used to analyze its results — is freely accessible at https://github.com/ilmarcopardo/SAD_analysis, allowing a full replication of the study and promoting its real impact on clinical and research practice. Ultimately, the potential of such multimodal, dynamic assessment tools is to revolutionize the evaluation and treatment of SAD, shifting the field towards a new era of psychometrics that is more objective, personalized, and sensitive to the moment-to-moment fluctuations that define the lived experience of anxiety.

ACKNOWLEDGMENT

We thank the Clinical Psychologist Sara Said and the Biomedical Engineers Arianna Galigani and Valerio Vatteroni for their contribution in collecting part of the data analyzed in the present paper. We also acknowledge the use of Google Gemini AI language model for English language revision.

REFERENCES

- [1] A. Diagnostic, "Statistical manual of mental disorders: Dsm-5 (ed.) washington," DC: American Psychiatric Association, 2013.
- [2] D. F. Santomauro, A. M. M. Herrera, J. Shadid, P. Zheng, C. Ashbaugh, D. M. Pigott, C. Abbafati, C. Adolph, J. O. Amlag, A. Y. Aravkin *et al.*, "Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the covid-19 pandemic," *The Lancet*, vol. 398, no. 10312, pp. 1700–1712, 2021.
- [3] M. R. Liebowitz, "Liebowitz social anxiety scale," *Journal of Anxiety Disorders*, 1987.
- [4] P. S. Brenner and J. DeLamater, "Lies, damned lies, and survey self-reports? identity as a cause of measurement bias," *Social psychology quarterly*, vol. 79, no. 4, pp. 333–354, 2016.
- [5] M. Zimmerman, "The value and limitations of self-administered questionnaires in clinical practice and epidemiological studies," *World Psychiatry*, vol. 23, no. 2, p. 210, 2024.
- [6] 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.
- [7] T. Horigome, S. Kurokawa, K. Sawada, S. Kudo, K. Shiga, M. Mimura, and T. Kishimoto, "Virtual reality exposure therapy for social anxiety disorder: a systematic review and meta-analysis," *Psychological medicine*, vol. 50, no. 15, pp. 2487–2497, 2020.
- [8] S. Shahid, J. Kelson, and A. Saliba, "Effectiveness and user experience of virtual reality for social anxiety disorder: systematic review," *JMIR mental health*, vol. 11, no. 1, p. e48916, 2024.
- [9] H. Choi, Y. Cho, C. Min, K. Kim, E. Kim, S. Lee, and J.-J. Kim, "Multiclassification of the symptom severity of social anxiety disorder using digital phenotypes and feature representation learning," *Digital Health*, vol. 10, p. 20552076241256730, 2024.
- [10] H. Kim, E. C. Han, P. B. Muntz, and J. Kemp, "Virtual reality in the treatment of anxiety-related disorders: A review of the innovations, challenges, and clinical implications," *Current Psychiatry Reports*, pp. 1–10, 2025.
- [11] D. Banakou, T. Johnston, A. Beacco, G. Senel, and M. Slater, "Desensitizing anxiety through imperceptible change: Feasibility study on a paradigm for single-session exposure therapy for fear of public speaking," *JMIR formative research*, vol. 8, p. e52212, 2024.
- [12] D. Mevleviöglu, S. Tabirca, and D. Murphy, "Real-time classification of anxiety in virtual reality therapy using biosensors and a convolutional neural network," *Biosensors*, vol. 14, no. 3, p. 131, 2024.
- [13] G. Norman, "Hi! how are you? response shift, implicit theories and differing epistemologies," *Quality of Life Research*, vol. 12, no. 3, pp. 239–249, 2003.
- [14] S. Schwartz, "Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research," *Social science & medicine*, vol. 48, pp. 1531–1548, 1999.
- [15] T. E. Raffegau, S. A. Brinkerhoff, M. Clark, A. D. McBride, A. Mark Williams, P. C. Fino, and B. Fawver, "Walking (and talking) the plank: dual-task performance costs in a virtual balance-threatening environment," *Experimental Brain Research*, vol. 242, no. 5, pp. 1237–1250, 2024.
- [16] L. C. Walz, M. H. Nauta, and M. aan het Rot, "Experience sampling and ecological momentary assessment for studying the daily lives of patients with anxiety disorders: A systematic review," *Journal of anxiety disorders*, vol. 28, no. 8, pp. 925–937, 2014.
- [17] S. Pilling, E. Mayo-Wilson, I. Mavranzeouli, K. Kew, C. Taylor, and D. M. Clark, "Recognition, assessment and treatment of social anxiety disorder: summary of nice guidance," *Bmj*, vol. 346, 2013.
- [18] M. Pardini, S. Frumento, M. Martini, G. Rho, V. Vatteroni, K. Tharun, M. Alaimo, F. A. Galatolo, M. De Marinis, E. P. Scilingo *et al.*, "Deep learning-based classification of social anxiety disorder using continuous self-reported anxiety in virtual reality," in *2025 IEEE Medical Measurements & Applications (MeMeA)*. IEEE, 2025, pp. 1–6.
- [19] J.-H. Park, Y.-B. Shin, D. Jung, J.-W. Hur, S. P. Pack, H.-J. Lee, H. Lee, and C.-H. Cho, "Machine learning prediction of anxiety symptoms in social anxiety disorder: utilizing multimodal data from virtual reality sessions," *Frontiers in Psychiatry*, vol. 15, p. 1504190, 2025.
- [20] R. Shaikat-Jali, N. van Zalk, D. E. Boyle *et al.*, "Detecting subclinical social anxiety using physiological data from a wrist-worn wearable: small-scale feasibility study," *JMIR Formative Research*, vol. 5, no. 10, p. e32656, 2021.
- [21] K. Chadaga, S. Prabhu, N. Sampathila, R. Chadaga, D. Bhat, A. K. Sharma, and K. Swathi, "Sadxai: Predicting social anxiety disorder using multiple interpretable artificial intelligence techniques," *SLAS technology*, vol. 29, no. 2, p. 100129, 2024.
- [22] N. K. Sahu, M. Yadav, and H. R. Lone, "Unveiling social anxiety: Analyzing acoustic and linguistic traits in impromptu speech within a controlled study," *ACM Journal on Computing and Sustainable Societies*, vol. 2, no. 2, pp. 1–19, 2024.
- [23] A. Al-Ezzi, N. Kamel, A. A. Al-Shargabi, F. Al-Shargie, A. Al-Shargabi, N. Yahya, and M. I. Al-Hiyali, "Machine learning for the detection of social anxiety disorder using effective connectivity and graph theory measures," *Frontiers in psychiatry*, vol. 14, p. 1155812, 2023.
- [24] M. Xing, J. M. Fitzgerald, and H. Klumpp, "Classification of social anxiety disorder with support vector machine analysis using neural correlates of social signals of threat," *Frontiers in psychiatry*, vol. 11, p. 144, 2020.
- [25] A. Prunas, I. Sarno, E. Preti, F. Madeddu, and M. Perugini, "Psychometric properties of the italian version of the scl-90-r: A study on a large community sample," *European psychiatry*, vol. 27, no. 8, pp. 591–597, 2012.
- [26] M. Martini, E. Viola, F. Bossi, S. Frumento, A. Iannizzotto, S. Said, A. L. Callara, F. Solari, E. P. Scilingo, A. Greco *et al.*, "Designing an immersive virtual reality scenario for social anxiety elicitation and modeling: a preliminary evaluation," in *2024 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRINE)*. IEEE, 2024, pp. 435–440.
- [27] M. P. Tarvainen, J.-P. Niskanen, J. A. Lipponen, P. O. Ranta-Aho, and P. A. Karjalainen, "Kubios hrv—heart rate variability analysis software,"

Computer methods and programs in biomedicine, vol. 113, no. 1, pp. 210–220, 2014.

- [28] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, “cvxeda: A convex optimization approach to electrodermal activity processing,” *IEEE transactions on biomedical engineering*, vol. 63, no. 4, pp. 797–804, 2015.
- [29] A. Baldini, S. Frumento, D. Menicucci, A. Gemignani, E. P. Scilingo, and A. Greco, “Modeling subjective fear using skin conductance: A preliminary study in virtual reality,” in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 3451–3454.
- [30] H. F. Posada-Quintero, J. P. Florian, A. D. Orjuela-Cañón, T. Aljama-Corrales, S. Charleston-Villalobos, and K. H. Chon, “Power spectral density analysis of electrodermal activity for sympathetic function assessment,” *Annals of biomedical engineering*, vol. 44, no. 10, pp. 3124–3135, 2016.
- [31] H. F. Posada-Quintero, J. P. Florian, Á. D. Orjuela-Cañón, and K. H. Chon, “Highly sensitive index of sympathetic activity based on time-frequency spectral analysis of electrodermal activity,” *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, vol. 311, no. 3, pp. R582–R591, 2016.
- [32] K. Yan and D. Zhang, “Feature selection and analysis on correlated gas sensor data with recursive feature elimination,” *Sensors and Actuators B: Chemical*, vol. 212, pp. 353–363, 2015.
- [33] P. M. Jenkinson, A. Fotopoulou, A. Ibañez, and S. Rossell, “Interception in anxiety, depression, and psychosis: a review,” *eClinicalMedicine*, vol. 73, p. 102673, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2589537024002529>
- [34] R. G. Heimberg, K. Horner, H. Juster, S. Safren, E. Brown, F. Schneier, and M. Liebowitz, “Psychometric properties of the liebowitz social anxiety scale,” *Psychological medicine*, vol. 29, no. 1, pp. 199–212, 1999.

Marco Pardini (Member, IEEE) is a PhD student at the University of Pisa, where he received a MS degree in Artificial Intelligence and Data Engineering. His research interests include artificial intelligence, Human-Robot Interaction, and physiological data modelling. His primary focus is on machine/deep learning applications in the field of affective computing and on the development of cognitive architectures for autonomous systems.

Sergio Frumento is a licensed Psychologist and postdoc research fellow at the University of Pisa, (Italy) in the Department of Information Engineering. His current research interests include the use of adaptive virtual reality as a form of exposure therapy for anxiety disorders (e.g., social anxiety, phobias, PTSD), and the neuroscience behind adaptive/maladaptive emotions, consciousness and perception.

Matteo Martini is a PhD student in Computer Science at the University of Genova, supervised by Prof. Manuela Chessa and Prof. Danilo Pani. His research focuses on innovative rehabilitative solutions using Virtual Reality for individuals with severe cognitive and motor disabilities, and other vulnerabilities. His interests include eXtended Reality, Human-Computer Interaction, Visual Perception, and Serious Games. He is part of the Perception and Interaction Laboratory (PILab).

Martina Alaimo holds a Master’s degree in Psychology and conducted experimental sessions as part of the BRAVE project, which was the focus of her thesis work.

Gianluca Rho is a postdoc researcher in Biomedical Engineering at the Department of Information Engineering, University of Pisa. His research focuses on biomedical signal processing, with particular interest in EEG activity and connectivity, as well as autonomic nervous system correlates such as electrodermal activity, electrocardiography, and photoplethysmography. His work spans the fields of neuroscience, affective computing, and personalized medicine. He is also involved in the development and application of wearable devices for the estimation of physiological parameters relevant to affective computing and personalized healthcare.

Noemi Paparo is a PhD student in the Department of Information Engineering at the University of Pisa. She received her MSc degree in Biomedical Engineering at the University of Pisa. Her research focuses on the analysis and processing of physiological signals and on machine learning approaches for affective state assessment. Her research interests include affective computing, neuroscience, and signal processing, with a particular focus on developing predictive models for biofeedback applications.

Mario G.C.A. Cimino (Senior Member, IEEE) is an associate professor at the Department of Information Engineering, University of Pisa. His research focuses on Information Systems and Artificial Intelligence. He is an associate editor of the international journals *Granular Computing* (Springer) and *Ambient Intelligence and Humanized Computing* (Springer). He is the chair of the IEEE CIS Task Force “Intelligent Agents,” IEEE Computational Intelligence Society. He is a visiting academic at the Cognitive Robotics and Autonomous Systems Laboratory, School of Computing, University of Kent, UK.

Enzo Pasquale Scilingo (Senior Member, IEEE) received the PhD degree. He is a full professor in electronic and information bioengineering with the University of Pisa. He coordinated the European projects “PSYCHE”, “NEVERMIND”, and “POTION”. His main research interests are in wearable monitoring systems, human-computer interfaces, biomedical and biomechanical signal processing, modeling, control, and instrumentation. He is the author of more than 300 papers on peer-review journals, contributions to international conferences, and chapters in international books.

Danilo Menicucci is an associate professor who received the MS degree in applied physics, and the PhD degree in basic neuroscience from the University of Pisa, Pisa, Italy. He is currently a researcher in psychophysiology with the Department of Surgical, Medical and Molecular Pathology and Critical Care Medicine, University of Pisa. His current research interests include social anxiety disorders, sleep psychophysiology, cognitive and emotional modulation of brain, cognitive neuroscience, and experimental psychology and neurophysiology.

Manuela Chessa (Member, IEEE) is Associate Professor in Computer Science at Dept. of Informatics, Bioengineering, Robotics, and Systems Engineering of the University of Genoa. She is the Principal Investigator of the Perception&Interaction Lab @DIBRIS (PILab). Her research interests are focused on the study and development of natural human-machine interfaces based on virtual, augmented, and extended reality and on the perceptual and cognitive aspects of interaction in VR, AR, and XR.

Alberto Greco (Senior Member, IEEE), is an Associate Professor at the Department of Information Engineering, University of Pisa, and a member of the “E. Piaggio” Research Center. He coordinated the PRIN project BRAVE and co-coordinated the H2020 project POTION. His research focuses on physiological signal analysis and modeling, machine learning, and wearable systems, with applications in affective computing, social interaction, and mental health. He has authored over 150 scientific publications and co-founded CARTESIA s.r.l. to develop AI-based solutions for emotional support and well-being.