Discriminating Stress and Mental Workload Using Multimodal Physiological Signals

IEEE Publication Technology, *Staff, IEEE,*

***Abstract*—Reliable inference of cognitive–affective states such as stress and mental workload (MWL) is essential for adaptive systems in domains ranging from education and healthcare to safety-critical operations. However, these states often co-occur and share physiological signatures, complicating their discrimination and limiting model generalizability. This study systematically varied stress and MWL in a controlled (n=40), ecologically valid task environment using virtual reality (VR), recording multimodal physiological signals including electroencephalography (EEG), electrodermal activity (EDA), heart rate, and pupillometry. Specifically, this study developed an immersive VR version of the Trier Social Stress Test incorporating a 2×2 factorial design that systematically varied stress (high/low) and MWL (high/low) to examine their overlapping yet partially discrim- inable physiological responses. Using interpretable machine learning models and rigorous cross-participant validation, we demonstrate that stress and MWL can be partially dissociated based on distinct patterns in neural and autonomic features. EEG-derived entropy and connectivity measures provided the strongest predictive value for both states, while autonomic indices contributed complementary information for stress. These findings clarify the overlapping and unique physiological correlates of stress and MWL, offering practical guidance for the design of more reliable and generalizable affective computing systems. The methodological framework and analytic insights presented here are applicable across a range of immersive and non-immersive contexts, advancing the field toward robust mental state inference in real-world settings.**

***Index Terms*—Affective computing, stress, mental workload, EEG, EDA, HRV, pupillometry, multimodal physiology, machine learning, model interpretability, virtual reality.**

I. INTRODUCTION

Physiological computing systems infer users' cognitive–affective states (e.g., stress, engagement, fatigue) from neurophysiological and autonomic signals such as electroencephalography (EEG), electrodermal activity (EDA), heart rate and heart rate variability (HR/HRV), and pupillometry [1]–[4]. These systems enable continuous, unobtrusive monitoring of the user's internal states without requiring explicit input [2], [3], [5], [6]. Physiological computing is particularly relevant to safety-critical, high-demand settings—such as aviation [7], driving [6], [8], and defence [9], [10]—where awareness of the user's cognitive–affective state is vital for system performance and safety. Yet, inferring user states through explicit reporting can disrupt task execution and increase risk [2], [7], [11].

These implicit inferences must remain reliable under realistic conditions. Errors in state inference can lead systems to behave erroneously, such as intervening when unnecessary, or failing to act when required [5], [12], thus producing unintended adverse effects on user experience and performance [12], [13]. Such events also undermine user trust in the system's reliability and adaptive capabilities [14]–[16]. However, achieving the level of accuracy required for reliable state inference remains challenging. Models that perform well on training data often experience substantial degradation in accuracy when applied to new users or contexts, reflecting poor cross-participant/user and cross-dataset/session generalisation [16]–[18]. For instance, a recent cross-dataset stress-detection study reported reductions in classification accuracy of roughly 20–40 % when models trained on one type of stressor were tested on another, illustrating how contextual changes can markedly degrade model performance [19].

A growing body of work suggests that poor cross-participant and cross-context generalisation is partly explained by state discriminability — the extent to which a physiological feature or model output is sensitive to the target mental state rather than concurrent influences such as task structure, general physiological arousal, or environmental artefacts. Low discriminability occurs when features encode either irrelevant patterns spuriously correlated with the task or non-unique responses expressed across multiple mental states. Recent evidence suggests that limited discriminability may be a key contributor to poor generalisability [20]–[22]. For example, EEG models trained to detect one mental state lost up to approximately 15% accuracy when the training and test data differed in the other concurrent state dimension as demostrated in [21], indicating that even modest variation in co-occurring states can substantially degrade detection reliability. These findings highlight that overlapping physiological responses across mental states can impair state discriminability and, in turn, model performance. However, the phenomenon of overlapping physiological responses, and their impact on state discriminability, remains poorly characterised and relatively underexplored within affective computing, despite its potential to improve the reliability and generalisability of such systems across users and contexts

Accounting for overlapping physiological responses is especially salient in the context of stress and mental workload (MWL), two foundational yet often conflated constructs in physiological computing. These states represent distinct but interacting dimensions of cognitive–affective function that influence attention, decision-making, and overall task performance [1]–[3], [23]. Stress reflects an affective–autonomic response to perceived threat or evaluation, whereas MWL reflects the cognitive demands placed on limited attentional

This paper was produced by the IEEE Publication Technology Group. They are in Piscataway, NJ.

Manuscript received April 19, 2021; revised August 16, 2021.

resources [1]–[3], [23]. Because both modulate arousal and engage cognitive control systems via overlapping autonomic and cortical mechanisms [8], [24], [25], their physiological signatures converge across modalities [1], creating ambiguity for machine-learning models that must distinguish between cognitive and affective sources of physiological activation [20], [21]. Despite these overlaps, converging evidence suggests that stress and MWL are elicited by different kinds of challenges and engage distinct neurophysiological processes. MWL primarily increases activation within the executive-control network as cognitive demands intensify [2], whereas stress involves affective–autonomic responses [26] that can further recruit these networks under social evaluation or threat [27]. Understanding the shared and distinct physiological signatures of stress and MWL is crucial for informing the design of systems that can infer and respond appropriately to each state.

Most of the prior research on the physiological discriminabiliy of stress and MWL has relied on conventional computer-based laboratory paradigms that simulate stress or MWL through onscreen tasks [20]–[22], [28]–[32]. Within these paradigms, stress was typically induced through tightly controlled manipulations, such as aversive auditory stimuli [20] or anticipated public speaking [21], [22], which, while effective for eliciting acute stress under controlled conditions, offer limited generalisation to more ecologically valid contexts. In addition, many investigations have drawn on a narrow range of physiological measures, often a single modality [21], [22], [28], [30], [33] or sparse sensor arrays [20], [30], yielding a limited account of the neural and autonomic processes involved.

To address these limitations, we introduce an interpretable multimodal framework built on a virtual-reality adaptation of the Trier Social Stress Test (VR-TSST) to examine how overlapping yet discriminable neural and autonomic responses to stress and MWL contribute to physiological state inference. First, the framework employs a 2×2 factorial design to vary stress (high/low) and MWL (high/low) while maintaining comparable sensory and task settings, balancing ecological realism with experimental control. Second, it integrates multimodal physiological sensing—neurological (EEG), autonomic (HR/HRV, EDA), and oculometric (pupillometry)—within a unified analytic pipeline, enabling direct comparison of neural and peripheral indicators of state discriminability. Third, it incorporates machine-learning models alongside traditional inferential analyses, including group-level statistics (ANOVAs, correlations) and regression. Together, these components provide a principled and interpretable basis for advancing the discrimination of co-occurring mental states in physiological computing systems.

The contributions are as follows:

1) Paradigm – A 2×2 VR adaptation of the TSST that manipulates stress and MWL independently within one interactive loop.
2) Pipeline – A multimodal acquisition and preprocessing framework (EEG, EDA, HR/HRV, pupil) designed for subject-independent modelling.
3) Inference – Gradient-boosted decision tree (XGBoost) models trained to predict continuous subjective ratings of stress and MWL under LOSO evaluation, benchmarked against linear baselines.

Empirical analysis shows that factorial manipulations elicit the intended subjective changes. Predictive models achieve moderate but non-trivial accuracy gains relative to baselines. These findings indicate that multimodal models can provide partial discrimination of overlapping states in ecologically plausible tasks, thereby advancing the reliability of physiological inference for adaptive systems.

The remainder of the paper is organised as follows. Section II details the paradigm, acquisition pipeline, and modelling framework. Section III reports manipulation checks and predictive analyses. Section IV examines state discriminability and broader implications.

II. Methods

A. Experimental Design

The study employed a 2 × 2 within-subjects design with factors of Stress (Low vs. High) and MWL (Low vs. High), resulting in four experimental conditions: **Low-Stress/Low-MWL**, **Low-Stress/High-MWL**, **High-Stress/Low-MWL**, and **High-Stress/High-MWL**. Participants completed all four conditions in a fully counterbalanced order, with two alternate versions of the **High-MWL** task used to minimise practice effects. Across participants, the 24 counterbalanced condition orders were doubled to account for these task variants, yielding 48 unique experimental configurations.

B. Immersive Environment and Task Design

The Trier Social Stress Test (TSST) is a widely used and well-validated protocol for eliciting psychosocial stress, reliably inducing increases in subjective, physiological, and hormonal stress markers [34], [35]. This study implemented a virtual adaptation of the TSST to reproduce two of its core components: social evaluation by observers and a cognitively demanding arithmetic task. The immersive VR format preserved these elements while providing enhanced experimental control and ecological validity [36].

Stress was manipulated by varying the intensity of social evaluation conveyed through the immersive social environment and accompanying pre-task instructions. In the **High-stress** condition, participants performed arithmetic tasks before avatars that displayed attentive, evaluative nonverbal behavior and were formally dressed, simulating a job-interview-like scenario to evoke social-evaluative stress. To reinforce the evaluative context, the experimenter explicitly informed participants that their verbal responses would be recorded and subsequently evaluated by experts, although no actual recordings were made.

In contrast, in the **Low-stress** condition, avatars were casually dressed and displayed disengaged behavior, such as avoiding eye contact and using laptops, while participants were explicitly told that their performance was neither recorded nor evaluated. Previous work has shown that social-evaluative stress is markedly reduced when observers display neutral or inattentive behavior [37]. Participants were encouraged to proceed at their own pace and reassured that mistakes were expected. Such framing has been shown to attenuate stress responses in modified TSST paradigms [38].

MWL was manipulated by varying the difficulty of the arithmetic task, consistent with established TSST-based and MWL paradigms. In the **Low-MWL** condition, participants counted aloud in steps of 15, a task intended to maintain verbal engagement while imposing minimal cognitive demand [39]. In the **High-MWL** condition, participants performed serial subtractions of 13 or 17 from four-digit numbers, tasks that impose sustained working-memory and arithmetic demands and are widely used in TSST variants to elicit high MWL [26]. Task variants were alternated to prevent participants from adopting predictable response patterns, as retrieval-based strategies impose lower cognitive demand than procedural computation [40].

To assess subjective experience following each task, participants rated their perceived stress using a single-item Likert scale, a method shown to correlate well with validated stress questionnaires and suitable for capturing acute stress responses [41]. Subjective workload was assessed after each condition using the mental-demand subscale of the NASA-TLX, rated on an 11-point scale (0–10 inclusive), which reliably indexes perceived MWL in task-based settings [42].

A three-minute relaxation scene was presented after each condition to minimize physiological carryover effects. The scene featured naturalistic forest visuals, ambient nature sounds, and a brief guided breathing instruction. Such short restorative breaks are known to support affective and physiological recovery following stress-inducing tasks [43]–[45].

To strengthen the perception of social evaluation while avoiding auditory feedback that could introduce EEG artifacts through evoked responses and muscle activity, a visual feedback cue was implemented [46]. A vertical progress bar was displayed during each task to simulate evaluative feedback, and participants were told it reflected their performance relative to pilot participants. The bar updated after each verbal response, creating the impression of continuous assessment. Its feedback rule was identical across participants and enforced a condition-specific bias: the bar remained entirely below the midpoint in the high-stress condition and entirely above it in the low-stress condition. To enhance credibility, the display responded noticeably to performance fluctuations within these bounds, so trajectories varied slightly between participants while adhering to the condition-specific bias.

1) Randomisation, Counterbalancing, and Blinding: The order of the four experimental conditions (2×2 Stress × Workload) was fully counterbalanced across participants, yielding 24 unique condition sequences. Two variants of the high-MWL task were alternated between participants to minimise practice effects, resulting in 48 distinct experimental configurations overall. Condition order and task variant assignment were independently randomised for each participant. The experimenter was aware of the condition but provided no feedback or performance cues during the trials to maintain consistency of instruction and social context across participants. Participants were naïve to the purpose of the manipulations and were only informed that they might complete arithmetic-based tasks during the session, enabling brief pre-task practice without revealing the experimental aims.

C. Apparatus

The experiment was conducted within a virtual reality environment using an HTC Vive Pro Eye headset (combined resolution 2880 × 1600 pixels, refresh rate 90 Hz) with six-degree-of-freedom (6 DoF) positional tracking using Valve Lighthouse 2.0 base stations (SteamVR Tracking 2.0). Physiological signals were recorded using multiple devices: the headset's built-in eye tracker for pupillometry, a Shimmer3 GSR+ for electrodermal activity (EDA), and a Polar H10 heart-rate monitor for heart rate and heart rate variability (HR and HRV). Facial motion data were also captured using a Vive Face Tracker but were not included in the final analysis.

EEG signals were recorded using an ANT Neuro eego mylab system with a 128-channel waveguard$^{\text{TM}}$ net arranged in an equidistant montage referenced to Cz. Data were sampled at 500 Hz. Electrode impedances were typically kept below 30 kΩ per channel, consistent with manufacturer guidance for the saline-based Waveguard$^{\text{TM}}$ system. In line with the study's focus on stress- and MWL-related activity, greater care was taken to optimise contact over frontal and parietal regions. Ideal impedance levels across all 128 electrodes were not always achievable, as setup efficiency and participant comfort were prioritised over complete optimisation, but slightly higher impedances were acceptable given the system's tolerance.

The VR simulation was based on the Entertainment Computing Group's VR-TSST (ECG VR-TSST, Version 2.3), a virtual adaptation of the Trier Social Stress Test (TSST) validated to induce social-evaluative stress [47]. The ECG VR-TSST was obtained from https://github.com/MIEC/vr-tsst and modified extensively to integrate physiological data acquisition and adapt the virtual environment for the present study (see Section II-B for details).

D. Participants

Forty-eight participants were recruited using convenience sampling through university-wide emails and posters at the University of Bath between August and December 2024. All participants were aged 18 years or older, fluent in English, and reported normal or corrected-to-normal vision. Exclusion criteria were a self-reported history of severe motion sickness, limited mobility, epilepsy, recurrent fainting spells, or febrile convulsions in infancy. Eight participants were excluded following automated signal-quality checks across all physiological modalities, leaving 40 participants in the final analysis ($M = 31.0$, $SD = 6.2$; 20 female). All participants provided written informed consent prior to participation, and the study was approved by the University of Bath Ethics Committee (reference: 2614-2411). Each participant received £30 compensation for their time. Experimental sessions took place in an electromagnetically shielded room within the Department of Psychology to minimise electromagnetic interference during data acquisition.

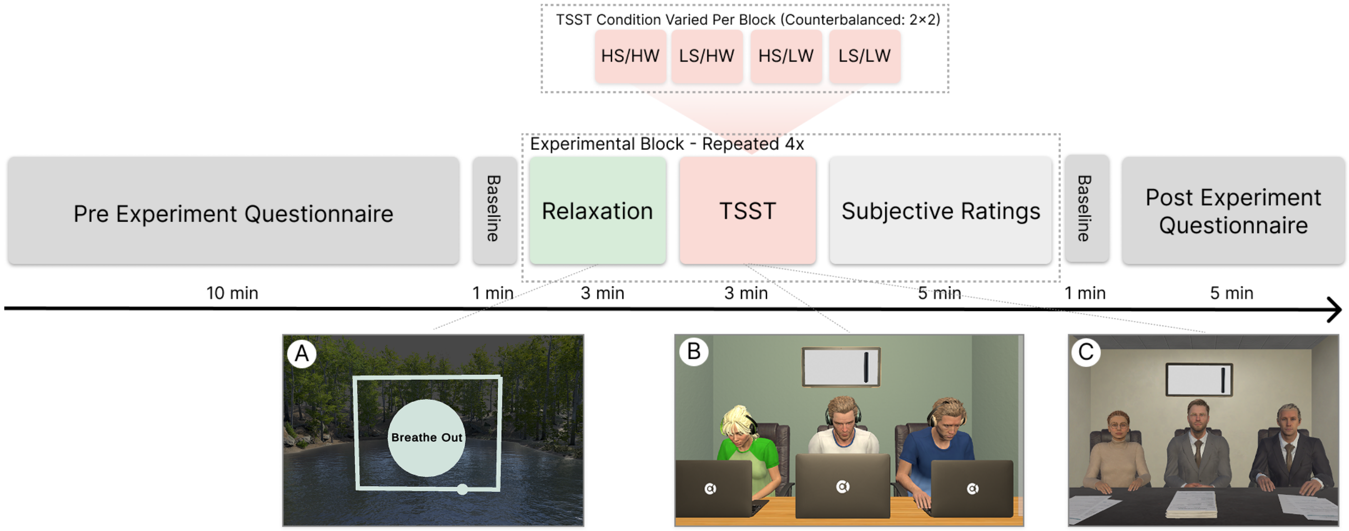
TSST Condition Varied Per Block (Counterbalanced: 2×2)
HS/HW
LS/HW
HS/LW
LS/LW
Experimental Block - Repeated 4x
Pre Experiment Questionnaire
Baseline
Relaxation
TSST
Subjective Ratings
Baseline
Post Experiment Questionnaire
10 min
1 min
3 min
3 min
5 min
1 min
5 min
A
Breathe Out
B
C

Fig. 1: Overview of the experimental protocol. The experimental block was repeated four times with counterbalanced TSST conditions (2×2: stress × workload). Panels (a)–(c) show the relaxation, high-stress, and low-stress scenes, respectively.

E. Procedure

Each session lasted approximately two hours and followed a structured sequence of setup, baseline calibration, four experimental conditions, and debriefing. Before calibration, participants were fitted with EEG (ANT Neuro eego mylab), EDA (Shimmer3 GSR+), and heart rate (Polar H10) sensors. EEG signal quality was verified via impedance checks and live monitoring in the acquisition software, while EDA and HR readings were inspected in Unity to confirm realistic streaming values prior to baseline recording.

A ten-minute calibration and baseline phase was conducted within VR. This included calibration of the eye-tracking system and pupil baseline recording across controlled luminance levels, followed by two one-minute EEG baselines. During the first baseline, participants fixated on a cross presented on a blank background; during the second, they fixated within a neutral virtual room designed as an intermediate environment between the stress and calm scenes used later in the experiment. These baselines provided reference measures for both resting-state and context-specific neural activity.

Each experimental block comprised a 3-minute relaxation scene, a 3-minute arithmetic task, and immediate subjective ratings, with physiological signals recorded continuously throughout. Before each task, the experimenter verbally delivered brief condition-specific instructions and refrained from interaction during the task. This sequence was repeated four times according to the counterbalanced condition order.

After completing all conditions, participants removed the VR headset and sensors, completed final questionnaires, were fully debriefed, and received compensation for their participation.

F. Data Acquisition and Synchronisation

Data from the physiological devices were streamed into Unity at an effective rate of approximately 30 Hz (one sample per rendered frame) and synchronised via the Lab Streaming Layer (LSL). EEG was recorded separately at 500 Hz to an XDF file and later temporally aligned with the Unity-based physiological streams. Data from the wireless devices (Shimmer3 GSR+ and Polar H10) were transmitted via Bluetooth Low Energy (BLE) to a VR-capable desktop workstation ([specifications to be inserted: CPU, GPU, RAM, and operating system]) running the Unity-based recording application. All physiological and behavioural data streams were time-synchronised using the Lab Streaming Layer (LSL; [48]) framework, which provided a unified clock across acquisition systems. All data streams were therefore temporally aligned on the common LSL clock, yielding a unified multimodal dataset for subsequent preprocessing and feature extraction. Although LSL provides convenient software-based synchronisation, minor offsets due to clock drift and network latency typically remain within a few milliseconds. However, this limitation was not considered critical given that all physiological features were computed over relatively long time windows, where sub-millisecond timing differences are negligible.

G. Signal Pre-processing and Feature Extraction

Each modality then underwent a modality-specific sequence of resampling, artifact removal or signal cleaning, and segmentation by experimental condition. Features were subsequently extracted from the cleaned segments and baseline-corrected relative to the preceding relaxation period, as detailed in the following subsections.

1) EEG: EEG data were recorded from 128 electrodes at 500 Hz and processed in MATLAB (EEGLAB v2024.1) following standard high-density EEG preprocessing procedures. Noisy or flat channels were automatically rejected using the EEGLAB *clean_rawdata* algorithm, which employs robust Random Sample Consensus (RANSAC) reconstruction to detect channels poorly correlated with neighbouring sensors or exhibiting excessive line noise. Excluded channels were later interpolated after ICA to preserve spatial consistency. Continuous EEG data were band-pass filtered between 1–49 Hz using a zero-phase finite impulse response (FIR) filter and notch filtered at 50 Hz to remove line noise. Participants 1–7 exhibited a stable 25 Hz artifact, which was attenuated using an additional notch filter. The filtered data were then downsampled to 125 Hz prior to artifact rejection. Independent components were estimated using Adaptive Mixture ICA (AMICA), and artifact classification was performed with ICLabel [49]. Components labeled with $\geq$90% probability as ocular or muscle activity were removed using ICFlag, while other classes (e.g., cardiac, line noise) were retained to minimize over-rejection. Following component removal, previously excluded channels were interpolated using spherical splines and data were re-referenced to the common average. Spectral features were derived from the cleaned, re-referenced EEG using Welch's method with 2 s Hamming windows and 50% overlap, as implemented in EEGLAB and MATLAB. Power spectral density (PSD) was averaged within canonical frequency bands (δ: 1–4 Hz, θ: 4–8 Hz, α: 8–13 Hz, β: 13–30 Hz, γ: 30–49 Hz), and derived metrics including θ/β ratios and spectral entropy were computed for each condition. For each participant and condition, spectral features were averaged across all Welch windows within the 3-minute task segment. Power values were $\log_{10}$-transformed prior to statistical analysis to normalize their distribution.

2) Electrodermal Activity (EDA): Skin-conductance signals were recorded using a Shimmer3 GSR+ sensor and resampled to 10 Hz.n Signals outside the physiological range 0.01–30 μS were considered artefactual. Flat segments longer than 5 s were removed, and short gaps ($\leq$5 s) were linearly interpolated. Signals were detrended and low-pass filtered (Butterworth, $\approx$ 5 Hz) before decomposition into tonic and phasic components following standard procedures [50]. Tonic activity was indexed by the mean and standard deviation of skin-conductance level, and phasic activity by the rate, total area, and total count of skin-conductance responses. All features were baseline-corrected to the preceding relaxation phase.

EEG entropy features.: Two entropy measures were derived from each regional EEG segment to quantify temporal and spectral irregularity.

Sample entropy (SampEn) estimates the unpredictability of EEG amplitude fluctuations as the negative natural logarithm of the conditional probability that two sequences of length m that match within a tolerance r will also match at length m+1 [51]:

$$\mathrm{SampEn}(m, r, N) = -\ln\left(\frac{A}{B}\right), \tag{1}$$

where A and B are the numbers of matching template pairs of length $m{+}1$ and m, respectively. SampEn was computed per channel ($m{=}2$, $\tau{=}1$, $r{=}0.2\sigma_x$) following standard physiological conventions [51]. Channel-wise values were averaged within each anatomical region, omitting non-finite values.

Spectral entropy (SpecEn) quantified the uniformity of the normalized EEG power spectral density (PSD) across 1–30 Hz (excluding 8–13 Hz) [52].

$$\mathrm{SpecEn} = -\sum_i p(f_i)\log_2 p(f_i), \quad p(f_i) = \frac{P(f_i)}{\sum_j P(f_j)}, \tag{2}$$

where $P(f_i)$ is the mean PSD at frequency f_i. Higher values indicate flatter (more irregular) spectra, consistent with the Shannon formulation of informational entropy [53]. Both entropy metrics were averaged across channels within each cortical region to yield a single regional feature per condition.

3) Heart Rate and Heart Rate Variability (HR/HRV): Heart-rate and interbeat-interval data were recorded using a Polar H10 sensor. Values outside physiological ranges (40–220 bpm; 100–2000 ms) were treated as artefactual and corrected using linear or cubic interpolation across breif discontinuities. Ectopic beats were corrected using the Kubios method [54] implemented in *NeuroKit2* [50], followed by a rolling median–absolute–deviation filter for residual outlier removal. Time-domain HRV features (mean HR, SDNN, RMSSD) were derived from the cleaned interbeat intervals, and all metrics were baseline-corrected to the preceding relaxation phase.

4) Pupillometry: Binocular pupil diameter was recorded at an effective rate of 30 Hz using the Vive Pro Eye tracker. Subject-specific luminance calibration was conducted during the baseline phase to estimate each participant's resting pupillary response to varying scene luminance levels. During each condition, instantaneous scene luminance at the participant's gaze point was logged, and pupil diameter was corrected by subtracting the participant's calibrated resting pupil response to that luminance, obtained during the pre-experiment calibration phase. Values outside the range of $\pm$4 mm were discarded, and eye closures were identified from rapid negative velocity and low-magnitude drops in both eyes. Short gaps ($\leq$0.3 s) were linearly interpolated, and the cleaned left and right traces were then averaged and low-pass filtered at 4 Hz (Butterworth). Mean pupil diameter and its standard deviation were computed directly.

5) Covariates and Confounds: Speech rate and head-motion velocity were extracted from VR logs and included as covariates in confirmatory mixed-effects analyses to account for behavioural differences between workload conditions.

H. Data Segmentation and Baseline Correction

Each participant's recording was divided into contiguous segments corresponding to the four experimental conditions and their preceding 3-minute baselines. For every physiological measure, feature values from each condition were baseline-corrected by subtraction of the participant's own preceding relaxation mean. This approach isolates task-related reactivity while controlling for slow physiological drift.

I. Quality Control and Participant Exclusion

Automated quality-control (QC) logs were generated for each participant and modality using modality-specific thresholds applied during preprocessing. For peripheral physiological data, participants were excluded if less than 70% of samples were retained in any condition, sensor, or overall dataset. Specifically, exclusion occurred if data retention for (i) any sensor within a condition, (ii) any condition within a sensor, (iii) across all sensors, or (iv) across all conditions fell below this threshold.

For EEG, participants were excluded if, after preprocessing, less than 75% of data samples were retained, more than 15% of channels were interpolated, or over 35% of independent components were removed. EEG waveform plots were automatically generated and visually inspected by one researcher after each major cleaning step (channel rejection, ASR correction, and ICA component removal) to confirm satisfactory data quality.

In total, eight participants were excluded from analysis based on these criteria: 3 for EEG quality; 5 for physiological data quality. All QC metrics and exclusion decisions were logged for reproducibility.

J. Exploratory Data Analysis

Exploratory analyses assessed the effectiveness of the experimental manipulations and examined relationships between subjective ratings and physiological measures.

1) Subjective Manipulation Checks: Subjective stress and MWL ratings were analysed using 2 × 2 repeated-measures ANOVAs with factors *Stress* (High, Low) and *MWL* (High, Low). Effect sizes are reported as partial η^2 with 95% confidence intervals. All tests were two-tailed.

To assess the independence of subjective dimensions, correlations were computed between stress and MWL ratings within congruent (*High–High*, *Low–Low*) and incongruent (*High–Low*, *Low–High*) condition pairs. A Fisher r-to-z comparison tested whether these correlations differed across congruency types.

K. Factorial Analysis of Physiological Measures

To evaluate how stress and MWL manipulations influenced physiological responses, each canonical feature was analysed using a 2×2 repeated-measures design with Stress (low, high) and MWL (low, high) as within-participant factors. Because residuals from standard repeated-measures ANOVA showed substantial non-normality across features, all inferential tests used the Aligned Rank Transform, which preserves factorial interpretability under non-normal distributions. For each feature, we obtained F-statistics, p-values, and partial eta-squared for the main effects and interaction. Significant effects were followed by ART-compatible post-hoc contrasts to characterise effect patterns.

1) Physiology-Subjective Correlations: To assess how the level of one factor modulated the association between the other factor and physiological features, we computed within-participant correlations between subjective ratings and the canonical physiological feature set. Stress–feature correlations were estimated separately under low and high MWL, and MWL–feature correlations were estimated separately under low and high stress. This stratified approach allowed us to quantify how physiological–subjective relationships changed as a function of concurrent mental state. For each feature in each subset, correlation coefficients and p-values were obtained, and false discovery rate (Benjamini–Hochberg) correction was applied within each stratified analysis.

L. Statistical and Machine-Learning Analyses

1) Feature Integration and Dataset Construction: Condition-level EEG, EDA, HR/HRV, and pupillometry features were merged by participant and condition using synchronized timestamps. The resulting multimodal dataset comprised one feature vector per participant per condition (4 per participant), used in both statistical and machine-learning analyses.

2) Model Training and Cross-Validation: To assess whether physiological features could discriminate stress and MWL states, we trained Support Vector Machine classifiers with radial-basis kernels using a fully nested leave-one-subject-out (LOSO) procedure. Within each outer fold, features were first pruned within the training set to remove near-zero-variance, missing, and highly collinear predictors, and the remaining features were ranked by their univariate association with the target label. Models were trained using the top k features (k = 20, 10, 5).

Hyperparameters (cost and gamma) were tuned within an inner LOSO loop using the training participants only. For each (cost, gamma, k) combination, inner-loop performance was averaged across folds, and the combination with the highest mean AUC was selected. The final model for that outer fold was then refit on the full training set and evaluated on the held-out participant. Performance was quantified using accuracy, F1 score, and area under the ROC curve (AUC), averaged across outer folds for each target.

Support Vector Machine (SVM) models were used to evaluate whether multimodal physiological and EEG features could classify the experimentally defined stress and mental workload (MWL) conditions. Classification was conducted separately for stress (high-stress vs. low-stress) and workload (high-MWL vs. low-MWL). To assess subject-independent generalisability, all analyses employed a leave-one-subject-out (LOSO) evaluation framework, in which each participant served once as the held-out test case.

The models used features derived from EEG power spectral density (PSD), electrodermal activity (EDA), heart rate (HR), heart rate variability (HRV), and pupillometry. All features were baseline-corrected using the pre-condition relaxation period by subtracting each participant's aggregated baseline value from their aggregated task value, and were then z-scored within participant. To prevent information leakage, all preprocessing and feature pruning were performed independently within each LOSO training fold.

Feature pruning followed a multi-step procedure. First, features with near-zero variance were removed. Second, features with high pairwise correlations were excluded using a threshold of $|r| > 0.75$. The remaining features were then rank-ordered by their absolute Pearson correlation with the target label (stress or workload) computed on the training set. Four candidate feature-set sizes—the top 5, 10, 15, and 20 ranked features—were evaluated to identify the smallest feature subset achieving the highest classification performance. This procedure was repeated independently within each LOSO fold.

All models were implemented in R using an SVM with a radial basis function (RBF) kernel. The SVM hyperparameters C and γ were tuned within each fold using a grid search applied only to the training data. Feature scaling was also performed inside each fold using summary statistics computed solely from the training participants. For each target, the tuned model was applied to the held-out participant to obtain unbiased predictions.

Model performance was assessed by aggregating predictions across all LOSO folds. The primary evaluation metric was classification accuracy, computed as the proportion of correctly classified samples across all held-out participants.

M. Ethics, Data Availability, and Reproducibility

All preprocessing and modelling code (MATLAB, Python, and R scripts) is available at https://github.com/JoeBathVR/StressWorkloadPipeline. Anonymised participant-level feature matrices and analysis scripts are available upon reasonable request, in line with University of Bath data-sharing policy. The VR environment was based on the open-source VR-TSST framework (v2.3) and modified under its MIT licence.

All procedures were approved under the same protocol described in Section **??** by the University of Bath Data & Digital Science Research Ethics Committee (reference: 2614-2411) and were conducted in accordance with the Declaration of Helsinki. The study involved mild deception: participants were informed that their performance and speech were being recorded and evaluated, although no such recordings were made. This procedure and its associated risk assessment were reviewed and approved by the ethics committee. Participants were fully debriefed after the experiment and informed that no recordings had occurred and that evaluative feedback had been simulated.

III. Results

A. Participant Characteristics

Forty participants (mean age = 31.0 years, SD = 6.2; 20 female) were included in the final analysis after exclusion of eight due to poor-quality physiological data based on the predefined quality-control criteria described in Section II-I

B. Factorial Analysis of Subjective Measures

As intended, self-reported stress was higher in the high-than low-stress condition (mean difference = 0.86, $t_{39} = 3.2$, $p = 0.003$, $d_z = 0.51$; Fig. 1a). MWL also increased stress ratings ($p < .001$), with no stress × MWL interaction ($p = .55$), indicating additive rather than interactive effects.

MWL manipulations strongly increased NASA-TLX mental demand scores (mean difference = 2.33, $t_{39} = 8.0$, $p < 0.001$, $d_z = 1.27$; Fig. 1b). Stress also raised MWL ratings ($p < 0.001$), and here an interaction was observed ($p = 0.012$), such that stress amplified the subjective burden of high MWL.

Correlations between stress and MWL ratings further indicated partial differentiation: associations were stronger in congruent conditions ($r = .52$, $p < 0.001$) than incongruent conditions ($r = .31$, $p = 0.005$), although this difference was not significant (Fisher's $z = 1.56$, $p = .12$).

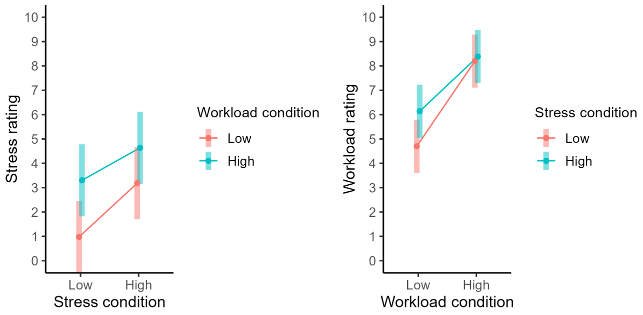

Fig. 2: Subjective stress and workload ratings under 2×2 manipulations. (a) Stress ratings by stress condition (x-axis), with separate lines for workload condition. (b) Workload ratings by workload condition (x-axis), with lines for stress condition. Error bars indicate 95% confidence intervals. Stress ratings are shown as change scores relative to a pre-experiment baseline (-10 = much less stressed than baseline, $+10$ = much more stressed), whereas workload ratings were collected on a 0–10 scale.

C. Factorial Analysis of Physiological Measures

Stress produced a single reliable physiological effect after baseline adjustment. Tonic EDA increased under high stress. No other autonomic, pupil, or EEG measure showed evidence of stress-related modulation, indicating a selective rather than multimodal response. MWL did not reliably influence any physiological or EEG measure. No feature showed a stress × workload interaction. The tonic EDA stress effect was unchanged by MWL level. Post-hoc contrasts confirmed higher tonic EDA under high stress, with this elevation present at both MWL levels. The pattern of consistently higher tonic activity in all high-stress conditions corroborates the robustness of the stress effect.

D. Correlations

Repeated-measures correlations were computed separately within each level of stress and workload to assess whether associations between subjective states and physiological or EEG features varied across conditions.

No modality showed meaningful associations with MWL under either stress level, with correlations remaining uniformly weak and unchanged across autonomic, EEG, and pupil features.

Stress demonstrated a clear and stable pattern: tonic EDA showed a moderate positive association with stress under both workload levels, and this relationship was statistically reliable in each stratum. EDA peak height and heart rate also exhibited positive associations with stress, reaching significance in at least one workload condition. EEG, HRV, and pupil measures did not show reliable stress correlations.

Taken together, autonomic markers—particularly tonic EDA—showed selective and robust coupling with stress, whereas MWL demonstrated no comparable physiological associations. These patterns were consistent across all stress and workload strata, indicating that their presence or absence did not depend on experimental condition.

TABLE I: ANOVA results for subjective stress and NASA-TLX mental demand under stress and workload manipulations. Each row reports $F(1, 41)$, p value, generalized eta squared (η_G^2), condition means, SDs, and 95% CIs where applicable.

Factor	Measure	$F(1, 39)$	p	η_G^2	Condition	Mean	SD	95% CI
Stress	Subjective stress	12.30	.001	.032	High	3.52	2.76	[2.79, 3.96]
					Low	2.52	3.09	[1.87, 3.16]
Workload	Subjective stress	27.40	$< .001$	.101	–	–	–	–
Stress × Workload	Subjective stress	0.837	.366	.002	–	–	–	–
Workload	Mental demand (NASA-TLX)	64.2	$< .001$	.286	High	8.02	1.49	[7.61, 8.27]
					Low	5.69	2.26	[5.28, 6.23]
Stress	Mental demand (NASA-TLX)	19.5	$< .001$	.038	–	–	–	–
Stress × Workload	Mental demand (NASA-TLX)	9.54	.004	.025	–	–	–	–

TABLE II: ART-ANOVA stress main effects for baseline-adjusted physiological and EEG features.

Feature	F	p	η_p^2	sig
HRV RMSSD (ms)	0.090	0.765	0.00	
Heart rate (median, bpm)	1.165	0.283	0.01	
EDA tonic level (mean, μS)	4.339	0.039	0.04	*
EDA peak height (mean, μS)	0.208	0.650	0.00	
Frontal midline θ power (mean)	0.346	0.557	0.00	
Frontal β power (mean)	0.277	0.600	0.00	
Parietal α power (mean)	1.144	0.287	0.01	
Pupil dilation (median, mm)	0.210	0.647	0.00	

TABLE III: ART-ANOVA workload main effects for baseline-adjusted physiological and EEG features.

Feature	F	p	η_p^2	sig
HRV RMSSD (ms)	1.813	0.181	0.02	
Heart rate (median, bpm)	1.149	0.286	0.01	
EDA tonic level (mean, μS)	1.138	0.288	0.01	
EDA peak height (mean, μS)	0.146	0.703	0.00	
Frontal midline θ power (mean)	0.019	0.889	0.00	
Frontal β power (mean)	0.155	0.695	0.00	
Parietal α power (mean)	0.328	0.568	0.00	
Pupil dilation (median, mm)	0.058	0.810	0.00	

TABLE IV: ART-ANOVA interaction effects (stress × workload) for baseline-adjusted physiological and EEG features.

Feature	F	p	η_p^2	sig
HRV RMSSD (ms)	0.002	0.965	0.00	
Heart rate (median, bpm)	0.480	0.490	0.00	
EDA tonic level (mean, μS)	1.062	0.305	0.01	
EDA peak height (mean, μS)	0.00003	0.996	0.00	
Frontal midline θ power (mean)	0.040	0.841	0.00	
Frontal β power (mean)	0.006	0.939	0.00	
Parietal α power (mean)	0.786	0.377	0.01	
Pupil dilation (median, mm)	0.025	0.873	0.00	

TABLE V: Post hoc contrast and cell means for tonic EDA (mean, μS).

Post hoc contrast (Stress main effect)				
Contrast	Estimate	SE	$t(114)$	p
Low – High	-11.33	5.44	-2.08	0.039

Cell means (Stress × Workload)			
Stress	Workload	Mean	SE
Low	Low	0.685	0.128
Low	High	1.205	0.208
High	Low	1.211	0.175
High	High	1.305	0.219

TABLE VI: Best-performing LOSO SVM models for stress and MWL classification.

Target	k (features)	Accuracy	F1	AUC
Stress	20	0.58	0.55	0.59
MWL	5	0.57	0.57	0.56

E. Support Vector Machine Classification

SVM models provided above-chance discrimination for both stress and MWL (Table X). Stress classification performed best with larger feature sets, with accuracy and AUC highest at k = 20. In contrast, MWL classification was strongest with the smallest feature set (k = 5), with larger sets yielding reduced performance. Overall, both states were moderately separable, with differing feature-set sensitivities.

a) Summary: These findings show that stress and workload can be predicted from multimodal physiological data with moderate accuracy and statistically reliable generalisation.

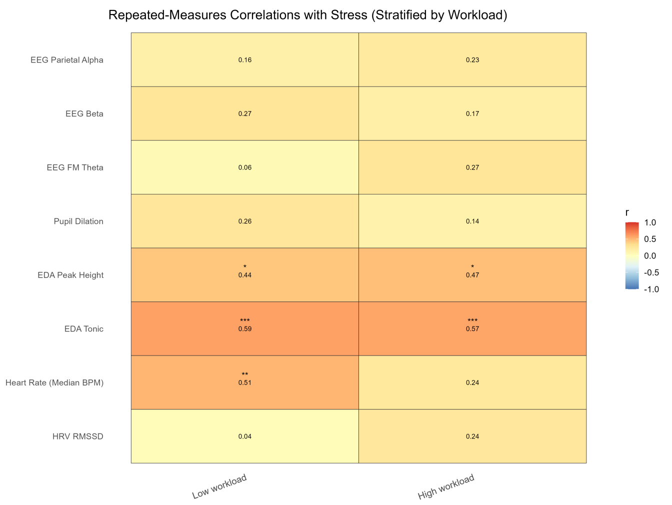

Fig. 3: Stratified correlations between physiological features and subjective stress, shown separately for low- and high-MWL conditions.

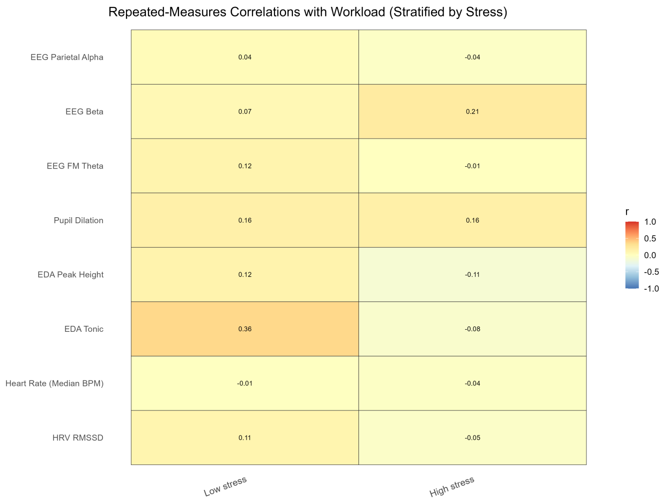

Fig. 4: Stratified correlations between physiological features and subjective MWL, shown separately for low- and high-stress conditions.

IV. Discussion

A. Overview and Key Findings

Stress and MWL are frequently conflated in affective computing research, despite engaging partially distinct physiological systems. This study manipulated stress and MWL within an immersive VR adaptation of the TSST, enabling controlled examination of their overlapping yet partially dissociable effects.

The present results highlight both the promise and the challenges of disentangling stress and mental workload (MWL) via physiology in an ecologically valid VR setting. Subjective measures confirmed that our manipulations successfully elevated the targeted states (high-stress conditions felt more stressful; high MWL conditions felt more demanding), yet they also revealed overlap: stress and MWL ratings were positively correlated and each factor modestly "leaked" into the other's appraisal. Physiologically, we observed a selectively stress-driven response profile. Only tonic electrodermal activity (EDA) showed a reliable main effect of stress (higher under high-stress), whereas no canonical feature showed a significant MWL effect or any stress×MWL interaction. This indicates that at the level of mean responses, our multimodal signals reflected psychosocial stress much more than cognitive workload. Stratified correlation analyses reinforced this pattern: autonomic markers (especially tonic EDA) tracked subjective stress robustly across different workload levels, while no measured feature consistently tracked subjective workload in either stress condition. Finally, subject-independent classification of the two states yielded above-chance but modest accuracy (57–59% correct), underscoring the difficulty of generalizable stress/ MWL detection with generic features. In the following, we interpret these findings in detail, considering their implications for discriminating stress vs. workload and for the design of affective computing systems.

B. Subjective Dissociation

Our paradigm aimed to manipulate stress and MWL as separate factors, and the subjective results demonstrate both dissociation and interdependence. On one hand, participants clearly differentiated the two: high-MWL tasks were rated as more mentally demanding (NASA-TLX) than low-MWL tasks, and high stress scenarios induced higher stress ratings than low-stress scenarios.

There was no interaction on subjective stress: the stress induced by social evaluation was not modulated by task difficulty. In contrast, MWL ratings exhibited a clear interaction, with participants experiencing the arithmetic task as more demanding when it was performed under social-evaluative stress. This pattern is consistent with models proposing that acute stress reduces cognitive efficiency and lowers the threshold at which mental workload is experienced as effortful. Prior work shows that distress can increase the subjective burden of cognitive tasks by impairing working-memory processes and executive control [**?**]; meta-analytic evidence similarly demonstrates that acute stress reliably disrupts working-memory capacity and cognitive flexibility, even when overt performance is preserved through compensatory effort [55]. TSST findings further indicate that acute social stress can attenuate or eliminate typical neural load effects, including reduced differentiation in ERP indices of working-memory operations [**?**]. Together, these results provide a coherent account of why participants in our paradigm perceived the high-MWL task as more demanding when performed under stress: acute social-evaluative threat likely diminished cognitive efficiency, amplifying the subjective cost of an already challenging task.

Importantly, however, the moderate correlation between stress and demand ratings in all conditions indicates only a partial overlap – high MWL did not always produce high stress, nor vice-versa. This partial separation is encouraging from a state-discriminability perspective: participants could distinguish feeling "stressed" from feeling "mentally taxed" to a degree. Yet the correlation also reflects concurrent activation of stress and MWL in our combined paradigm, mirroring the real-world tendency for challenging tasks to carry emotional strain . For affective computing, this implies that purely subjective differentiation of the two states is feasible but imperfect, a ground truth challenge that inevitably carries over to physiological measures.

C. Physiological Dissociation

Despite the strong subjective MWL effect, our physiological features showed surprisingly little sensitivity to workload once baseline differences were removed. Instead, the autonomic nervous system responses were dominated by the stress manipulation. In particular, tonic EDA exhibited a clear elevation under high stress (relative to low stress) across both levels of task difficulty. This finding is consistent with the role of EDA as a indicator of sympathetic arousal: the social-evaluative threat in the high-stress VR scenarios evoked sustained sympathetic activation, which was reflected in higher skin conductance level. Notably, this stress effect in EDA was small-to-moderate in magnitude, indicating that while reliable, it did not involve a dramatic physiological surge for all participants. Other autonomic indices such as heart rate and HRV (RMSSD) did not show significant overall differences between high- and low-stress conditions.

One reason may be that both the low-stress and high-stress conditions involved active arithmetic tasks, which themselves elevate cardiovascular metrics to some extent; any additional impact of stress or MWL may have been subtle and overshadowed by inter-individual variability. Indeed, it is well documented that mental effort and psychological stress produce overlapping patterns in cardiovascular measures (both tend to raise heart rate and suppress HRV) [56], [57], making it difficult for simple between-condition comparisons to isolate each effect. This pattern is consistent with prior work showing that stress can mask physiological signatures of MWL. for instance, differences in GSR observed across workload levels disappear once stress is introduced [28], [58].

Crucially, we found no reliable physiological signature of MWL in this study. High-MWL arithmetic (serial subtraction) did not significantly differ from low-MWL addition on any feature when controlling for the pre-task baseline. This is somewhat surprising, as many controlled studies link increases in MWL to changes in, e.g., frontal midline theta EEG or decreased parietal alpha power, as well as modest rises in EDA or heart rate. In our immersive and stress-modulated context, however, those typical MWL related shifts were absent or masked. The lack of EEG spectral effects is particularly noteworthy. Frontal theta power, often cited as an index of working memory load , did not show a dependable increase with task difficulty, nor did we observe the expected alpha suppression or beta increase. It appears that the presence of a realistic social stressor and other VR elements introduced enough noise or additional arousal to blunt these neural workload indicators. EEG workload indices are often more difficult to detect in immersive VR due to motion-related and EMG artifacts; high-frequency components from scalp tension often require suppression, inadvertently reducing sensitivity to MWL-related neural activity [59].

Similarly, pupil diameter did not differ between easy and hard tasks after baseline correction, possibly due to the competing influence of stress-induced arousal. The overall physiological pattern thus suggests that the stress manipulation elicited a fairly specific autonomic response (tonic EDA), whereas the MWL manipulation, as implemented, did not elicit a consistent peripheral or EEG response.

D. Stratified Correlations

By examining within-condition correlations, we further probed whether physiology–state relationships hold regardless of the other factor. These analyses confirmed that subjective stress had consistent autonomic correlates, whereas subjective workload remained physiologically elusive. Tonic EDA stood out as the most reliable individual indicator of stress levels: within both low-MWL and high-MWL scenarios, participants with higher self-reported stress tended to have higher skin conductance. This held true under both task difficulties (with correlation coefficients in the moderate range), suggesting that the coupling between EDA and the experience of psychosocial stress was robust to variations in MWL. In other words, even when participants were performing a demanding task, those who were more stressed showed the expected EDA increases, a promising result for using EDA in naturalistic stress monitoring.

We also observed that phasic EDA features (e.g. peak amplitudes) and heart rate exhibited positive correlations with stress in at least some strata. Notably, heart rate's correlation with stress was stronger in low-MWL conditions and weakened when workload was high. This trend implies an interaction: under low MWL, an elevated heart rate is more clearly a sign of stress, whereas under high MWL, even non-stressed individuals may have elevated heart rates, compressing the range and diluting the correlation. Such an effect underscores how concurrent MWL can confound stress inference from certain signals. In contrast, both tonic and phasic EDA appears less susceptible to this particular confound in our data, since these features remained responsive to stress irrespective of MWL level.

Perhaps the most striking finding is the near-zero correlation between physiological features and subjective MWL ratings in any stratification. Neither autonomic measures (EDA, heart metrics, pupil size) nor the EEG band powers showed any meaningful relationship to how mentally demanding participants found the task, whether under calm or stressed conditions. This null result reinforces the earlier point: the physiological manifestations of MWL in this scenario were too subtle or too entwined with stress to be separately detectable.

E. Implications for State Discrimination and Physiological Computing

F. Relation to Prior Work

These results are consistent with prior evidence that subject-independent models perform substantially worse than personalised ones. [22] simultaneously classified stress and workload across both levels of the other state using EEG and peripheral measures, reporting mean cross-participant accuracies of approximately 58 % under leave-one-subject-out validation and without transfer learning. This design, in which each state varied while the other also fluctuated, highlights the difficulty of generalising across individuals when multiple cognitive–affective dimensions interact. [20] obtained comparable accuracies when modelling stress and workload jointly in a 2×3 (threat × n-back) design using fNIRS and ECG—43–47 % for three-level workload (chance = 33 %) and roughly 62 % for two-level stress (chance = 50 %)—again reflecting the limits of cross-participant generalisation even under controlled factorial manipulations.

G. Methodological and Conceptual Limitations

Methodological Limitations

We acknowledge several methodological constraints that must be considered when interpreting our findings and planning future work. Because we did not record respiratory signals, our HRV metrics may include respiratory influences in addition to autonomic changes, which could reduce the specificity of our inferences about sympathetic/parasympathetic activation. Although independent-component-analysis (ICA)-based artefact rejection substantially reduced ocular and muscular contamination in the EEG signals, residual electromyography (EMG) activity—especially in the frontal β-bands—may still have influenced our neural-feature estimates; incorporating dedicated EMG channels in subsequent studies would enable more precise isolation of cortical vs. muscular sources.

On the subjective-data side, our reliance on single post-condition self-reports limits the temporal resolution at which we capture fluctuations in perceived stress and MWL, weakening our ability to align subjective dynamics with physiological changes and thereby reducing the sensitivity of our models to moment-to-moment variation. Moreover, the subtlety of our manipulations resulted in low variance in subjective ratings, and the absence of a genuine neutral baseline condition means even our “low–stress/low MWL” state likely involved mild activation—both factors diminish label contrast and may hamper model learnability, reducing model robustness when applied to more diverse or extreme conditions. Additionally, each participant contributed only approximately 12 minutes of data, limiting the total training-data volume, constraining within-participant variability and feature stability over time, and potentially causing our models to under-perform once deployed in longer or more varied immersive settings.

From a modelling and validation perspective, we performed feature-pruning and hyper-parameter tuning outside of a fully nested cross-validation (CV) framework. Although we avoided target-label leakage, this approach may still introduce a mild optimistic bias in performance estimates; future research should adopt fully nested optimisation and bootstrap confidence intervals to provide more rigorous uncertainty quantification and guard against over-estimation of generalisability. Despite these constraints, this study offers a valuable initial demonstration of multimodal physiological detection of stress and MWL in immersive contexts, and the insights we gained—regarding feature behaviour, modality synergy and pilot model feasibility— remain valid within the defined scope. Addressing the outlined limitations in follow-up work will enhance the reliability, external validity and practical applicability of adaptive affective-computing systems in real-world settings.

Conceptual Limitations.

The present analysis focused on static inference rather than adaptive feedback; model interpretability and generalisation were evaluated offline, without real-time environmental adaptation. Although stress and MWL were both varied, these factors may not capture the full spectrum of cognitive–affective interactions that occur in naturalistic settings, where demands, motivation, and affect fluctuate continuously. Extending this framework to continuous, context-varying, and closed-loop environments will be essential for establishing causal validity, enhancing ecological generalisability, and advancing toward practical adaptive systems.

H. Theoretical and Applied Implications

The present results indicate that stress and MWL are distinguishable using multimodal signals, with EEG features providing most of the predictive information and autonomic cues offering complementary value. Future work should broaden ground truth beyond laboratory stressors and test generalisability across task structures and individuals, using time-resolved labels and added behavioural/respiratory measures where appropriate

V. Conclusion

This study examined stress and MWL in a combined VR-TSST with multimodal sensing and subject-independent modelling using LOSO validation. Methodologically, the combined design helps separate stress from MWL relative to single-factor paradigms, and attribution analyses offer practical guidance for feature selection while cautioning against interpreting peripheral signals as state-specific. Practically, . Limitations include the absence of respiration data, brief recording duration, and limited temporal resolution of subjective ratings. Future work should extend recording time, adopt time-resolved labelling, and evaluate generalisability across task structures, frequency bands, and individuals.

Acknowledgments

The authors thank [NAMES / institutions] for their contributions and support.

References

[1] S. H. Fairclough, “Fundamentals of physiological computing,” *Interacting with Computers*, vol. 21, no. 1, pp. 133–145, Jan. 2009.

[2] F. Dehais, A. Lafont, R. Roy, and S. Fairclough, “A Neuroergonomics Approach to Mental Workload, Engagement and Human Performance,” *Frontiers in Neuroscience*, vol. 14, Apr. 2020.

[3] M. Alimardani and K. Hiraki, “Passive Brain-Computer Interfaces for Enhanced Human-Robot Interaction,” *Frontiers in Robotics and AI*, vol. 7, p. 125, Oct. 2020.

[4] J. A. Mark, A. Curtin, A. E. Kraft, M. D. Ziegler, and H. Ayaz, “Mental workload assessment by monitoring brain, heart, and eye with six biomedical modalities during six cognitive tasks,” *Frontiers in Neuroergonomics*, vol. 5, p. 1345507, Mar. 2024.

[5] P. Aricò, G. Borghini, G. Di Flumeri, A. Colosimo, S. Bonelli, A. Golfetti, S. Pozzi, J.-P. Imbert, G. Granger, R. Benhacene, and F. Babiloni, "Adaptive Automation Triggered by EEG-Based Mental Workload Index: A Passive Brain-Computer Interface Application in Realistic Air Traffic Control Environment," *Frontiers in Human Neuroscience*, vol. 10, Oct. 2016.

[6] Q. Meteier, M. Capallera, S. Ruffieux, L. Angelini, O. Abou Khaled, E. Mugellini, M. Widmer, and A. Sonderegger, "Classification of Drivers' Workload Using Physiological Signals in Conditional Automation," *Frontiers in Psychology*, vol. 12, Feb. 2021.

[7] F. Dehais, S. Ladouce, L. Darmet, T.-V. Nong, G. Ferraro, J. Torre Tresols, S. Velut, and P. Labedan, "Dual Passive Reactive Brain-Computer Interface: A Novel Approach to Human-Machine Symbiosis," *Frontiers in Neuroergonomics*, vol. 3, Apr. 2022.

[8] G. Masi, G. Amprimo, C. Ferraris, and L. Priano, "Stress and Workload Assessment in Aviation—A Narrative Review," *Sensors (Basel, Switzerland)*, vol. 23, no. 7, p. 3556, Mar. 2023.

[9] H. Liu, Y. Zhang, Y. Li, and X. Kong, "Review on Emotion Recognition Based on Electroencephalography," *Frontiers in Computational Neuroscience*, vol. 15, 2021.

[10] K. Huttunen, H. Keränen, E. Väyrynen, R. Pääkkönen, and T. Leino, "Effect of cognitive load on speech prosody in aviation: Evidence from military simulator flights," *Applied Ergonomics*, vol. 42, no. 2, pp. 348–357, Jan. 2011.

[11] T. O. Zander and C. Kothe, "Towards passive brain-computer interfaces: Applying brain-computer interface technology to human-machine systems in general," *Journal of Neural Engineering*, vol. 8, no. 2, p. 025005, Apr. 2011.

[12] V. D. Novak, D. Hass, M. S. Hossain, A. F. Sowers, and J. D. Clapp, "Effects of adaptation accuracy and magnitude in affect-aware difficulty adaptation for the multi-attribute task battery," *International Journal of Human-Computer Studies*, vol. 183, p. 103180, Mar. 2024.

[13] K. Forbes-Riley and D. Litman, "Using Performance Trajectories to Analyze the Immediate Impact of User State Misclassification in an Adaptive Spoken Dialogue System," in *Proceedings of the SIGDIAL 2011 Conference*, J. Y. Chai, J. D. Moore, R. J. Passonneau, and D. R. Traum, Eds. Portland, Oregon: Association for Computational Linguistics, Jun. 2011, pp. 216–226.

[14] M. S. Hossain, A. F. Sowers, J. D. Clapp, and V. D. Novak, "Different adaptation error types in affective computing have different effects on user experience: A Wizard-of-Oz study," *International Journal of Human-Computer Studies*, vol. 196, p. 103440, Feb. 2025.

[15] C. Holland, G. Perry, and H. F. Neyedli, "Calibrating Trust, Reliance and Dependence in Variable-Reliability Automation," *Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting. Human Factors and Ergonomics Society. Annual Meeting*, vol. 68, no. 1, pp. 604–610, Sep. 2024.

[16] G. Demirezen, T. Taşkaya Temizel, and A.-M. Brouwer, "Reproducible machine learning research in mental workload classification using EEG," *Frontiers in Neuroergonomics*, vol. 5, Apr. 2024.

[17] G. Vos, K. Trinh, Z. Sarnyai, and M. R. Azghadi, "Generalizable machine learning for stress monitoring from wearable devices: A systematic literature review," *International Journal of Medical Informatics*, vol. 173, p. 105026, May 2023.

[18] M. Benchekroun, P. E. Velmovitsky, D. Istrate, V. Zalc, P. P. Morita, and D. Lenne, "Cross Dataset Analysis for Generalizability of HRV-Based Stress Detection Models," *Sensors*, vol. 23, no. 4, p. 1807, Jan. 2023.

[19] P. Prajod, B. Mahesh, and E. André, "Stressor Type Matters! – Exploring Factors Influencing Cross-Dataset Generalizability of Physiological Stress Detection," May 2024.

[20] M. Parent, V. Peysakhovich, K. Mandrick, S. Tremblay, and M. Causse, "The diagnosticity of psychophysiological signatures: Can we disentangle mental workload from acute stress with ECG and fNIRS?" *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, vol. 146, pp. 139–147, Dec. 2019.

[21] M. Bagheri and S. D. Power, "EEG-based detection of mental workload level and stress: The effect of variation in each state on classification of the other," *Journal of Neural Engineering*, vol. 17, no. 5, p. 056015, Oct. 2020.

[22] ——, "Simultaneous Classification of Both Mental Workload and Stress Level Suitable for an Online Passive Brain–Computer Interface," *Sensors (Basel, Switzerland)*, vol. 22, no. 2, p. 535, Jan. 2022.

[23] B. Reimer and B. Mehler, "The Impact of Cognitive Workload on Physiological Arousal in Young Adult Drivers: A Field Study and Simulation Validation," Rochester, NY, Jan. 2011.

[24] B. Mehler, B. Reimer, and J. F. Coughlin, "Physiological Reactivity to Graded Levels of Cognitive Workload across Three Age Groups: An On-Road Evaluation," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 54, no. 24, pp. 2062–2066, Sep. 2010.

[25] I. S. C. Man, R. Shao, W. K. Hou, S. Xin Li, F. Y. Liu, M. Lee, Y. K. Wing, S.-Y. Yau, and T. M. C. Lee, "Multi-systemic evaluation of biological and emotional responses to the Trier Social Stress Test: A meta-analysis and systematic review," *Frontiers in Neuroendocrinology*, vol. 68, p. 101050, Jan. 2023.

[26] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on Psychological Stress Detection Using Biosignals," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 440–460, Jan. 2022.

[27] M. Causse, E. Lepron, K. Mandrick, V. Peysakhovich, I. Berry, D. Callan, and F. Rémy, "Facing successfully high mental workload and stressors: An fMRI study," *Human Brain Mapping*, vol. 43, no. 3, pp. 1011–1031, Nov. 2021.

[28] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Troster, and U. Ehlert, "Discriminating Stress From Cognitive Load Using a Wearable EDA Device," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 410–417, Mar. 2010.

[29] C. Mühl, C. Jeunet, and F. Lotte, "EEG-based workload estimation across affective contexts," *Frontiers in Neuroscience*, vol. 8, Jun. 2014.

[30] N. M. Ehrhardt, J. Fietz, J. Kopf-Beck, N. Kappelmann, and A.-K. Brem, "Separating EEG correlates of stress: Cognitive effort, time pressure, and social-evaluative threat," *European Journal of Neuroscience*, vol. 55, no. 9-10, pp. 2464–2473, 2022.

[31] G. Vanhollebeke, M. Kappen, R. De Raedt, C. Baeken, P. van Mierlo, and M.-A. Vanderhasselt, "Effects of acute psychosocial stress on source level EEG power and functional connectivity measures," *Scientific Reports*, vol. 13, p. 8807, May 2023.

[32] D. Pei, S. Tirumala, K. T. Tun, A. Ajendla, and R. Vinjamuri, "Identifying neurophysiological correlates of stress," *Frontiers in Medical Engineering*, vol. 2, Oct. 2024.

[33] F. Gioia, M. A. Pascali, A. Greco, S. Colantonio, and E. P. Scilingo, "Discriminating Stress From Cognitive Load Using Contactless Thermal Imaging Devices," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Nov. 2021, pp. 608–611.

[34] W. K. Goodman, J. Janson, and J. M. Wolf, "Meta-analytical assessment of the effects of protocol variations on cortisol responses to the Trier Social Stress Test," *Psychoneuroendocrinology*, vol. 80, pp. 26–35, Jun. 2017.

[35] G. Vanhollebeke, S. De Smet, R. De Raedt, C. Baeken, P. van Mierlo, and M.-A. Vanderhasselt, "The neural correlates of psychosocial stress: A systematic review and meta-analysis of spectral analysis EEG studies," *Neurobiology of Stress*, vol. 18, p. 100452, May 2022.

[36] E. C. Helminen, M. L. Morton, Q. Wang, and J. C. Felver, "Stress Reactivity to the Trier Social Stress Test in Traditional and Virtual Environments: A Meta-Analytic Comparison," *Biopsychosocial Science and Medicine*, vol. 83, no. 3, p. 200, Apr. 2021.

[37] T. Ye, R. Elliott, M. McFarquhar, and W. Mansell, "The impact of audience dynamics on public speaking anxiety in virtual scenarios: An online survey," *Journal of Affective Disorders*, vol. 363, pp. 420–429, Oct. 2024.

[38] U. S. Wiemers, D. Schoofs, and O. T. Wolf, "A friendly version of the Trier Social Stress Test does not activate the HPA axis in healthy men and women," *Stress*, vol. 16, no. 2, pp. 254–260, Mar. 2013.

[39] J. Guez, R. Saar-Ashkenazy, E. Keha, and C. Tiferet-Dweck, "The Effect of Trier Social Stress Test (TSST) on Item and Associative Recognition of Words and Pictures in Healthy Participants," *Frontiers in Psychology*, vol. 7, Apr. 2016.

[40] H. M. Sokolowski, Z. Hawes, and D. Ansari, "The neural correlates of retrieval and procedural strategies in mental arithmetic: A functional neuroimaging meta-analysis," *Human Brain Mapping*, vol. 44, no. 1, pp. 229–244, 2023.

[41] A. J. Littman, E. White, J. A. Satia, D. J. Bowen, and A. R. Kristal, "Reliability and Validity of 2 Single-Item Measures of Psychosocial Stress," *Epidemiology*, vol. 17, no. 4, p. 398, Jul. 2006.

[42] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," in *Advances in Psychology*, ser. Human Mental Workload, P. A. Hancock and N. Meshkati, Eds. North-Holland, Jan. 1988, vol. 52, pp. 139–183.

[43] S. H. M. Chan, L. Qiu, G. Esposito, K. P. Mai, K.-P. Tam, and J. Cui, "Nature in virtual reality improves mood and reduces stress: Evidence from young adults and senior citizens," *Virtual Reality*, pp. 1–16, Nov. 2021.

[44] P. Albulescu, I. Macsinga, A. Rusu, C. Sulea, A. Bodnaru, and B. T. Tulbure, "”Give me a break!" A systematic review and meta-analysis on

the efficacy of micro-breaks for increasing well-being and performance," *PLOS ONE*, vol. 17, no. 8, p. e0272460, Aug. 2022.

[45] S. Kumpulainen, S. Esmaeilzadeh, and A. J. Pesola, "Assessing the well-being benefits of VR nature experiences on group: Heart rate variability insights from a cross-over study," *Journal of Environmental Psychology*, vol. 97, p. 102366, Aug. 2024.

[46] E. Ahissar, S. Nagarajan, M. Ahissar, A. Protopapas, H. Mahncke, and M. M. Merzenich, "Speech comprehension is correlated with temporal response patterns recorded from auditory cortex," *Proceedings of the National Academy of Sciences*, vol. 98, no. 23, pp. 13 367–13 372, Nov. 2001.

[47] S. Liszio, "Relaxation, Distraction, and Fun: Improving Well-being in Situations of Acute Emotional Distress with Virtual Reality," Sep. 2021.

[48] C. Kothe, S. Y. Shirazi, T. Stenner, D. Medine, C. Boulay, M. I. Grivich, T. Mullen, A. Delorme, and S. Makeig, "The Lab Streaming Layer for Synchronized Multimodal Recording," p. 2024.02.13.580071, Feb. 2024.

[49] L. Pion-Tonachini, K. Kreutz-Delgado, and S. Makeig, "ICLabel: An automated electroencephalographic independent component classifier, dataset, and website," *NeuroImage*, vol. 198, pp. 181–197, Sep. 2019.

[50] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen, "NeuroKit2: A Python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 1689–1696, Aug. 2021.

[51] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 278, no. 6, pp. H2039–H2049, Jun. 2000.

[52] T. Inouye, K. Shinosaki, H. Sakamoto, S. Toi, S. Ukai, A. Iyama, Y. Katsuda, and M. Hirano, "Quantification of EEG irregularity by use of the entropy of the power spectrum," *Electroencephalography and Clinical Neurophysiology*, vol. 79, no. 3, pp. 204–210, Sep. 1991.

[53] C. E. Shannon, "A Mathematical Theory of Communication."

[54] J. A. Lipponen and M. P. Tarvainen, "A robust algorithm for heart rate variability time series artefact correction using novel beat classification," *Journal of Medical Engineering & Technology*, vol. 43, no. 3, pp. 173–181, Apr. 2019.

[55] G. S. Shields, M. A. Sazma, and A. P. Yonelinas, "The Effects of Acute Stress on Core Executive Functions: A Meta-Analysis and Comparison with Cortisol," *Neuroscience and biobehavioral reviews*, vol. 68, p. 651, Sep. 2016.

[56] H.-G. Kim, E.-J. Cheon, D.-S. Bai, Y. H. Lee, and B.-H. Koo, "Stress and Heart Rate Variability: A Meta-Analysis and Review of the Literature," *Psychiatry Investigation*, vol. 15, no. 3, p. 235, Mar. 2018.

[57] R. L. Charles and J. Nixon, "Measuring mental workload using physiological measures: A systematic review," *Applied Ergonomics*, vol. 74, pp. 221–232, Jan. 2019.

[58] D. Conway, I. Dick, Z. Li, Y. Wang, and F. Chen, "The Effect of Stress on Cognitive Load Measurement," in *Human-Computer Interaction – INTERACT 2013*, P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, and M. Winckler, Eds. Berlin, Heidelberg: Springer, 2013, pp. 659–666.

[59] C. Tremmel, C. Herff, T. Sato, K. Rechowicz, Y. Yamani, and D. J. Krusienski, "Estimating Cognitive Workload in an Interactive Virtual Reality Environment Using EEG," *Frontiers in Human Neuroscience*, vol. 13, p. 401, Nov. 2019.