

# VR-FOCUS: Investigating eye tracking during a virtual reality N-back task as a predictor of cognitive load in chronic pain

Jordan Tsigarides

[j.tsigarides@uea.ac.uk](mailto:j.tsigarides@uea.ac.uk)

University of East Anglia

Jennifer Bowler

University of East Anglia

Jack Dainty

University of East Anglia

Anthony Bagnall

University of Southampton

Alexander Macgregor

University of East Anglia

---

## Article

### Keywords:

**Posted Date:** February 9th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-8680786/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** Competing interest reported. J.T. holds an honorary Chief Medical Officer role at Revolve Labs Ltd. Revolve Labs Ltd provided financial support for J.T.'s PhD programme during which this study was conducted. The company had no role in the study design, data collection, analysis, interpretation, manuscript preparation, or decision to submit for publication. The other authors declare no competing interests.

---

**Version of Record:** A version of this preprint was published at Scientific Reports on October 14th, 2026. See the published version at <https://doi.org/10.1038/s41598-026-74975-6>.



# Abstract

Immersive virtual reality (VR) is a promising medium for adaptive pain therapeutics, but objective markers of cognitive load suitable for real-time adaptation remain insufficiently characterised in people living with chronic pain. This study evaluated whether eye tracking embedded within a consumer-available VR system provides signatures of cognitive load during a VR N-back task in healthy controls and a pragmatic chronic musculoskeletal pain cohort. A total of 84 participants (42/group) completed five levels (Baseline to 4-Back) while ocular responses were recorded. Performance declined and subjective workload increased with higher N-back levels, confirming successful manipulation of cognitive demand. In mixed-effects models, larger pupil diameter and higher blink rate were associated with higher task level and workload ratings after covariate adjustment. Logistic regression using summary pupil and blink features showed limited discrimination of load (1-Back vs 4-Back), with near-chance accuracies (0.51–0.60). In contrast, time-series classifiers exploiting temporal structure achieved higher participant-level accuracy in healthy controls (0.81–0.87) and in chronic pain (0.60–0.66) using pupil diameter alone. Adding blinks produced small, reproducible improvements in model accuracy for the chronic pain group. These findings support VR-embedded eye tracking for cognitive load estimation, but suggest closed-loop applications in chronic pain will require personalised models.

## Introduction

Chronic pain is one of the leading causes of disability worldwide, affecting 1.5 billion globally and posing a substantial socioeconomic and healthcare burden<sup>1, 2</sup>. The experience of chronic pain is heterogeneous and complex, with symptom variability driven by a dynamic interplay between biological, psychological, and social factors such as mood, sleep, and occupation<sup>3</sup>. This variability contributes to considerable day-to-day fluctuations in function and perceived pain intensity, making chronic pain particularly difficult to treat through conventional means<sup>4, 5</sup>. Although pharmacological management remains common despite international recommendations, effectiveness is limited and side effects can be significant<sup>6–9</sup>. There is therefore an urgent need for non-pharmacological and personalised treatment strategies.

Virtual reality (VR) therapeutics have emerged as a promising non-pharmacological option in chronic pain, offering an immersive and engaging medium for the delivery of behavioural and cognitive interventions<sup>10, 11</sup>. Through the use of a head-mounted display (HMD), users are transported into interactive, computer-generated environments with the potential to modulate attention, reduce pain perception, and enhance therapeutic engagement. While early studies have shown encouraging results in chronic pain<sup>12–15</sup>, the content and configuration of VR interventions are rarely personalised. Many existing applications rely on fixed, pre-defined virtual environments and activities, often using a ‘one-size-fits-all’ approach. This lack of personalisation and dynamic content adaptation fails to address the complex, variable nature of symptoms and cognitive capacity in those living with chronic pain<sup>16</sup>.

Modern VR platforms allow users to select different virtual environments, difficulty levels, or task types, thereby providing limited personalisation. However, this relies on users’ self-assessment of their own

cognitive and emotional state within each session, an inherently subjective process that may result in sub-optimal engagement or even fatigue. A more sophisticated approach would be to develop closed-loop systems in which the VR content dynamically adapts to the user's ongoing physiological or cognitive state<sup>17</sup>. Such systems could use real-time biosignals to identify markers of cognitive load or emotional engagement, enabling automated adjustments to maintain an optimal therapeutic state<sup>18</sup>.

Cognitive load represents a particularly promising target for such closed-loop modulation. The concept of flow<sup>19</sup>; a state of deep, effortless engagement occurring at an optimal level of challenge, has been linked to enhanced learning, performance, and immersive flow-like states have also been associated with reduced pain perception in engaging tasks<sup>20–22</sup>. Conversely, excessive or insufficient cognitive load may limit immersion and diminish therapeutic benefit. Accurately quantifying cognitive load during VR experiences could therefore form the foundation of adaptive, personalised VR therapeutics for pain.

Among candidate biosensors, eye tracking holds unique promise for integration with VR. Modern VR systems increasingly incorporate embedded eye-tracking for gaze-based interaction, foveated rendering, and device calibration, providing continuous, high-frequency measurements of ocular metrics without additional hardware burden<sup>23, 24</sup>. In non-VR research, pupil dilation and blink suppression have been consistently associated with higher cognitive load across a range of controlled tasks<sup>25–27</sup>. Despite this, few studies have examined whether these relationships persist within immersive VR environments<sup>28, 29</sup>, and none to our knowledge have specifically investigated this in chronic musculoskeletal pain, where fatigue, medication use, pain and altered attentional allocation may influence oculomotor behaviour<sup>30</sup>.

To unlock the potential of such biosignals for adaptive therapeutics, machine learning is critical. Classical statistical approaches are typically limited to low-dimensional, summary metrics and cannot capture the rich temporal structure inherent in eye-tracking data<sup>31</sup>. Recent advances in time-series approaches, including convolutional<sup>32</sup>, interval-based<sup>33</sup> and ensemble<sup>34</sup> architectures, now allow high-resolution physiological signals to be modelled with substantially greater fidelity. Toolkits such as AEON<sup>35</sup> have further accelerated progress by providing efficient, scalable implementations of state-of-the-art classifiers designed for time-series data. These developments make it increasingly feasible to infer internal cognitive states from complex ocular dynamics in near real time. Importantly, classifiers that can detect cognitive load directly from eye-tracking patterns, without requiring users to provide subjective ratings or behavioural inputs, offer particular value for chronic pain populations where fatigue and symptom burden may limit the reliability or feasibility of repeated self-report<sup>36</sup>. Analysing eye-tracking data with machine learning approaches represents a critical step towards real-time, closed-loop VR systems capable of autonomously recognising and responding to fluctuations in cognitive load.

Building on this rationale, the VR-FOCUS study aims to provide an essential foundation for future closed-loop pain therapeutics by investigating the relationship between eye-tracking features and cognitive load in both healthy controls and a pragmatic, real-world sample of individuals with chronic musculoskeletal pain. Using consumer-available VR technology and a VR N-Back paradigm, this study combines

subjective cognitive load measures with objective eye-tracking and performance data, analysed through both conventional and machine learning approaches, to explore VR-based eye tracking as a measure of cognitive load. We hypothesised that eye-tracking features would reliably predict high versus low cognitive load during VR N-back tasks, including within a clinically heterogeneous chronic pain population, and that these signatures could feasibly be used to inform future closed-loop VR therapeutics.

## Results

A total of 84 participants were included with equal numbers in the healthy control and chronic pain groups (n=42 per group). Table 1 summarises participant characteristics between groups. For those with previous VR exposure (n=56, 67%), most reported rare or one-off use (n=47, 84%).

Most participants in the chronic pain group had complex multimorbidity, with 93% (n=39) reporting two or more long-term conditions and 95% (n=40) taking at least two regular medications. The most common diagnoses were inflammatory arthritis such as Rheumatoid or Psoriatic Arthritis (n=35, 83%), Osteoarthritis (n=21, 50%), Fibromyalgia Syndrome (n=12, 29%) and autoimmune connective tissue disease (n=6, 14%). 26% (n=11) reported an eye condition such as dry eyes or Sjogren's syndrome. Opioid use was reported in 45% (oral morphine equivalent  $2.11 \pm 6.49$ mg, range 1.5-40mg daily). Regular eye-drop use (n=9, 21%), medications with anticholinergic effects (n=12, 29%), and selective serotonin reuptake inhibitors (SSRIs) or serotonin-norepinephrine reuptake inhibitors (SNRIs) (n=13, 31%) were also frequent. In contrast, only 33% (n=14) of the healthy control group were diagnosed with a medical condition (none of which caused regular pain), with most on no regular medication or a single medication.

Baseline symptom burden in the chronic pain group was moderate, with mean pain and fatigue scores of  $3.9 \pm 1.8$  and  $5.6 \pm 2.4$  respectively. Only four participants (10%) reported severe pain (NRS  $\geq 7/10$ ). PROMIS Pain Intensity T-scores were mildly elevated relative to population norms ( $51.7 \pm 5.9$ ), whereas Pain Interference was in the moderate range ( $61.6 \pm 8.3$ ). Pain catastrophising levels were modest overall, with mean PCS scores ( $20.1 \pm 13.5$ ) below commonly used thresholds for high catastrophising ( $>30$ ).

### Descriptive Task-level Outcomes Across VR N-Back Levels

#### Performance Indices

Performance declined with increasing task level in both groups, with the steepest reductions occurring between 1-Back and 3-Back (Figure 1).

When comparing the groups, healthy controls achieved more correct responses ("hits") at all task levels (Figure 1, Panel A). Ability to discriminate between true targets and false alarms ( $d'$ ) showed a similar pattern, except at the 3-Back level (Figure 1, Panel B). Response bias ( $c$ ) became more conservative with

increasing task load in both groups. The chronic pain group demonstrated more conservative bias across 1-Back through 3-Back, with higher bias at lower task load. (Figure 1, Panel C).

Mean reaction times for both “hits” and “false alarms” were faster in the healthy control group (difference in means 36.7ms and 17.5ms, respectively), with mean “hit” reaction times being faster than with “false alarms” overall.

## **Participant-Reported Cognitive Load and Post-Task Measures**

Similar patterns were observed across all self-reported post-task measures of cognitive load, with the NASA-TLX total, mental demand subscale and PAAS measures increasing sharply from Baseline through 3-Back before stabilising (Figure 2, Panels A-C).

Group differences were uniformly small across all levels and measures, with similar reporting of cognitive load in both the healthy control and chronic pain groups.

Perceived difficulty maintaining task focus increased with the task level. Clear between-group differences emerged from 3-Back onwards, with participants in the chronic pain group reporting greater difficulty with the higher N-back levels.

Motivation and engagement ratings remained high across all task levels, with median scores above 90/100 from 1-Back onwards in both groups.

## **Pupil Diameter and Blink Rate**

Mean pupil diameter increased from Baseline to 2-Back in both groups and then remained stable through 3-Back and 4-Back (Panel A-B, Figure 3). Blink rate showed a similar early increase in both groups, but whereas it continued to rise across higher task levels in the chronic pain group, it plateaued from 2-Back onwards in healthy controls (Panel C, Figure 3).

### **Relationship Between Eye-tracking Features and Cognitive Load**

Mixed-effects models showed that higher subjective cognitive load predicted larger pupil dilation, with significant effects for NASA-TLX total ( $\beta$ -coefficient 0.009, 95% CI 0.009-0.010,  $P < 0.001$ ) and PAAS scores ( $\beta$ -coefficient 0.051, 95% CI 0.045-0.058,  $P < 0.001$ ). These effects were consistent across both groups, with no group difference in overall pupil change from Baseline to 4-Back ( $P = 0.196$ ). Higher cognitive load also predicted increased blink rate, with significant effects for NASA-TLX ( $\beta$ -coefficient 0.187, 95% CI 0.139-0.234,  $P < 0.001$ ) and PAAS ( $\beta$ -coefficient 1.081, 95% CI 0.827-1.335,  $P < 0.001$ ), again with no group-level differences.

Replacing subjective ratings with N-back level as the predictor produced the same overall pattern. Relative to Baseline, each task level was associated with greater pupil dilation (model-estimated mean difference in pupil diameter: 1-Back +0.252mm; 2-Back +0.435mm; 3-Back +0.478mm; 4-Back +0.455mm; all  $P < 0.001$ ), with most of the increase occurring by 3-Back. Using 1-Back as the reference condition yielded analogous results, with pupil dilation significantly greater in the 2-, 3- and 4-Back tasks (all  $P < 0.001$ ), supporting the use of these contrasts in subsequent classification analyses. Blink rate demonstrated a similar increase with N-back level, with significantly higher blink rate at each task level compared with Baseline (all  $P < 0.001$ ). For both eye tracking features, no group effect or group-by-task interaction was observed after adjustment.

Age was positively associated with both pupil size ( $P = 0.019$ ) and blink rate ( $P = 0.015$ ), although including age as a covariate did not alter the task or group effects. Pain and fatigue scores were not associated with either pupil size or blink rate. However, several behavioural and oculomotor indicators showed associations with both pupil size and blink rate, including performance indices ( $d'$  and  $c$ ) and self-reported difficulty focussing on the task (all  $P < 0.001$ ).

## Predictive Models

### Logistic Regression

A logistic regression model using summary eye-tracking features showed limited discriminative value for classifying low-load (1-Back) versus high-load (4-Back). Using mean pupil size alone, accuracy was 0.56 (95% CIs 0.45-0.66, AUC 0.57) in healthy controls and 0.51 (95% CIs 0.41-0.61, AUC 0.53) in the chronic pain group. Including blink rate provided only a small improvement, with accuracies reaching 0.60 (AUC 0.62-0.63). These values remain close to chance-level performance, indicating that simple aggregate pupil and blink metrics have limited discriminative value for this contrast.

## Time-series Classifiers

Evaluation of a diverse range of time-series classifiers (Table 2) showed consistently high accuracy for distinguishing 1-Back from 4-Back task levels in healthy controls using pupil-derived signals, with broadly comparable performance across classifiers. In contrast, classification performance in the chronic pain group was attenuated and showed greater inter-individual variability. Adding a blink-mask as a second time-series feature had negligible impact in healthy controls but produced small, reproducible gains in the chronic pain group, suggesting additional discriminative information carried by blink dynamics in this group.

Accuracy across classifiers ranged between 0.81-0.87 in healthy controls and 0.56-0.70 in the chronic pain group. A comparison of ROC curves between groups for classifiers that produce probabilistic estimates is shown in Figure 4. Given its computational efficiency, consistent behaviour across folds and performance, ROCKET was selected as the primary exemplar, with corresponding contingency tables

presented in Figure 5. In healthy controls, ROCKET's pooled confusion matrix showed a clear diagonal structure with minimal misclassification, and both sensitivity and specificity on held-out participants were substantially above chance.

Including blinks in the form of a blink mask as an additional time-series feature made little difference in healthy controls but yielded small, consistent gains in the chronic pain group. ROCKET showed a modest improvement in accuracy from 0.65 to 0.70 in the 1-Back versus 4-Back comparison with similar effects observed across other classifiers. These findings suggest that blink dynamics may provide complementary discriminative information in chronic pain, possibly reflecting altered oculomotor stability or compensatory blink patterns under cognitive load.

Given that earlier mixed-effects models identified age as a potential confounder, exploratory linear regression was conducted to evaluate whether age contributed to classification accuracy. In a model including both age and group, age showed no meaningful association with accuracy ( $\beta$ -coefficient  $-0.0006$ ,  $P=0.793$ ), accounting for less than 0.1% of explained variance. In contrast, group membership was associated with accuracy after adjusting for age ( $\beta$ -coefficient  $+0.1882$ ,  $P=0.018$ ), indicating higher classification accuracy in healthy controls than in participants with chronic pain. No age-group interaction was observed, suggesting that age does not differentially affect accuracy across groups

### **Adverse Effects and Task Acceptability**

VR tasks were well tolerated, with overall adverse effects remaining low across participants. Median Virtual Reality Sickness Questionnaire (VRSQ) scores were 8.3 [4.2–15.8] out of 100, indicating minimal cybersickness symptoms during the protocol. Although participants with chronic pain reported higher VRSQ scores than healthy controls (12.5 [6.7–17.3] versus 4.2 [4.2–12.5],  $p=0.006$ ,  $r=0.34$ ), absolute symptom levels were still low in both groups. Pain and fatigue VAS scores were higher in participants with chronic pain but remained stable across N-back levels, with a slight downward trend in median pain scores between Baseline (18.0 [9.3-25.3]) and 4-Back (10.5 [1.0-19.0]).

## **Discussion**

This study investigated whether consumer-grade, VR-integrated eye tracking can provide objective markers of cognitive load that could inform future closed-loop pain therapeutics, in healthy adults and a pragmatic, clinically heterogeneous cohort with chronic musculoskeletal pain. We hypothesised that ocular features would reliably index cognitive load during a VR N-back task in both groups, and that models exploiting temporal dynamics would outperform summary metrics. Increasing N-back level led to poorer performance and higher subjective workload; in parallel, pupil diameter and blink rate changed systematically with task level and workload under stable luminance. Simple models based on aggregate pupil and blink features showed limited discrimination of low versus high load, whereas time-series classifiers achieved higher participant-level performance, particularly in healthy controls. Although mixed-effects models indicated broadly similar mean task-evoked ocular responses across groups, classification performance was lower and more variable in chronic pain. This variability did not appear

attributable to age and underscores the additional challenges of modelling cognitive load in chronic pain groups, motivating cohort-sensitive and participant-calibrated approaches. Overall, these findings support VR-embedded eye tracking as a feasible biosignal for cognitive-load estimation, while highlighting the need for personalised models for translation to closed-loop applications in chronic pain.

The present results extend a large evidence-base demonstrating task-evoked pupillary responses (TEPRs) as sensitive indices of cognitive effort in working-memory paradigms, including the N-back<sup>37, 38</sup>. TEPRs are widely interpreted as reflecting engagement of arousal and cognitive control systems, including the locus coeruleus–noradrenergic system, with pupil diameter typically increasing with cognitive load and stabilising or attenuating as demands approach capacity<sup>25, 37, 38</sup>. In this N-back paradigm, mean pupil diameter increased from Baseline to 2-Back and then remained comparatively stable through 3-Back and 4-Back, broadly mirroring trajectories in performance and subjective workload, and consistent with a capacity-related plateau at higher levels.

The magnitude of pupillary change (approximately 0.2 to 0.6 mm) falls within ranges reported in laboratory pupillometry studies of mental workload<sup>39, 40</sup>. Metrological work also indicates that video-based eye trackers can resolve sub-millimetre changes under controlled conditions, supporting the physiological interpretability of fractional-millimetre effects when luminance and data quality are appropriately managed<sup>41</sup>. In VR, additional measurement considerations apply (for example, gaze-angle dependent pupil-size artefacts), reinforcing the value of centrally presented stimuli and stabilised luminance across blocks<sup>23, 42</sup>. Together with prior VR studies indicating that pupil dynamics can track cognitive load when luminance is controlled or modelled<sup>28, 43</sup>, these findings support the practicality of headset-integrated pupillometry for workload estimation in immersive applications.

Blink rate also increased with task level and subjective workload, but its interpretation is more context-dependent than pupil dilation. Blink suppression is frequently reported when continuous visual sampling demands are high<sup>44-46</sup>, whereas other studies have shown increased blink rate with tasks requiring sustained mental effort and internally focused processing<sup>47, 48</sup>. In this task, blink rate increased systematically from Baseline to 2-Back, with a relative plateau thereafter. Notably, blink rate was higher overall in the chronic pain cohort. While this difference was not explained by pain intensity or fatigue in the present analyses, it may reflect broader clinical factors (for example, medication exposure, comorbidity burden, autonomic regulation, or ocular surface symptoms)<sup>49</sup>. These results argue against blink counts as a standalone marker of cognitive load in VR but support their inclusion within multivariate models that account for task structure, viewing conditions, and individual differences.

The chronic pain cohort showed marginally poorer task performance, slower reaction times and greater subjective difficulty maintaining focus at higher N-back levels, despite broadly similar group-level ocular responses. This is consistent with evidence that chronic pain can be associated with attentional dysregulation and impairments in working memory and executive function, particularly in older adults and those with higher pain interference<sup>50, 51</sup>. Experimental studies also indicate that pain can disrupt N-

back performance and increase false alarms, consistent with competition between nociceptive and task-related processing for limited cognitive resources<sup>52</sup>.

Physiologically, the absence of marked group differences in mean TEPRs suggests that chronic pain did not simply blunt pupillary responsivity to cognitive load. Instead, the machine learning analyses highlight more nuanced alterations in how ocular dynamics relate to task demands at the individual level. Time-series classifiers discriminated low versus high load more reliably in healthy controls than in the chronic pain group, with increased off-diagonal misclassifications under leave-one-subject-out cross-validation. Despite age differences between groups, linear regression showed no significant association between age and classifier performance, proving reassurance that this pattern is not solely age-driven. A plausible interpretation is that chronic pain is associated with increased variability in the temporal coupling between cognitive demand and ocular responses, rather than a uniform shift in mean physiology. This supports the view that chronic pain does not represent a single cognitive phenotype, and motivates participant-calibrated and cohort-sensitive modelling approaches for future biomarker-informed or adaptive systems.

Immersive VR is increasingly used as a non-pharmacological adjunct for both acute and chronic pain, typically through reallocation of attentional resources away from pain (attentional modulation), movement-based activities, skills training or psychoeducational approaches. Systematic reviews and recent randomised controlled trials report improvements in pain-related outcomes across multiple clinical contexts, although effect sizes are variable and uncertainty remains around dosing, mechanisms and target populations<sup>12, 15, 53-55</sup>. Most deployed or evaluated VR interventions remain open-loop, with fixed content that does not adapt to patients' cognitive load, symptom burden or fatigue.

The present findings suggest a pathway toward adaptive VR therapeutics that explicitly monitor and regulate cognitive load. A closed-loop system could combine pupillometry blink dynamics and performance features to estimate whether a user is under-challenged, optimally engaged or overloaded, then adjust task difficulty, interaction demands, or sensory richness in real time. When load is below a personalised engagement target or "flow range", the system might increase goal-directed interaction to sustain attention. When overload is detected, it could simplify tasks or transition to calmer content to limit fatigue and disengagement. In chronic pain populations, where baseline cognitive strain and symptoms may fluctuate, the observed variability supports designing algorithms that can accommodate heterogeneity rather than assuming a fixed mapping between ocular dynamics and cognitive load.

In addition, cognitive-load estimation may benefit from integrating complementary signals that capture different aspects of task engagement. Prior work in non-clinical domains suggests that no single physiological measure generalises across all tasks and contexts, and that combining complementary features may improve robustness or interpretability of workload estimation in certain settings<sup>56, 57</sup>. Multimodal approaches, including complementary biosignals such as heart-rate variability or EEG where feasible, may help disentangle cognitive effort from arousal and affective responses, though their incremental benefit should be tested against added complexity. Critically, cognitive-load adaptation must

be aligned with therapeutic intent: some interventions may aim to sustain moderate-high demand and engagement (for example, attention-based and CBT modules), while others may target reduced cognitive load (for example, mindfulness-oriented VR). Closed-loop control should therefore be guided by theory-driven therapeutic targets, not engagement alone.

Several methodological elements strengthen the interpretation of this work. The study used a consumer-available VR headset with integrated binocular eye tracking, demonstrating that informative pupil and blink signals can be obtained within a realistic deployment rather than a under tightly constrained laboratory conditions. Cognitive load was manipulated with a well-characterised N-back task and quantified using complementary measures, including validated self-report instruments (NASA-TLX, PAAS) and signal-detection performance metrics. Eye-tracking was analysed as continuous time series across task blocks to preserve temporal ocular dynamics, luminance was logged and controlled at the block level, and preprocessing and feature extraction were embedded within cross-validation folds to reduce leakage. Performance was evaluated across multiple open-access, state-of-the-art time-series classifiers under leave-one-subject-out cross-validation, providing a conservative estimate of participant-level generalisation and supporting reproducibility.

Limitations remain. The single-session design precludes conclusions about within-person stability across days and fluctuating symptom states. The chronic pain cohort was clinically representative but heterogeneous, including multiple diagnoses, comorbidities, and medication exposures that may influence cognition and ocular physiology, so larger studies with stratified sampling and richer phenotyping are needed. Pain severity and impact were generally low-to-moderate, limiting generalisability to more severe or disabling pain. The N-back task provides a controlled probe of cognitive load but does not capture the full complexity of therapeutic VR experiences. While luminance was controlled here, therapeutic VR will typically involve greater visual and luminance variability that will need to be explicitly modelled to support future real-world closed-loop use. Finally, all machine-learning models were trained and evaluated using internal cross-validation, and no independent external dataset was available for validation. As such, the generalisability of the classification results beyond the present sample remains to be established. Moreover, all analyses were conducted offline on desktop hardware; translation to real-time implementation on standalone VR devices will require additional engineering, model optimisation and prospective validation.

Future research should prioritise replication in larger and more diverse chronic pain cohorts, longitudinal studies across varying pain and fatigue states, and study of alternative VR hardware and eye-tracking pipelines. Prospective trials should test whether cognitive-load-adaptive control improves engagement, tolerability, and clinical outcomes compared with open-loop VR within ecologically valid therapeutic applications. Translational progress will require close collaboration between clinicians, VR developers, and machine-learning researchers to ensure robustness to device heterogeneity and missing or noisy data, alongside acceptability, interpretability, and strong data governance.

This study provides initial evidence that consumer-grade VR-integrated eye tracking captures task-evoked changes consistent with cognitive load during a VR N-back task in adults with and without chronic musculoskeletal pain. While summary pupil and blink features offered limited discrimination of low versus high load, time-series machine learning classifiers that utilise temporal dynamics achieved higher discriminative performance, with performance attenuated and more variable in the chronic pain group. These findings support VR-embedded eye tracking as a practical route to estimating cognitive load in immersive settings, and motivate participant-calibrated, externally validated approaches as next steps toward safe, clinically meaningful closed-loop VR interventions.

## Methods

### Participants and Ethical Approval

This repeated-measures experimental study included adults with chronic musculoskeletal pain and healthy controls. Participants were recruited between March 2025 and June 2025. Ethical approval was gained through the University of East Anglia's (UEA) Faculty for Medicine and Health Research Ethics Sub-committee (REF: ETH2324-2498). All procedures followed relevant guidelines and regulations, including the Declaration of Helsinki. Written informed consent was obtained from all participants prior to enrolment.

Healthy controls were current UEA students or staff members affiliated to the School of Medicine or School of Health Sciences, recruited through institutional mailing lists, posters, and social media. Specific healthy control exclusion criteria included current acute or chronic pain, and current medical conditions or use of medications known to affect eye movements or pupillary responses.

Participants with chronic pain were identified through two existing ethically approved research databases with prior consent for re-contact related to the Norfolk Arthritis Register<sup>58</sup> and the VIPA study<sup>15</sup>. Specific inclusion criteria for participants with chronic pain included current pain lasting  $\geq 3$  months at the time of recruitment<sup>6</sup>.

General inclusion criteria for both groups were: age  $\geq 18$  years, conversational English proficiency, and capacity to provide informed consent. General exclusion criteria included any condition exacerbated by flashing lights or screens, significant visual or hearing impairment that would preclude use of the VR system, diagnosis of cognitive impairment, facial injury or other condition preventing comfortable VR system use.

All screening and consent procedures were conducted via the VR-FOCUS project website, which hosted the participant information sheet and online eligibility questionnaire. Eligible individuals completed electronic consent and were contacted by the research team to schedule a single on-site testing session at the UEA.

# Apparatus and Virtual Environment

The consumer available Pico Neo 3 Pro Eye<sup>®</sup> VR system (Pico Interactive, San Francisco, USA)<sup>59</sup> containing an integrated Tobii<sup>®</sup> binocular eye-tracking system (sampling rate 60-90 Hz) was used to deliver the VR N-Back tasks. This included dual 1832 x 1920 pixel LCD display panels, 90Hz refresh rate and 98° horizontal field of view. The headset was used offline, with data captured during the task written to the device's persistent data path.

The VR N-Back application was developed in Unity<sup>®</sup> (v2021.3.45f1, Unity Technologies, San Francisco, USA)<sup>60</sup> with use of the Tobii Ocumen SDK<sup>61</sup> to enable capture of advanced eye tracking metrics. Headset output was mirrored via USB-C to a research laptop to allow real-time observation. Participants used one of the handheld VR controllers to interact in VR. Testing rooms included minimal external noise pollution and participants remaining seated on a comfortable chair during VR use (Figure 6, Panel A).

An eye-tracking calibration was performed before commencing the first task. This included standardisation of the headset position and a 5-point gaze-based calibration, using the Tobii Ocumen configuration tool included with the Ocumen SDK. Calibration was considered successful when all points showed bias  $\leq 3^\circ$  and precision  $\leq 1^\circ$  for both eyes, with all samples valid and used, triggering a green indicator next to the VR task menu. When calibration was unsuccessful, participants re-completed this step with the goal of gaining a successful calibration. Headset fit and interpupillary distance were individually adjusted.

## VR N-Back Task

The N-Back task is a validated neurocognitive task used to deliver different 'levels' of cognitive load. Participants are presented with a continuous sequence of letters and respond when the current letter matches one shown N steps earlier. Task difficulty increases with higher N levels, requiring greater working memory and cognitive load.

A bespoke VR N-Back task was developed to include five levels: Baseline, 1-Back, 2-Back, 3-Back, and 4-Back (Figure 6, Panel C). Letters within each task were randomly generated prior to each completion of the task, with 'matches' inserted at random locations. Each level presented 40 letters, including 10 matches (none in Baseline). For higher-load conditions (2-, 3-, 4-Back), 'lures' were inserted at random locations in the sequence (aiming for 15-20% of stimuli). 'Lures' were defined as a letter that appears one step before a match would typically present. Each letter appeared for 750ms, with a 2000ms inter-trial interval. Task parameters were chosen as a compromise to limit total VR exposure to <20 minutes (given risk of fatigue and other VR-related side effects), deliver appropriate levels of cognitive load to a chronic pain cohort, remain within widely used N-back timing ranges<sup>62, 63</sup>, and provide sufficient data for statistical and machine-learning analyses.

The virtual environment consisted of a neutral grey surrounding space with a darker virtual floor. Task information and letter stimuli appeared as dull white text on a darker grey rectangular panel centred in

the user's field of view (Figure 6, Panel B). There was no requirement for participants to rotate or move forwards/backwards during the task to enable minimisation of gaze-angle variability and prevent participant neck strain, particularly in the chronic pain population.

To account for the influence of scene luminance on pupil diameter, mean on-screen luminance was calculated for each task block. Luminance was derived from frame-level RGB values exported from the Unity rendering pipeline and converted to a standardised luminance estimate using a linear RGB-to-luminance transform ( $Y = 0.2126R + 0.7152G + 0.0722B$ )<sup>64, 65</sup>. For each task level, luminance values were averaged across frames to produce a single mean luminance estimate, alongside its standard deviation and coefficient of variation (CV). Luminance remained stable across all N-back levels (0.376-0.378; CV <1%), ensuring that observed pupil-dilation effects<sup>61</sup> reflected cognitive processing rather than scene luminance changes.

## Experimental Procedure

Prior to VR use, participants completed the baseline questionnaires and were briefed on the equipment and tasks. This included being shown short tutorial videos explaining each task. Understanding was confirmed verbally before using VR.

After putting the headset on, participants first completed the standardised eye tracking calibration. They were given the option to refresh their memory of a task by watching the short video tutorial again in VR before starting. Each task began with a 10-second fixation cross to collect that task's baseline eye tracking metrics and give a period of time for the pupil to stabilise. Participants pressed the controller trigger when they detected a match. If a letter was present on screen at the time, that letter underlined to provide input feedback (Figure 6, Panel B). No accuracy feedback was given.

Tasks were performed in a fixed order for the first two conditions (Baseline, 1-Back), followed by a pre-randomised allocation (counterbalanced, six pre-defined task orders) for the order in which they completed the 2-, 3-, and 4-Back tasks. This was to account for the risk of order effect on task difficulty, pupil metrics and cognitive load.

After each task, participants completed brief VR questionnaires using the handheld controller. Following all tasks, participants removed the VR system and completed and completed the Virtual Reality Sickness Questionnaire (VRSQ) on paper prior to debrief.

## Measures

A comprehensive set of subjective, behavioural and physiological measures was collected during a single in-person session to characterise baseline health status, task performance, cognitive load and ocular responses. All questionnaires were administered in English. Digital forms were used for all measures except the Pain Catastrophising Scale, which was completed on paper.

## Baseline Pre-VR Measures

Participants first completed a baseline characteristics questionnaire that captured demographic information, previous exposure to technology, medical comorbidities and regular medications. Individuals with chronic musculoskeletal pain additionally completed validated patient-reported outcome measures to quantify symptom severity and functional impact. These comprised the PROMIS Pain Intensity (Short Form 3a)<sup>66</sup> and PROMIS Pain Interference (Short Form 8a)<sup>66</sup> instruments, providing structured assessments of pain severity and its disruption to day-to-day activity. They also completed the Pain Catastrophising Scale<sup>67</sup> and provided single-item numeric ratings of current pain and fatigue intensity on a 0–10 scale. Together, these assessments enabled detailed characterisation of the chronic pain cohort and provided data on potential covariates for subsequent analyses.

## Performance Measures During VR

Task performance was recorded automatically for each N-back level except Baseline. For each stimulus sequence, the system logged the number of correctly identified targets, those responded to that were incorrect, false alarms to lure stimuli, and reaction times. These raw metrics were used to derive discrimination index ( $d'$ ; a participant's ability to distinguish targets from non-targets) and response bias ( $c$ ; a participant's tendency to respond more liberally or conservatively) according to standard signal-detection conventions<sup>68, 69</sup>. These measures allowed quantification of working-memory performance and supported analyses of how behaviour related to subjective load and physiological responses.

## Eye-tracking Measures

Continuous eye-tracking data were captured during every task block, with timestamped pupil diameter and gaze direction in three-dimensional space recorded at 60–90 Hz for both eyes and written directly to device storage. Missing samples arising from blinks were processed according to predefined temporal criteria as described in the Data Pre-processing section. These ocular metrics comprised the primary physiological measures used to investigate cognitive load, evaluate relationships with subjective and behavioural metrics, and train time-series classifiers.

## Post-task Subjective Measures Collected in VR

After completing each N-back block, participants remained in the VR environment and completed a brief set of subjective ratings using a continuous sliding scale. Cognitive workload was assessed using the NASA Task Load Index<sup>70</sup>, which includes six dimensions rated from 0 to 100, and an adapted version of the one-item PAAS scale<sup>71</sup>, recorded on a 0–9 scale in VR and subsequently rescaled to align with conventional scoring (1-9). Participants also rated current pain and fatigue intensity using blinded 0–100 visual analogue scales. A seven-item subjective experience questionnaire assessed immersion, comfort,

anxiety, ability to maintain attention, difficulty focusing, motivation and engagement. These additional measures provided contextual information about users' experiences and potential sources of variance in eye-tracking or performance data.

## Post-VR Measures

Immediately after removal of the headset, participants completed the Virtual Reality Sickness Questionnaire<sup>44</sup>. This nine-item measure quantified the presence and severity of VR side effects, including disorientation, nausea and oculomotor discomfort. These data allowed assessment of tolerability of the VR protocol and examination of group-level differences in adverse effects.

## Data Pre-processing & Analysis

### Sample size

Given the exploratory nature of this study and the use of machine-learning analyses, a formal power calculation was not performed. A target sample of 30–50 participants per group was selected to provide sufficient variability for modelling while remaining feasible for an intensive VR-based protocol.

### Data Pre-processing

Raw eye-tracking streams were exported as .ocumen files from Unity and converted into participant-level datasets using Python 3.12 within an Anaconda 3<sup>72</sup> environment. Data were acquired at 60 or 90 Hz, downsampled to 60 Hz using nearest-neighbour interpolation when required and annotated by task epoch with a peri-stimulus window (–200 to +1300 ms) to account for pupil-dilation latency<sup>73</sup>. Blink events were defined as gaps of 50–500ms<sup>74</sup> and reconstructed using a cubic-spline interpolation with linear fallback<sup>75</sup>; longer artefacts were removed. Unweighted NASA-TLX scores were used. PAAS scores (recorded on 0–9 due to a Unity constraint) were converted so that 0 corresponded to the traditional lower bound of 1. Performance indices, including discrimination index ( $d'$ ; a participant's ability to distinguish targets from non-targets) and response bias ( $c$ ; a participant's tendency to respond more liberally or conservatively), were computed according to signal-detection-theory conventions<sup>68, 69</sup>.

### Statistical Analysis

Conventional analyses were conducted in R (v4.4.3)<sup>76</sup>. Continuous variables were assessed for normality using the Shapiro–Wilk test and summarised as mean  $\pm$  SD for normal or median [IQR] for non-normal data. Categorical variables were reported as frequencies and percentages. Associations between pupil diameter, blink rate, and cognitive-load scores (NASA-TLX, PAAS) were evaluated using linear mixed-effects models with participant ID as a random effect. Models adjusted for group, age, task order, sampling resolution, and fatigue, with exploratory analyses including anxiety, engagement, perceived difficulty, attention consistency, and performance indices ( $d'$ ,  $c$ ). Effect sizes were reported using  $\beta$ -coefficients with 95% confidence intervals. Model diagnostics were inspected visually. All tests of statistical significance were two-tailed with  $\alpha=0.05$  and p-values reported where appropriate.

## Predictive Models

As an interpretable baseline, a logistic regression classifier was implemented using pre-specified task-level summary eye-tracking features (mean left pupil size and blink rate). No automated feature selection, interaction terms or non-linear transformations were applied. Summary features were winsorised and z-scored globally prior to cross-validation. Models were evaluated using leave-one-subject-out cross-validation, with performance quantified using pooled out-of-fold accuracy with Wilson 95% confidence intervals and receiver operating characteristic area under the curve (ROC AUC).

Time-series classification was performed in Python (v3.12)<sup>77</sup> using the time series machine learning toolkit aeon (v1.1)<sup>35</sup>. Continuous eye-tracking streams were segmented into task epochs (Baseline to 4-Back) and harmonised to a common length to ensure comparable temporal structure across participants. The primary multivariate feature set comprised left-eye pupil diameter and a binary blink-indicator feature. Missing samples were imputed using the per-case, per-feature mean. Outliers were winsorised and data were normalised using per-feature z-scores, with all preprocessing parameters estimated on the training data only and applied unchanged to the held-out participant to avoid information leakage. The same preprocessing pipeline was applied uniformly to both healthy controls and participants with chronic pain to ensure methodological symmetry across groups.

For classification tasks, N-back levels (e.g. 1-Back vs 4-Back) was used as class labels. A range of time-series specific classification algorithms were used in the evaluation. Classifiers representing the state of the art for different feature representations of the data were selected based on a recent comparative study<sup>78</sup>: Rotation Forest<sup>79</sup> is a standard classifier benchmark that trains an ensemble of decision trees on rotated feature spaces; ROCKET<sup>32</sup> is a pipeline classifier that uses a large sets of random convolutional kernels to generate features for a linear classifier. QUANT<sup>80</sup> and DrCIF<sup>34</sup> are classifiers that use summary statistics taken over subseries and are designed to find localised temporal discriminatory features, with QUANT providing markedly faster training and prediction because its quantile-based feature extraction is computationally minimal. HIVE-COTE 2.0<sup>34</sup> is a state-of-the-art hierarchical meta ensemble that combines four diverse time-series models. For computational feasibility, the maximum training time for each HIVE-COTE 2.0 fold was capped at eight minutes. Default aeon hyperparameters were used, and stochastic components were initialised with fixed random seeds to ensure reproducibility. All models were trained using leave-one-subject-out cross-validation so that all epochs from a participant were held out together.

Performance was evaluated using pooled out-of-fold predictions, with mean accuracy and confidence intervals as the primary metrics. Standard deviations across cross-validation folds were not reported because fold-level accuracies were highly discretised due to the small number of trials in the held-out test set per fold, making variance estimates mathematically inflated and not informative of model stability. Receiver Operating Characteristic (ROC) AUC values were computed only for classifiers that generate probabilistic predictions. ROCKET does not output class probabilities and therefore cannot be evaluated using ROC curves or AUC; its performance was summarised using accuracy, sensitivity,

specificity and confusion matrices. Model behaviour was examined using these metrics alongside ROC curves for the probabilistic models.

To assess whether age contributed to between-group differences in classifier performance, exploratory linear regression was used to model participant-level classification accuracy as a function of age and group membership, with an age-by-group interaction tested. Regression coefficients ( $\beta$ ) with corresponding two-tailed p-values were reported.

## Declarations

**Acknowledgements:** None

**Author contributions:** J.T. conceived and designed the study, conducted recruitment and data collection, performed all analyses, interpreted the results, and drafted the manuscript. J.B. provided methodological input and critically reviewed the manuscript. J.D. provided expertise and support with conventional data analyses. A.B. provided expertise and support for predictive modelling approaches. A.M. provided methodological guidance and critically reviewed the manuscript. All authors reviewed and approved the final manuscript.

**Data availability statement:** The data that support the findings of this study are available from the corresponding author [JT] upon reasonable request

## Additional Information

**Competing interests:** J.T. holds an honorary Chief Medical Officer role at Revolve Labs Ltd. Revolve Labs Ltd provided financial support for J.T.'s PhD programme during which this study was conducted. The company had no role in the study design, data collection, analysis, interpretation, manuscript preparation, or decision to submit for publication. The other authors declare no competing interests.

**Funding:** This work was completed as part of J.T.'s PhD programme and was supported by funding from the University of East Anglia and Revolve Labs Ltd.

## References

1. Gill TK, Mittinty MM, March LM, et al. Global, regional, and national burden of other musculoskeletal disorders, 1990–2020, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. *The Lancet Rheumatology* 2023; 5: e670-e682. DOI: 10.1016/S2665-9913(23)00232-1.
2. Lurie JM and Javaid A. Visualizing Global Chronic Pain. *Anesthesia & Analgesia* 2024; 138: 918-919. DOI: 10.1213/ane.0000000000006564.
3. Gatchel RJ, Peng YB, Peters ML, et al. The biopsychosocial approach to chronic pain: scientific advances and future directions. *Psychol Bull* 2007; 133: 581-624. DOI: 10.1037/0033-

2909.133.4.581.

4. Hutchings A, Calloway M, Choy E, et al. The Longitudinal Examination of Arthritis Pain (LEAP) study: relationships between weekly fluctuations in patient-rated joint pain and other health outcomes. *J Rheumatol* 2007; 34: 2291-2300. 20071015.
5. Rogers AH, Smit T, Bakhshaie J, et al. Impact of intraindividual pain variability on functional pain outcomes among adults with chronic pain: an ecological momentary assessment study. *Journal of Behavioral Medicine* 2025; 48: 1035-1042. DOI: 10.1007/s10865-025-00590-x.
6. Excellence NIHaC. *Chronic pain (primary and secondary) in over 16s: assessment of all chronic pain and management of chronic primary pain*. April 2023 2023. National Institute for Health and Care Excellence (NICE).
7. Häuser W, Walitt B, Fitzcharles MA, et al. Review of pharmacological therapies in fibromyalgia syndrome. *Arthritis Res Ther* 2014; 16: 201. 20140117. DOI: 10.1186/ar4441.
8. Macfarlane GJ, Kronisch C, Dean LE, et al. EULAR revised recommendations for the management of fibromyalgia. *Ann Rheum Dis* 2017; 76: 318-328. 20160704. DOI: 10.1136/annrheumdis-2016-209724.
9. Rivera J, Molina-Collada J, Martínez-Barrio J, et al. Opioids and fibromyalgia: frequency of use and factors associated with increased consumption in patients remitted to a tertiary care center. *BMC Musculoskeletal Disorders* 2024; 25: 121. DOI: 10.1186/s12891-024-07263-x.
10. Pretat T, Koller C and Hügler T. Virtual reality as a treatment for chronic musculoskeletal pain syndromes. *Joint Bone Spine* 2025; 92: 105769. DOI: <https://doi.org/10.1016/j.jbspin.2024.105769>.
11. Tsigarides J, Shenker NG and MacGregor A. Ready Patient One: the role of therapeutic virtual reality in the future management of chronic pain. *Rheumatology (Oxford)* 2022; 61: 482-483. DOI: 10.1093/rheumatology/keab594.
12. Eccleston C, Fisher E, Liikkanen S, et al. A prospective, double-blind, pilot, randomized, controlled trial of an "embodied" virtual reality intervention for adults with low back pain. *Pain* 2022; 163: 1700-1715. 20220325. DOI: 10.1097/j.pain.0000000000002617.
13. Garcia LM, Birckhead BJ, Krishnamurthy P, et al. Three-Month Follow-Up Results of a Double-Blind, Randomized Placebo-Controlled Trial of 8-Week Self-Administered At-Home Behavioral Skills-Based Virtual Reality (VR) for Chronic Low Back Pain. *J Pain* 2022; 23: 822-840. 20211211. DOI: 10.1016/j.jpain.2021.12.002.
14. Garcia LM, Birckhead BJ, Krishnamurthy P, et al. An 8-Week Self-Administered At-Home Behavioral Skills-Based Virtual Reality Program for Chronic Low Back Pain: Double-Blind, Randomized, Placebo-Controlled Trial Conducted During COVID-19. *J Med Internet Res* 2021; 23: e26292. 20210222. DOI: 10.2196/26292.
15. Tsigarides J, Grove V, Chipping J, et al. Feasibility of attention-based virtual reality interventions in fibromyalgia syndrome: comparing systems, virtual environments and activities. *Br J Pain* 2025: 20494637241310696. 20250104. DOI: 10.1177/20494637241310696.

16. Themelis K and Tang NKY. The Management of Chronic Pain: Re-Centring Person-Centred Care. *J Clin Med* 2023; 12 20231107. DOI: 10.3390/jcm12226957.
17. Baker C and Fairclough SH. Chapter 9 - Adaptive virtual reality. In: Fairclough SH and Zander TO (eds) *Current Research in Neuroadaptive Technology*. Academic Press, 2022, pp.159-176.
18. Serrano RR-M, Troughton JG, Mirkhani N, et al. From Neural Sensing to Stimulation: An Interdisciplinary Roadmap for Neurotechnology. *arXiv preprint arXiv:251007116* 2025.
19. Csikszentmihalyi M. *Flow: The Psychology of Optimal Experience*. New York: Harper-Collins, 1990.
20. Bantick SJ, Wise RG, Ploghaus A, et al. Imaging how attention modulates pain in humans using functional MRI. *Brain* 2002; 125: 310-319. DOI: 10.1093/brain/awf022.
21. Beals M. *Flow as a Tool to Manage Chronic Pain*. University of California, 2023.
22. Hoffman HG, Richards TL, Bills AR, et al. Using fMRI to study the neural correlates of virtual reality analgesia. *CNS Spectr* 2006; 11: 45-51. DOI: 10.1017/s1092852900024202.
23. Adhanom IB, MacNeilage P and Folmer E. Eye Tracking in Virtual Reality: a Broad Review of Applications and Challenges. *Virtual Real* 2023; 27: 1481-1505. 20230118. DOI: 10.1007/s10055-022-00738-z.
24. Schuetz I and Fiehler K. Eye Tracking in Virtual Reality: Vive Pro Eye Spatial Accuracy, Precision, and Calibration Reliability. *J Eye Mov Res* 2022; 15 20220907. DOI: 10.16910/jemr.15.3.3.
25. Beatty J. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin* 1982; 91: 276-292. DOI: 10.1037/0033-2909.91.2.276.
26. Van Orden KF, Limbert W, Makeig S, et al. Eye Activity Correlates of Workload during a Visuospatial Memory Task. *Human Factors* 2001; 43: 111-121. DOI: 10.1518/001872001775992570.
27. Yeung MK, Lee TL, Han YMY, et al. Prefrontal activation and pupil dilation during n-back task performance: A combined fNIRS and pupillometry study. *Neuropsychologia* 2021; 159: 107954. 20210709. DOI: 10.1016/j.neuropsychologia.2021.107954.
28. Lee JY, Jong Nd, Donkers J, et al. Measuring Cognitive Load in Virtual Reality Training via Pupillometry. *IEEE Transactions on Learning Technologies* 2024; 17: 704-710. DOI: 10.1109/TLT.2023.3326473.
29. Wei J, Siegel E, Sundaramoorthy P, et al. Cognitive Load Inference Using Physiological Markers in Virtual Reality. In: *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)* 8-12 March 2025 2025, pp.759-769.
30. Chan FHF, Suen H, Jackson T, et al. Pain-related attentional processes: A systematic review of eye-tracking research. *Clin Psychol Rev* 2020; 80: 101884. 20200616. DOI: 10.1016/j.cpr.2020.101884.
31. Rajula HSR, Verlato G, Manchia M, et al. Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment. *Medicina (Kaunas)* 2020; 56 20200908. DOI: 10.3390/medicina56090455.
32. Dempster A, Petitjean F and Webb GI. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery* 2020; 34:

- 1454-1495. DOI: 10.1007/s10618-020-00701-z.
33. Middlehurst M, Large J and Bagnall A. The Canonical Interval Forest (CIF) Classifier for Time Series Classification. In: *2020 IEEE International Conference on Big Data (Big Data)* 10-13 Dec. 2020 2020, pp.188-195.
34. Middlehurst M, Large J, Flynn M, et al. HIVE-COTE 2.0: a new meta ensemble for time series classification. *Machine Learning* 2021; 110: 3211-3243. DOI: 10.1007/s10994-021-06057-9.
35. Middlehurst M, Ismail-Fawaz A, Guillaume A, et al. aeon: a Python Toolkit for Learning from Time Series. *Journal of Machine Learning Research* 2024; 25: 1-10.
36. Berryman C, Stanton TR, Jane Bowering K, et al. Evidence for working memory deficits in chronic pain: A systematic review and meta-analysis. *PAIN@* 2013; 154: 1181-1196. DOI: <https://doi.org/10.1016/j.pain.2013.03.002>.
37. Johnson EL, Miller Singley AT, Peckham AD, et al. Task-evoked pupillometry provides a window into the development of short-term memory capacity. *Front Psychol* 2014; 5: 218. 20140313. DOI: 10.3389/fpsyg.2014.00218.
38. van der Wel P and van Steenbergen H. Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin & Review* 2018; 25: 2005-2015. DOI: 10.3758/s13423-018-1432-y.
39. Bauer R, Jost L, Günther B, et al. Pupillometry as a measure of cognitive load in mental rotation tasks with abstract and embodied figures. *Psychol Res* 2022; 86: 1382-1396. 20210812. DOI: 10.1007/s00426-021-01568-5.
40. Beatty J and Lucero-Wagoner B. The pupillary system. *Handbook of psychophysiology. Cambridge University Press*, xiii 2000; 1039: 142-162.
41. Klingner J. The pupillometric precision of a remote video eye tracker. *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*. Austin, Texas: Association for Computing Machinery, 2010, p. 259–262.
42. Salari M, Niehorster DC, Nyström M, et al. The effect of pupil size on data quality in head-mounted eye trackers. *Behav Res Methods* 2025; 58: 17. 20251203. DOI: 10.3758/s13428-025-02880-3.
43. Thomay C, Fermitsch A, Fessler J, et al. Towards Cognitive Load-Based Decision Making in VR Training. In: *2023 IEEE 2nd International Conference on Cognitive Aspects of Virtual Reality (CVR)* 26-27 Oct. 2023 2023, pp.000023-000028.
44. Kim HK, Park J, Choi Y, et al. Virtual reality sickness questionnaire (VRSQ): Motion sickness measurement index in a virtual reality environment. *Applied Ergonomics* 2018; 69: 66-73. DOI: <https://doi.org/10.1016/j.apergo.2017.12.016>.
45. Kim J, Hwang L, Kwon S, et al. Change in Blink Rate in the Metaverse VR HMD and AR Glasses Environment. *International Journal of Environmental Research and Public Health* 2022; 19: 8551.
46. Stern JA, Walrath LC and Goldstein R. The endogenous eyeblink. *Psychophysiology* 1984; 21: 22-33. DOI: 10.1111/j.1469-8986.1984.tb02312.x.

47. Biondi FN, Saberi B, Graf F, et al. Distracted worker: Using pupil size and blink rate to detect cognitive load during manufacturing tasks. *Appl Ergon* 2023; 106: 103867. 20220812. DOI: 10.1016/j.apergo.2022.103867.
48. Magliacano A, Fiorenza S, Estraneo A, et al. Eye blink rate increases as a function of cognitive load during an auditory oddball paradigm. *Neurosci Lett* 2020; 736: 135293. 20200806. DOI: 10.1016/j.neulet.2020.135293.
49. Okazaki Y and Yokoi N. Time-course observation of tear film dynamics during VR headset use. *Scientific Reports* 2025; 15: 33082. DOI: 10.1038/s41598-025-16634-w.
50. Chen J, Wang X and Xu Z. The Relationship Between Chronic Pain and Cognitive Impairment in the Elderly: A Review of Current Evidence. *J Pain Res* 2023; 16: 2309-2319. 20230707. DOI: 10.2147/jpr.S416253.
51. Zhang X, Gao R, Zhang C, et al. Evidence for Cognitive Decline in Chronic Pain: A Systematic Review and Meta-Analysis. *Front Neurosci* 2021; 15: 737874. 20210922. DOI: 10.3389/fnins.2021.737874.
52. Attridge N, Noonan D, Eccleston C, et al. The disruptive effects of pain on n-back task performance in a large general population sample. *Pain* 2015; 156: 1885-1891. DOI: 10.1097/j.pain.0000000000000245.
53. Dreesmann NJ, Su H and Thompson HJ. A Systematic Review of Virtual Reality Therapeutics for Acute Pain Management. *Pain Manag Nurs* 2022; 23: 672-681. 20220719. DOI: 10.1016/j.pmn.2022.05.004.
54. Lier EJ, de Vries M, Steggink EM, et al. Effect modifiers of virtual reality in pain management: a systematic review and meta-regression analysis. *Pain* 2023; 164: 1658-1665. 20230322. DOI: 10.1097/j.pain.0000000000002883.
55. Maddox T, Oldstone L, Sackman J, et al. Twelve-month results for a randomized sham-controlled effectiveness trial of an in-home skills-based virtual reality program for chronic low back pain. *Pain Rep* 2024; 9: e1182. 20240904. DOI: 10.1097/pr9.0000000000001182.
56. Charles RL and Nixon J. Measuring mental workload using physiological measures: A systematic review. *Applied Ergonomics* 2019; 74: 221-232. DOI: <https://doi.org/10.1016/j.apergo.2018.08.028>.
57. Hogervorst MA, Brouwer AM and van Erp JB. Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. *Front Neurosci* 2014; 8: 322. 20141014. DOI: 10.3389/fnins.2014.00322.
58. Gwinnutt JM, Symmons DPM, MacGregor AJ, et al. Have the 10-year outcomes of patients with early inflammatory arthritis improved in the new millennium compared with the decade before? Results from the Norfolk Arthritis Register. *Annals of the Rheumatic Diseases* 2018; 77: 848-854. DOI: 10.1136/annrheumdis-2017-212426.
59. Pico I. Pico Neo 3 Pro Eye virtual reality headset. San Francisco, USA: Pico Interactive, 2021.
60. Unity T. Unity Real-Time Development Platform, version 2021.3.45f1. San Francisco, USA: Unity Technologies, 2021.

61. Tobii AB. Tobii Ocumen SDK: Overview, <https://developer.tobii.com/xr/ocumen/overview/> (2025, accessed 27/11/2025).
62. Byrne EM, Gilbert RA, Kievit RA, et al. Evidence for separate backward recall and n-back working memory factors: a large-scale latent variable analysis. *Memory* 2024; 32: 1182-1198. DOI: 10.1080/09658211.2024.2393388.
63. Jaeggi SM, Buschkuhl M, Perrig WJ, et al. The concurrent validity of the N-back task as a working memory measure. *Memory* 2010; 18: 394-412. 20100419. DOI: 10.1080/09658211003702171.
64. International Telecommunication U. *Parameter values for the HDTV standards for production and international programme exchange (ITU-R BT.709-6)*. 2015. Geneva, Switzerland: International Telecommunication Union.
65. Mathôt S. Pupillometry: Psychology, Physiology, and Function. *Journal of Cognition* 2018. DOI: 10.5334/joc.18.
66. Cella D, Riley W, Stone A, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005&#x2013;2008. *Journal of Clinical Epidemiology* 2010; 63: 1179-1194. DOI: 10.1016/j.jclinepi.2010.04.011.
67. Sullivan MJL, Bishop SR and Pivik J. The Pain Catastrophizing Scale: Development and validation. *Psychological Assessment* 1995; 7: 524-532. DOI: 10.1037/1040-3590.7.4.524.
68. Haatveit BC, Sundet K, Hugdahl K, et al. The validity of d prime as a working memory index: Results from the "Bergen n-back task". *Journal of Clinical and Experimental Neuropsychology* 2010; 32: 871-880. DOI: 10.1080/13803391003596421.
69. Kane MJ, Conway ARA, Miura TK, et al. Working memory, attention control, and the N-back task: a question of construct validity. *J Exp Psychol Learn Mem Cogn* 2007; 33: 615-622. DOI: 10.1037/0278-7393.33.3.615.
70. Hart SG and Staveland LE. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In: Hancock PA and Meshkati N (eds) *Advances in Psychology*. North-Holland, 1988, pp.139-183.
71. Paas F. Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology* 1992; 84: 429-434.
72. Anaconda I. Anaconda Distribution, version 3. Austin, TX, USA: Anaconda Inc., 2023.
73. Hoeks B and Levelt WJM. Pupillary dilation as a measure of attention: a quantitative system analysis. *Behavior Research Methods, Instruments, & Computers* 1993; 25: 16-26. DOI: 10.3758/BF03204445.
74. Nyström M, Andersson R, Niehorster DC, et al. What is a blink? Classifying and characterizing blinks in eye openness signals. *Behavior Research Methods* 2024; 56: 3280-3299. DOI: 10.3758/s13428-023-02333-9.
75. Mathôt S, Fabius J, Van Heusden E, et al. Safe and sensible preprocessing and baseline correction of pupil-size data. *Behav Res Methods* 2018; 50: 94-106. DOI: 10.3758/s13428-017-1007-2.

76. Team RC. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2025.

77. Python Software F. Python Language Reference, version 3.12. Wilmington, DE, USA: Python Software Foundation, 2023.

78. Middlehurst M, Schäfer P and Bagnall A. Bake off redux: a review and experimental evaluation of recent time series classification algorithms. *Data Mining and Knowledge Discovery* 2024; 38: 1958-2031. DOI: 10.1007/s10618-024-01022-1.

79. Rodriguez JJ, Kuncheva LI and Alonso CJ. Rotation Forest: A New Classifier Ensemble Method. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2006; 28: 1619-1630. DOI: 10.1109/TPAMI.2006.211.

80. Dempster A, Schmidt DF and Webb GI. quant: a minimalist interval method for time series classification. *Data Mining and Knowledge Discovery* 2024; 38: 2377-2402. DOI: 10.1007/s10618-024-01036-9.

## Tables

**Table 1.** Baseline participant characteristics by group

| Characteristic                              | Healthy controls (n=42) | Chronic pain (n=42) |
|---------------------------------------------|-------------------------|---------------------|
| Age in years, mean (SD)                     | 35.3 (12.2)             | 57.8 (11.4)         |
| Gender, n (%)                               |                         |                     |
| Male                                        | 13 (31%)                | 8 (19%)             |
| Female                                      | 29 (69%)                | 34 (81%)            |
| Ethnicity, n (%)                            |                         |                     |
| White                                       | 31 (74%)                | 41 (98%)            |
| Non-white                                   | 11 (26%)                | 1 (2%)              |
| Education level, n (%)                      |                         |                     |
| Up to A level                               | 9 (21%)                 | 33 (79%)            |
| University                                  | 26 (62%)                | 16 (38%)            |
| Previous VR Exposure, n (%)                 | 34 (81%)                | 22 (52%)            |
| Number of Medical Conditions, median (IQR)  | 0 (0-1)                 | 5 (3-6)             |
| Number of Regular Medications, median (IQR) | 0 (0-1)                 | 5 (3-7)             |

Data are n (%) unless otherwise stated. IQR, interquartile range; SD, standard deviation; VR, virtual reality.

**Table 2.** Classifier accuracy metrics for 1-Back and 4-Back class labels with left pupil size as signal input

|                 | Accuracy |      | ROC AUC |      | Sensitivity |      | Specificity |      |
|-----------------|----------|------|---------|------|-------------|------|-------------|------|
| Classifier      | HC       | CP   | HC      | CP   | HC          | CP   | HC          | CP   |
| ROCKET          | 0.87     | 0.64 | N/A     | N/A  | 0.81        | 0.67 | 0.93        | 0.62 |
| HIVE COTE 2.0   | 0.83     | 0.61 | 0.91    | 0.66 | 0.79        | 0.57 | 0.88        | 0.64 |
| DrCIF           | 0.85     | 0.66 | 0.92    | 0.65 | 0.81        | 0.64 | 0.88        | 0.67 |
| Rotation Forest | 0.82     | 0.60 | 0.90    | 0.72 | 0.76        | 0.50 | 0.88        | 0.69 |
| Quant           | 0.71     | 0.56 | N/A     | N/A  | 0.71        | 0.52 | 0.71        | 0.60 |

HC, healthy controls; CP, chronic pain. Sensitivity indicates the proportion of true 4-back cases correctly classified, and specificity indicates the proportion of true 1-back cases correctly classified. ROC AUC is reported only for classifiers that generate probabilistic outputs.

Figures

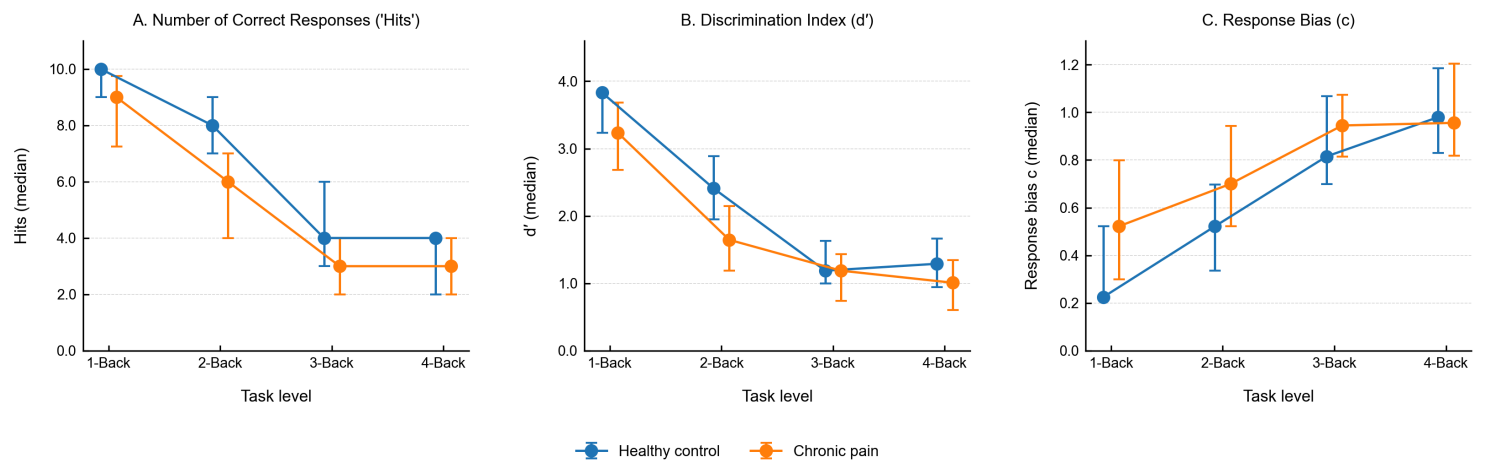


**Figure 1**

Performance indices across tasks and between groups

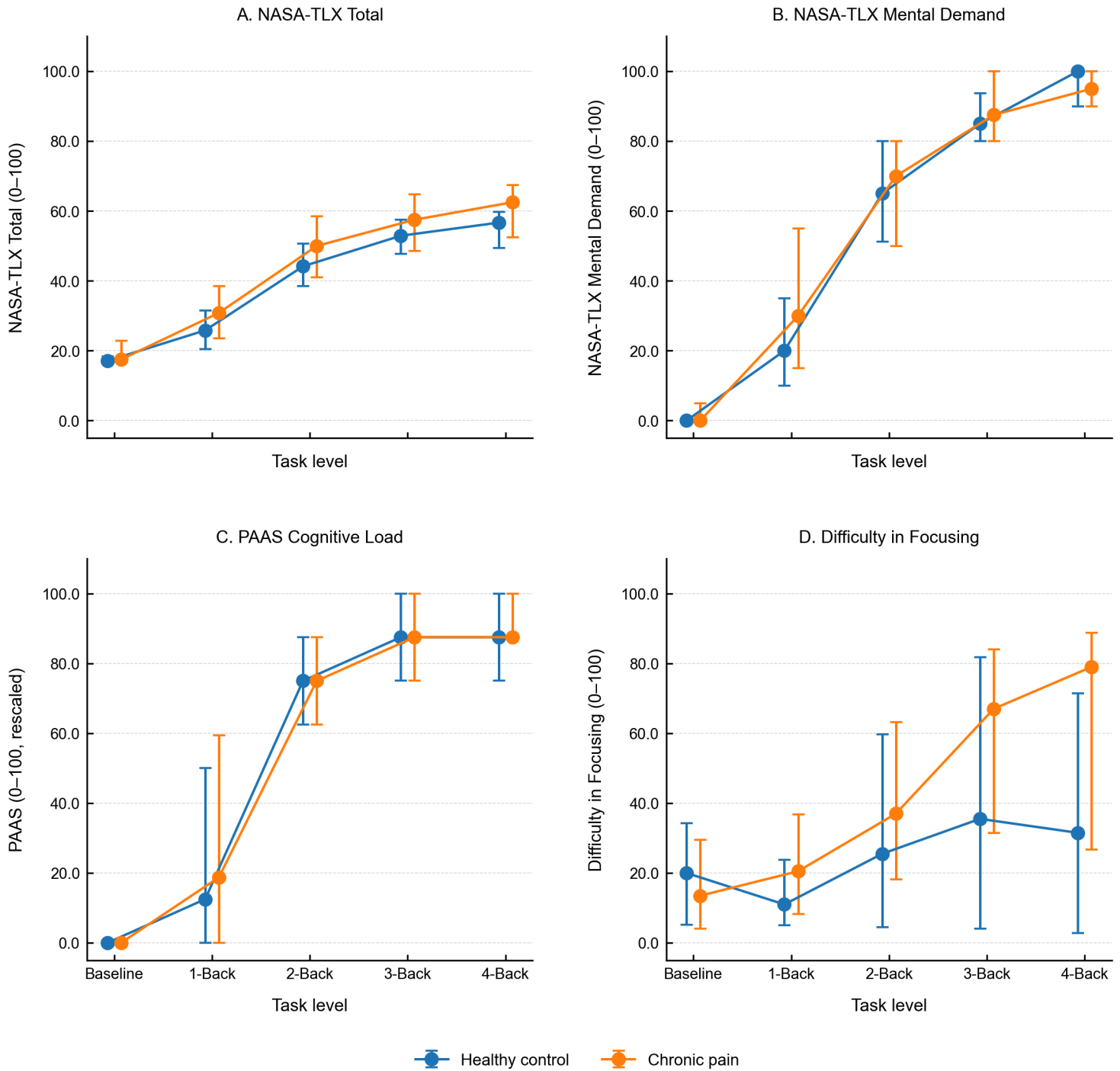


**Figure 2**

Participant-reported post-task measures across tasks and between groups

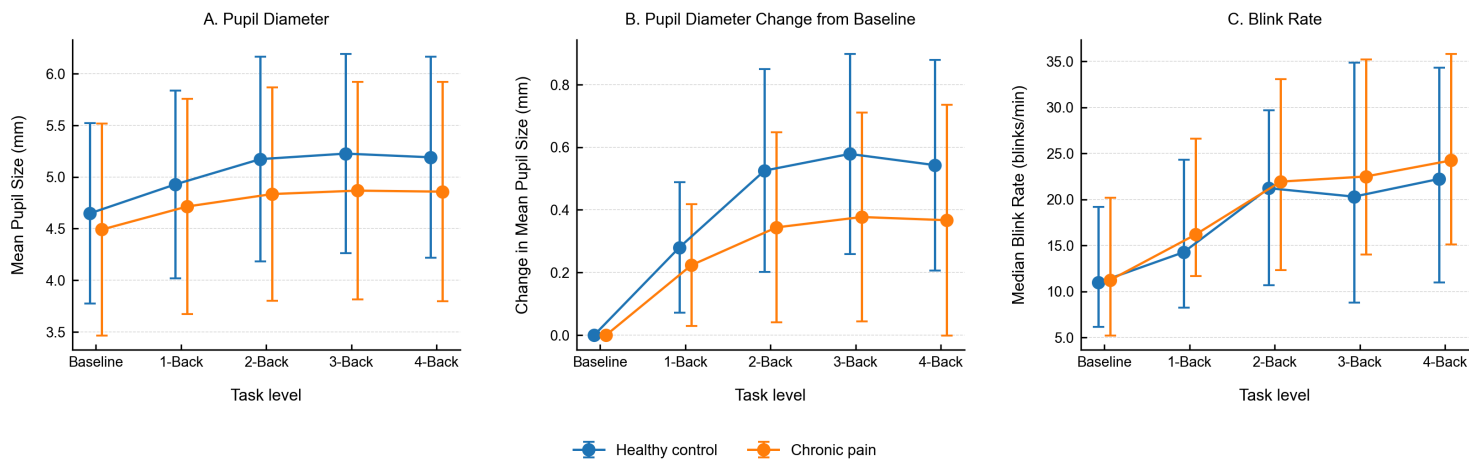


**Figure 3**

Trends in pupil diameter and blink rate across N-back levels and between groups

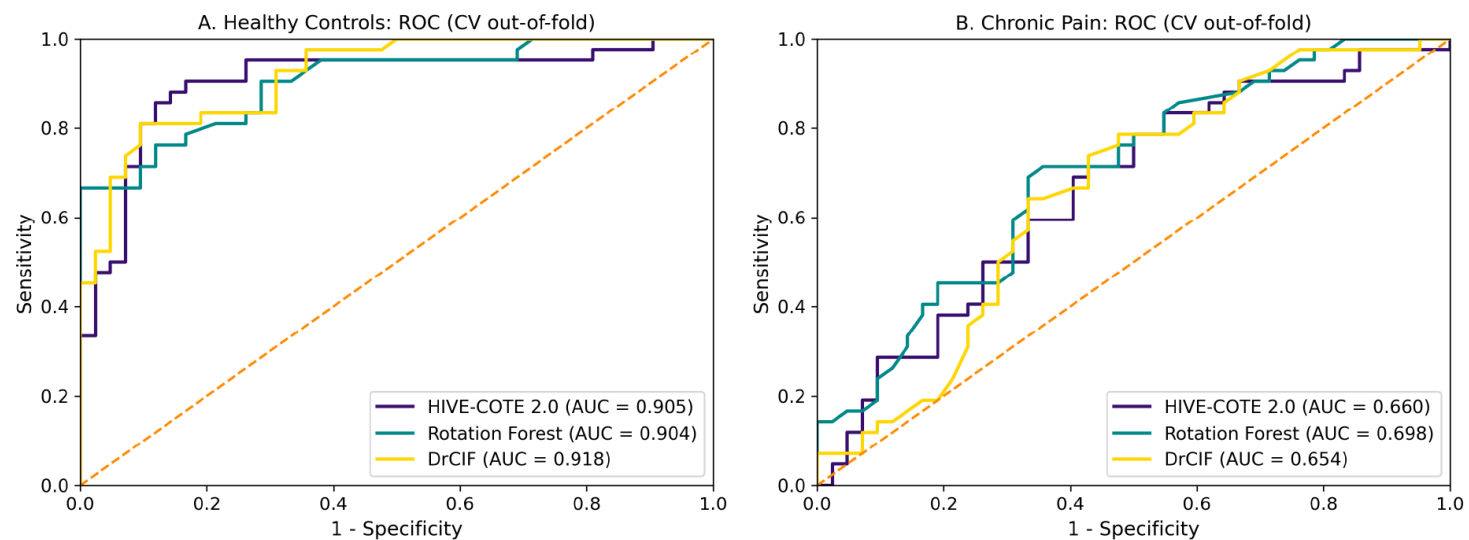


**Figure 4**

ROC curves for 1-Back vs 4-Back across HIVE-COTE 2.0, Rotation Forest and DrCIF classifiers (left pupil size only)

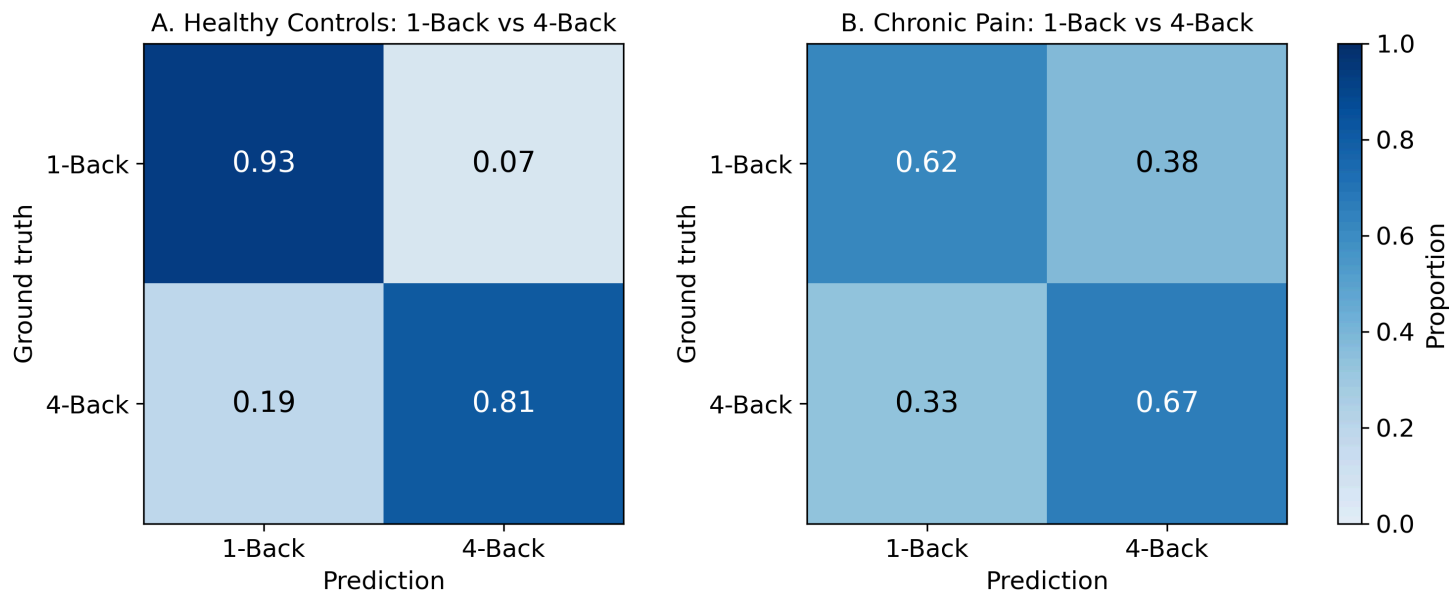


**Figure 5**

Normalised confusion matrices for 1-Back vs 4-Back ROCKET classification (left pupil size only)

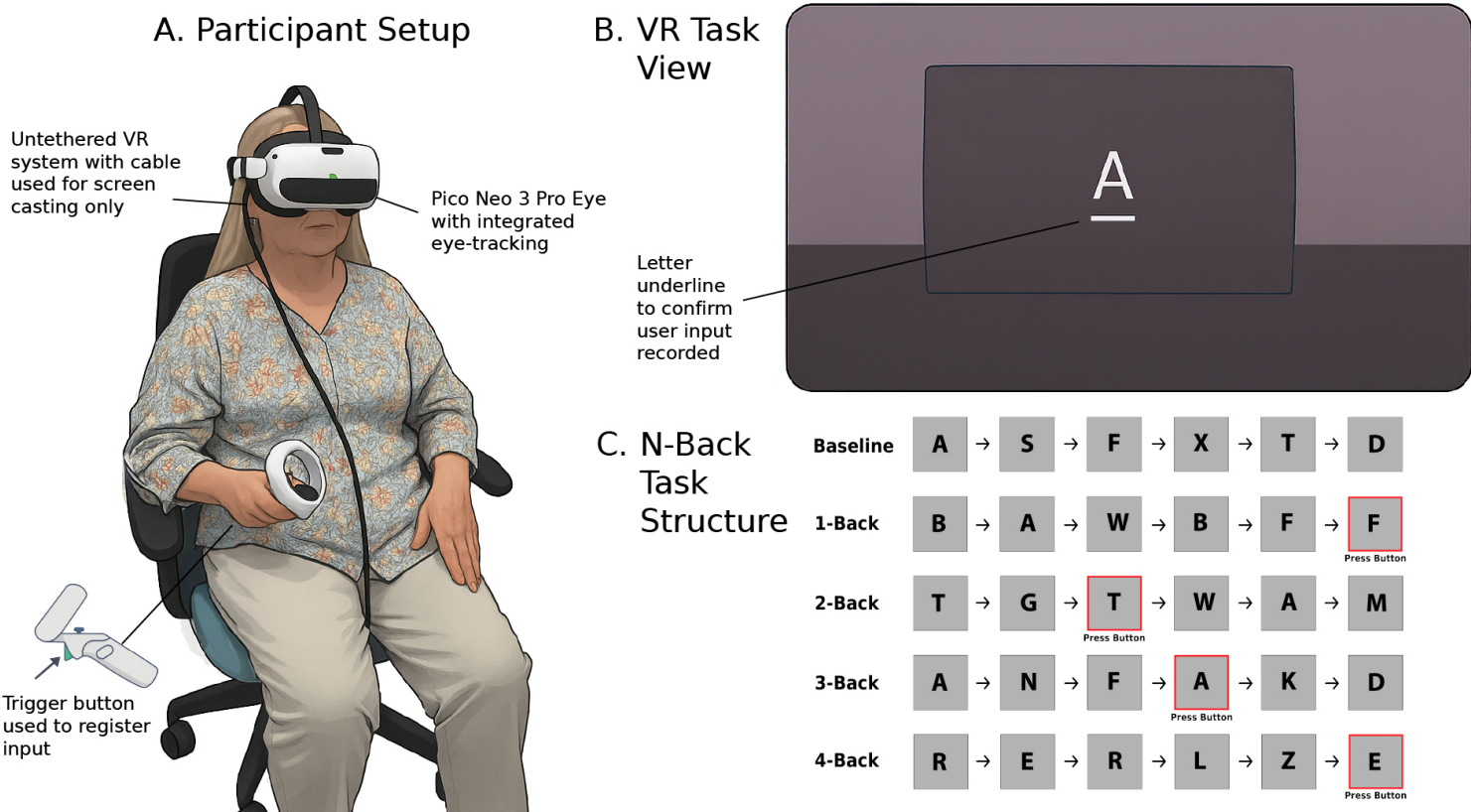


**Figure 6**

Overview of VR-FOCUS experimental setup, in-headset task display, and N-Back procedure

C. The baseline task acted as a control condition with minimal load, the 1-Back task was the least difficult active task and the 4-Back task was the most difficult active task.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformation.docx](#)