

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

Jordan Tsigarides^{ID} · Jennifer O. Bowler · Jack Dainty · Anthony Bagnall · Alexander Macgregor^{ID}

Received: 23 January 2026 / Accepted: 5 October 2026
© The Author(s) 2026

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 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. A total of 84 participants (42 healthy controls, 42 chronic musculoskeletal pain) completed five levels (Baseline to 4-Back) while ocular responses were recorded. Performance declined and subjective workload increased with task level, confirming successful manipulation of cognitive demand, although performance plateaued beyond 3-Back. In mixed-effects models, larger pupil diameter and higher blink rate were associated with task level and workload, without group differences. Logistic regression using summary features discriminated 1-Back from 4-Back at near-chance levels (0.51–0.60). Time-series classifiers exploiting temporal structure performed substantially better in healthy controls (0.74–0.88) but were attenuated in chronic pain (0.58–0.65). Age alone did not appear to explain this difference, although a contribution cannot be excluded. These findings support the feasibility of VR-embedded eye tracking for cognitive load estimation in healthy adults, but performance in chronic pain is currently insufficient for reliable individual-level inference; closed-loop application in this population will require participant-calibrated models and prospective validation.

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^{1,2}. 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 occupation³. 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^{4,5}. Although pharmacological management remains common despite international recommendations, effectiveness

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.



is limited and side effects can be significant^{6–9}. 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^{10,11}. 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^{12–15}, 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 pain¹⁶.

Modern VR platforms often allow users to select different virtual environments, difficulty levels, or task types, providing a limited form of personalisation. However, these choices depend on users accurately assessing and communicating their cognitive and emotional state during each session. Self-report remains the reference standard for subjective workload, but it is difficult to use as a continuous control signal for an adaptive system. Repeated ratings may interrupt immersion, increase demands on the user who is already fatigued, and become less reliable through repeated administration. A more sophisticated approach would use a closed-loop system in which the VR content dynamically adapts to the user’s ongoing physiological or cognitive state¹⁷. Such a system could combine real-time biosignals with intermittent self-report to identify changes in cognitive load or emotional engagement, enabling automated adjustments to maintain an optimal therapeutic state¹⁸.

Cognitive load represents a particularly promising target for such closed-loop modulation. The concept of flow¹⁹; 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^{20–22}. 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^{23,24}. In non-VR research, pupil dilation and blink suppression have been consistently associated with higher cognitive load across a range of controlled tasks^{25–27}. Despite this, there is limited evidence that these relationships persist within immersive VR environments^{28,29}. VR introduces measurement conditions that differ from conventional desktop pupillometry, including near-eye cameras operating at oblique angles, gaze-angle-dependent artefacts in apparent pupil size, variation in screen luminance and vergence demands, and consumer-available sensors that have been less extensively characterised than laboratory systems²⁴. Because an adaptive VR intervention would ultimately operate using a head-mounted display, candidate markers should be evaluated in that environment and using hardware representative of the intended delivery platform.

Evidence obtained from healthy people may not transfer directly to people living with chronic pain. Chronic pain has been associated with altered attentional processing and working-memory impairment, while fatigue, pain intensity, comorbidities, and commonly used medications may also affect autonomic and ocular responses³⁰. These factors could alter both the relationship between cognitive demand and eye-tracking measures and the ability of predictive models to generalise to an unseen user. Evaluating these relationships directly in a pragmatic chronic pain cohort is therefore necessary before eye tracking can be considered as a control signal for adaptive VR pain interventions.

Machine learning offers a means of modelling complex ocular signals for cognitive workload estimation. Previous studies have classified workload using engineered summaries of pupil diameter, blinks, fixations and saccades, including prediction in unseen participants during an N-back task^{31,32}. More recent approaches have

sought to retain temporal information that may be lost when recordings are reduced to block-level summaries, including subject-independent classification from short sequences of eye-region data and temporally segmented representations of pupil diameter^{33,34}. Within immersive VR, recent work used headset-derived ocular and physiological time series data to estimate cognitive load in a large cohort with independent participant-level testing²⁹. Toolkits such as AEON³⁵ now provide open access, scalable implementations of convolutional³⁶, interval-based³⁷ and ensemble³⁸ architectures, allowing these approaches to be evaluated against conventional models within a single framework. However, few studies have directly compared summary-feature models with state-of-the-art time-series classifiers using the same HMD-derived recordings, and none to our knowledge has evaluated performance in chronic musculoskeletal pain.

Building on this rationale, the present study, conducted under the short title VR-FOCUS, investigated 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. This study therefore provides, to our knowledge, the first evaluation of headset-embedded eye tracking as a marker of cognitive load in chronic musculoskeletal pain. It compares summary-feature models directly against state-of-the-art time-series classifiers trained on the same recordings, isolating the contribution of temporal structure rather than of measurement or task design. Every model is evaluated under leave-one-subject-out cross-validation on consumer hardware, establishing how far these signals generalise to an unseen user, the property on which closed-loop application would depend.

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% ($n=11$)

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%)	19 (45%)
University	33 (79%)	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.

reported an eye condition such as dry eyes or Sjogren's syndrome. Opioid use was reported in 45% (oral morphine equivalent 2.11 ± 6.49 mg, range 1.5–40 mg 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 ± 1.8 and 5.6 ± 2.4 respectively. Only four participants (10%) reported severe pain ($\text{NRS} \geq 7/10$). PROMIS Pain Intensity T-scores were mildly elevated relative to population norms (51.7 ± 5.9), whereas Pain Interference was in the moderate range (61.6 ± 8.3). Pain catastrophising levels were modest overall, with mean PCS scores (20.1 ± 13.5) below commonly used thresholds for high catastrophising (>30).

VR eye tracking calibration metrics were available for all 84 participants. Final calibration was obtained on the first attempt in 58 participants (69.0%), and 80 participants (95.2%) met all predefined criteria. Mean final precision error was 0.25° (SD 0.13, range 0.12–0.86), while mean final bias error was 0.62° (SD 0.27, range 0.28–1.66). Four participants did not meet the final precision criterion, but these were at individual calibration points while their mean precision remained within thresholds. For these participants, calibration was reviewed visually before task commencement and judged acceptable, and these participants were retained.

Blink-related interpolation accounted for a mean of 7.37% of task time (SD 4.27, range 0.59–21.49). Longer missing segments exceeding 500ms accounted for a mean of 0.14% of task time (SD 0.32, range 0.00–2.32), with 31 participants (36.9%) having at least one such segment. No participants or task blocks were excluded because of inadequate eye-tracking quality.

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 (Fig. 1).

When comparing the groups, healthy controls achieved more correct responses (“hits”) at all task levels (Fig. 1, Panel A). Ability to discriminate between true targets and false alarms (d') showed a similar pattern, except at the 3-Back level (Fig. 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 (Fig. 1, Panel C).

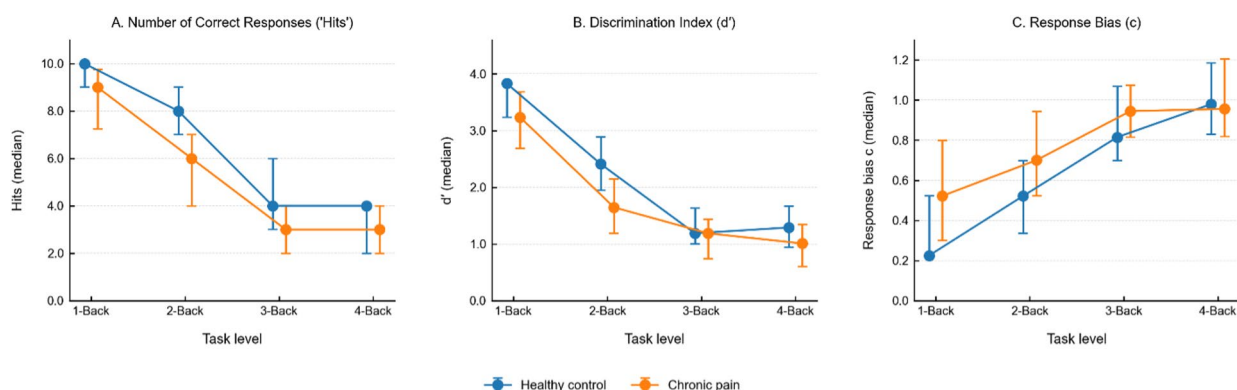


Fig. 1 Performance indices across tasks and between groups.

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.

Holm-corrected pairwise contrasts showed significant changes in performance between successive levels up to 3-Back, but between 3-Back and 4-Back hits ($p=0.051$), discrimination index ($p=0.178$) and reaction time ($p=0.564$) did not differ, indicating a plateau in measurable performance at the highest task levels (Supplementary Table 8).

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 (Fig. 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.

Subjective workload also varied significantly across task levels (Friedman tests, NASA-TLX and PAAS both $p<0.001$). In contrast to performance, workload continued to increase significantly between 3-Back and 4-Back (NASA-TLX $p=0.004$; PAAS $p=0.001$), indicating that participants experienced rising effort beyond the point at which measurable performance had plateaued.

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, Fig. 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, Fig. 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 (β -coefficient 0.009, 95% CI 0.009–0.010, $p<0.001$) and PAAS scores (β -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 (β -coefficient 0.187, 95% CI 0.139–0.234, $p<0.001$) and PAAS (β -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 +0.252 mm; 2-Back +0.435 mm; 3-Back +0.478 mm; 4-Back +0.455 mm; all $p<0.001$), with most of the increase occurring by 2-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.

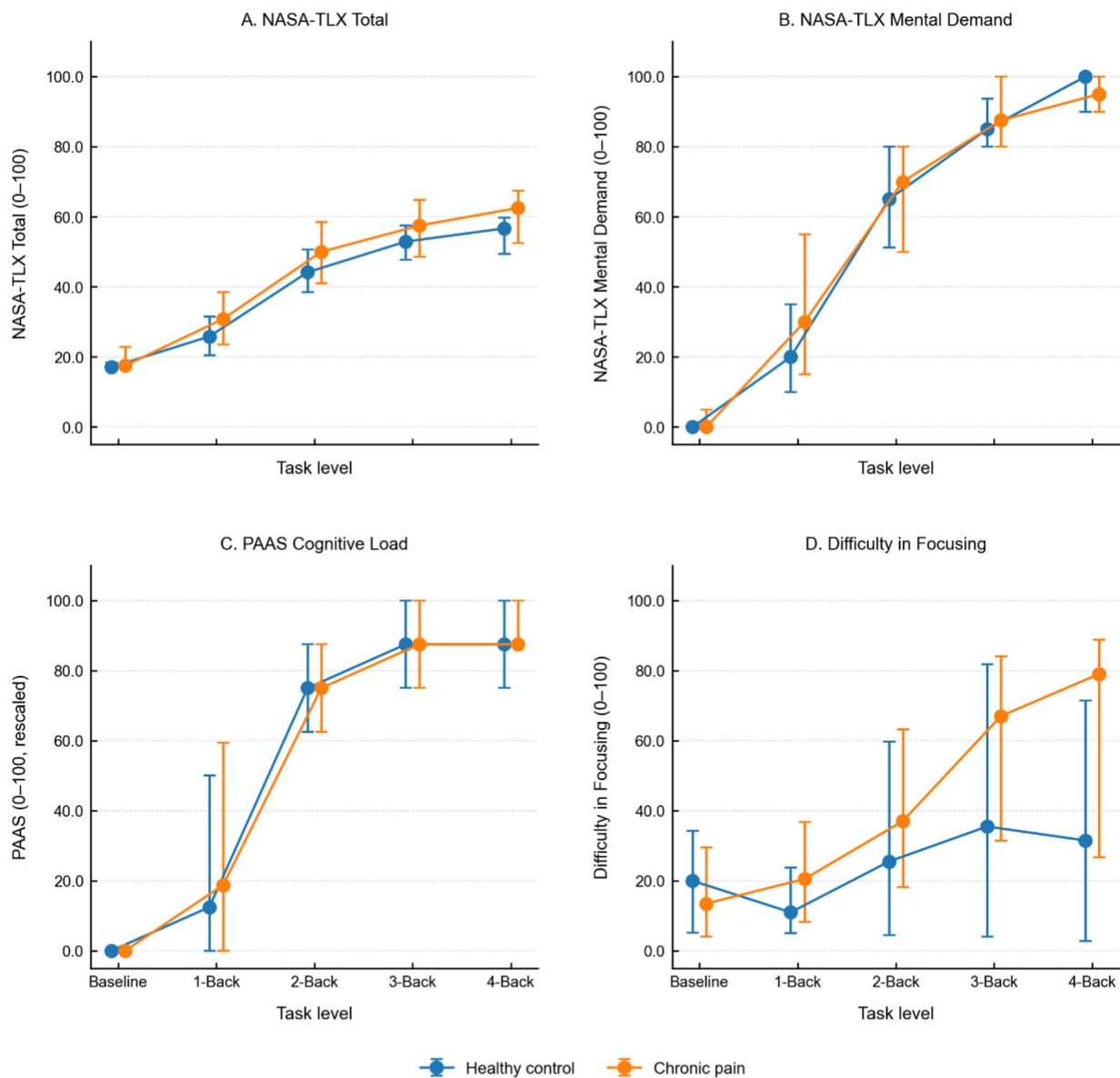


Fig. 2 Participant-reported post-task measures across tasks and between groups.

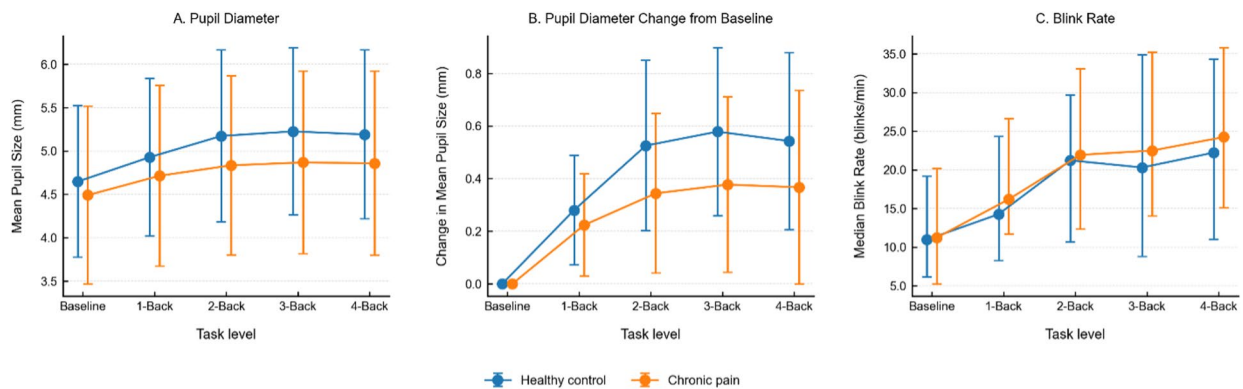


Fig. 3 Trends in pupil diameter and blink rate across N-back levels and between groups.

Age was negatively associated with pupil diameter ($p=0.012$) and positively associated with 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 $p<0.001$).

After adjustment for the other covariates, there was no clear evidence of an association between N-back task order and pupil diameter ($p=0.086$).

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% CI 0.45–0.66, AUC 0.57) in healthy controls and 0.51 (95% CI 0.41–0.62, AUC 0.53) in the chronic pain group. Including blink rate provided only a small improvement, with accuracy reaching 0.60 in both groups (95% CI 0.49–0.69 in both groups; AUC 0.62 and 0.63 respectively). 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.

Accuracy across classifiers ranged between 0.74 and 0.88 in healthy controls and 0.58–0.65 in the chronic pain group using left pupil diameter alone (Table 2). Confidence intervals in the chronic pain group approached chance for the weaker classifiers (e.g. QUANT 95% CI 0.48–0.68), indicating limited individual-level discriminative capacity in this cohort. A comparison of ROC curves between groups for classifiers that produce probabilistic estimates is shown in Fig. 4. Given its computational efficiency, consistent behaviour across folds and performance, ROCKET was selected as the primary exemplar, with corresponding contingency tables presented in Fig. 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 a blink mask as an additional time-series feature had divergent effects between groups. In healthy controls, accuracy fell from 0.88 (95% CI 0.79–0.93) to 0.79 (95% CI 0.69–0.86), whereas in the chronic pain group it rose from 0.64 (95% CI 0.54–0.74) to 0.69 (95% CI 0.59–0.78).

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

Classifier	Accuracy		ROC AUC		Sensitivity		Specificity	
	HC	CP	HC	CP	HC	CP	HC	CP
ROCKET	0.88	0.64	N/A	N/A	0.86	0.67	0.90	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	0.65	0.92	0.65	0.81	0.64	0.88	0.67
Rotation Forest	0.82	0.60	0.91	0.70	0.79	0.50	0.86	0.69
QUANT	0.74	0.58	0.76	0.63	0.76	0.55	0.71	0.62
Logistic Regression	0.56	0.51	0.57	0.53	0.52	0.50	0.60	0.52

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.

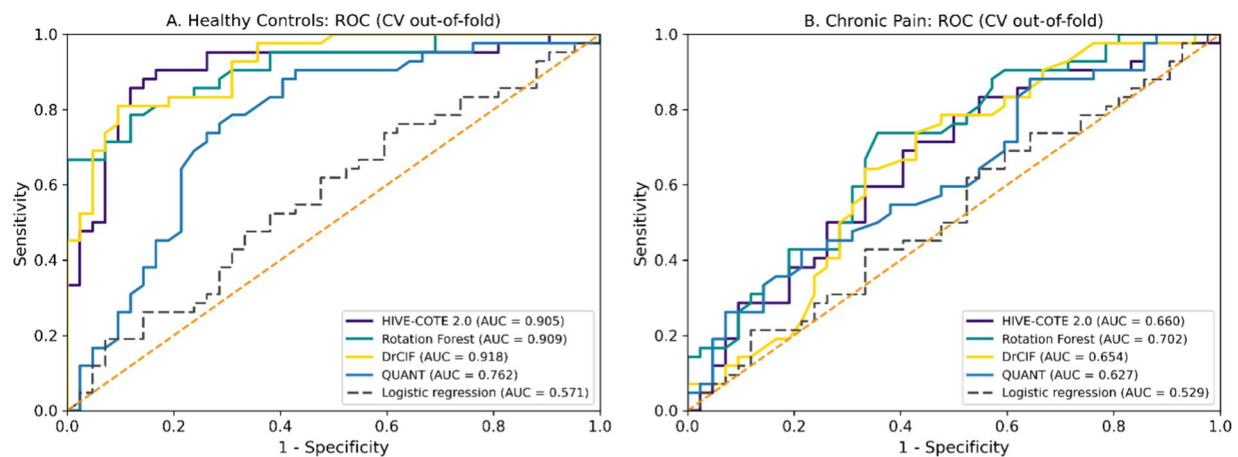


Fig. 4 ROC curves for 1-Back vs. 4-Back across HIVE-COTE 2.0, Rotation Forest, DrCIF, QUANT and logistic regression classifiers (left pupil size only).

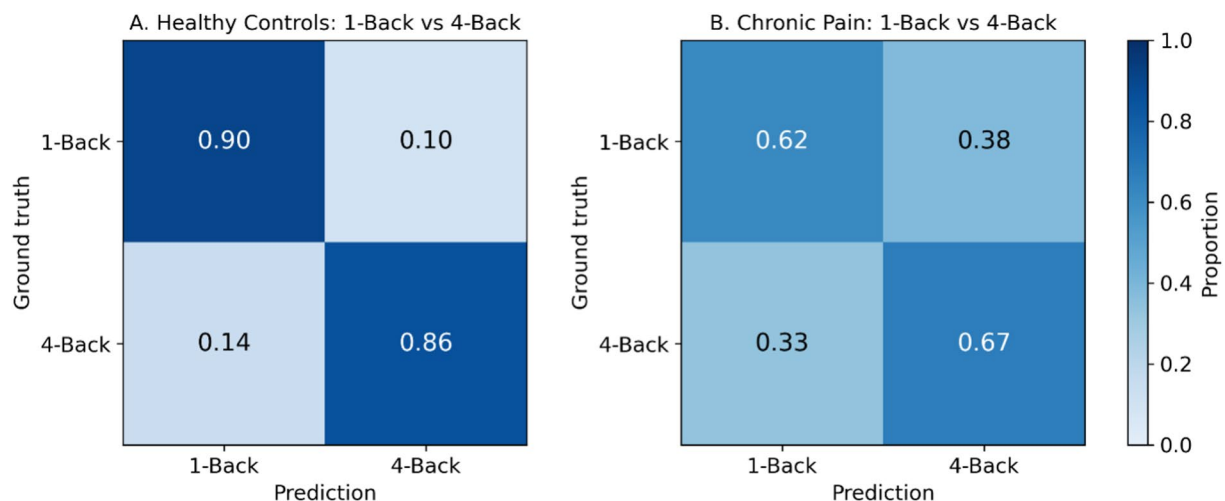


Fig. 5 Normalised confusion matrices for 1-Back vs. 4-Back ROCKET classification (left pupil size only).

Sensitivity analyses

The age difference between groups prompted three complementary sensitivity analyses assessing whether age could account for the difference in classification performance. When participant-level accuracy was modelled as a function of age and group, age was not associated with accuracy ($\beta = -0.0010$, $p = 0.646$), whereas group remained associated after adjustment ($\beta = 0.215$, $p = 0.004$), with no age-by-group interaction ($p = 0.296$). Participants were then optimally matched one-to-one by age, with a maximum permitted difference of two years, producing 14 pairs (mean within-pair difference 0.9 years). Among matched participants, accuracy from the primary analysis remained higher in healthy controls (0.86, 95% CI 0.69–0.94) than in chronic pain (0.68, 95% CI 0.49–0.82). Retraining the classifiers within each matched subgroup reduced accuracy in both groups (0.68 and 0.64), an effect attributable to the smaller training sets rather than to age matching. Across 50 randomly drawn, age-unrestricted subgroups of the same size, accuracy averaged 0.74 and 0.58 respectively, and the between-group difference persisted ($p < 0.001$; Supplementary Table 13). These analyses did not indicate that age alone accounted for the attenuated performance in chronic pain, although the matched sample was small and a contribution of age cannot be fully excluded.

Two analyses evaluated whether classification reflected the assigned load conditions. Because Baseline and 1-Back always appeared first, the analysis was repeated using only the counterbalanced levels. Point estimates for 2-Back versus 4-Back classification were above chance in both groups (0.61 and 0.56), although accuracy was lower than for the 1-Back versus 4-Back contrast and the chronic pain estimate remains uncertain (95% CI 0.45–0.66). This suggests that discrimination did not depend solely on the fixed position of the 1-Back condition. In a label-permutation sensitivity analysis, the mapping between each participant's two epochs and their labels was randomly exchanged and the full cross-validation repeated across 100 permutations. Permuted accuracy fell to chance in both groups, and observed accuracy exceeded every permuted value in healthy controls ($p < 0.01$) and all but one in chronic pain ($p = 0.02$), though the margin over chance was considerably narrower in chronic pain (Supplementary Table 9).

To test whether ocular signals supported graded rather than binary discrimination, classifiers were fitted to all four active N-back levels (Supplementary Table 10). In healthy controls, exact four-way accuracy was 0.46 (chance 0.25) and 0.79 of cases fell at the correct or an adjacent level (chance 0.63), with predicted level correlating positively with true level ($\rho = 0.49$, $p < 0.001$); recall was highest for the extreme conditions (1-Back and 4-Back) and lowest for intermediate levels. Performance was weaker in chronic pain, where exact accuracy was 0.35, adjacent accuracy 0.67 exceeded its chance level of 0.63 by only a small margin, and the correlation between predicted and true level was not significant ($\rho = 0.15$, $p = 0.056$). Ocular signals therefore contained graded information about cognitive load in healthy adults but provided little evidence of ordinal discrimination in chronic pain.

Two sensitivity analyses examined the robustness of the classification results. Repeating the primary analysis with right-eye pupil diameter produced similar results in both groups (0.83 and 0.60), differing from left-eye data by 0.048 in each group with confidence intervals including zero. Within the chronic pain group, participant-level accuracy did not differ significantly by opioid use, anticholinergic medication, eye-drop use, ocular conditions, antidepressant use or diagnostic category, and weak negative correlations with comorbidity and medication counts did not reach significance (Supplementary Table 11). Medication exposure and diagnosis therefore do not appear to explain the group difference, although these comparisons are underpowered and a modest contribution of overall clinical complexity cannot be excluded.

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-available, 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, although performance reached a plateau between the 3-Back and 4-Back levels while subjective workload continued to rise. In parallel, pupil diameter and blink rate varied systematically with task level and workload

under stable luminance. Models using only 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. Sensitivity analyses in age-matched subgroups, together with a random-subset control, did not indicate that the age disparity between cohorts alone accounted for this difference, although a contribution of age cannot be fully excluded. Overall, the results support VR-embedded eye tracking as a feasible biosignal for cognitive load estimation in healthy adults, but suggest that current performance in chronic pain is insufficient for reliable individual-level inference and that translation to closed-loop applications in this population will require participant-calibrated modelling and prospective evaluation.

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^{39,40}. 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^{25,39,40}. 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^{41,42}. 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 managed⁴³. 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^{23,44}. Together with prior VR studies indicating that pupil dynamics can track cognitive load when luminance is controlled or modelled^{28,45}, 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, although its interpretation is more task-dependent than pupil dilation. Blink suppression is frequently reported when continuous visual sampling demands are high^{46,47}, whereas increased blink rate has been reported during sustained mental effort and internally focused processing^{48,49}. The N-back task combines discrete, centrally presented stimuli with internal maintenance and updating of working memory between stimuli, making an increase in blink rate plausible. Blink rate rose from Baseline to 2-Back before plateauing and was higher overall in the chronic pain group. While the difference in groups was not explained by pain intensity or fatigue in the present analyses, it may reflect a combination of broader clinical factors present in the chronic pain group (for example, medication exposure, comorbidity burden, autonomic regulation, or ocular surface symptoms)⁵⁰. Temporal blink information also contributed differently to classification between groups. With the prespecified exemplar classifier, ROCKET, adding the binary blink-indicator channel reduced classification accuracy in healthy controls but increased it in participants with chronic pain. Although the mechanism cannot be determined from the present data, this dissociation suggests that the value of blink dynamics varies between cohorts. Aggregate blink measures therefore appear unsuitable as standalone markers of cognitive load in VR, but temporal blink information may contribute to multivariate models when task structure and population characteristics are considered.

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^{51,52}. 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 resources⁵³.

Physiologically, the absence of marked group differences in task-related pupil and blink responses suggests that the chronic pain cohort did not show a simple uniform attenuation of ocular responsivity. Nevertheless,

time-series classifiers discriminated low from high load less accurately and with greater inter-individual variability in this group. Population-level association and individual-level prediction are distinct properties. Mixed-effects models estimate average condition-related changes across repeated observations, whereas classification asks whether data from an unseen participant can be assigned to the correct condition. This depends on the separation and overlap of the condition-specific distributions, not on a difference between their means. The higher accuracy of time-series classifiers relative to summary-feature logistic regression is consistent with informative temporal structure being lost through aggregation. When comparing time-series classifiers, accuracy was broadly comparable across ROCKET, DrCIF, HIVE-COTE 2.0 and Rotation Forest, which represented time-resolved information through convolutional responses, characteristics of localised intervals, heterogeneous temporal representations or values at fixed temporal positions^{36–38,54}. QUANT, which compresses fixed multiscale intervals primarily into quantile summaries, produced the lowest point estimate of accuracy in both cohorts. Although quantiles discard the ordering of values within each interval, QUANT retains interval location and incorporates transformed representations of the series. The observed pattern is therefore compatible with richer local temporal morphology carrying discriminative information beyond aggregate distributional properties, but it does not provide a direct test of this explanation⁵⁵. The classifiers were not formally compared, HIVE-COTE 2.0 was trained under an eight-minute contract per fold, and differences in model flexibility may also have contributed.

Age-adjusted and age-matched analyses did not indicate that age alone accounted for the poorer classification performance, although the remaining demographic and clinical differences prevent this pattern from being attributed specifically to chronic pain. One possibility is that heterogeneity within the chronic pain cohort increased variability in the temporal coupling between cognitive demand and ocular responses rather than producing a uniform shift in mean physiology. An alternative explanation is that the signal-to-noise ratio of headset-derived ocular measures was insufficient under the present acquisition and modelling conditions. These explanations are not mutually exclusive. Pupil-only accuracies of 0.58–0.65, with confidence intervals including chance for the weaker classifiers, do not currently support reliable individual-level inference in this cohort. Distinguishing between these explanations will require larger, more deeply phenotyped samples, repeated within-person measurements across symptom states, and direct comparison of population-level and participant-calibrated models. We therefore interpret the chronic pain findings as demonstrating measurement feasibility while identifying a performance gap, rather than as evidence of a deployable marker.

This distinction between measurement feasibility and deployability is clinically important. 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^{12,15,56–58}. Most deployed or evaluated VR interventions remain open-loop, with fixed content that does not adapt to patients' cognitive load, symptom burden or fatigue.

Although the present results do not establish readiness for clinical deployment, they identify a potential pathway towards VR systems that monitor cognitive load and adapt task demands. A simple closed-loop system could use binary discrimination between lower and higher load to make coarse adjustments or provide recommendations regarding task difficulty or other characteristics. However, a more flexible system would ideally distinguish several ordered levels of demand, without needing to recover the nominal task label or estimate an exact continuous value. The multiclass analysis provided modest evidence of this graded information in healthy controls, where predicted and true levels were correlated and most predictions were correct or adjacent to the true level, although intermediate levels were frequently confused. Evidence of ordering was weaker in the chronic pain group, where the correlation was not statistically significant and adjacent-level accuracy exceeded chance by only a small margin. The binary findings therefore support the feasibility of threshold-based adaptation, while the multiclass results identify the further development required for more nuanced control. Future systems could combine pupillometry, blink dynamics and, where available, performance information to estimate whether load

is below, within or above a personalised therapeutic target. They might increase interaction when load is below this range or simplify the experience when overload is detected. In chronic pain populations, where cognitive strain and symptoms may fluctuate, the observed variability motivates algorithms that accommodate individual heterogeneity rather than assuming a fixed relationship 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^{59,60}. 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 cognitive behavioural therapy 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 setting rather than 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 all preprocessing parameters, including winsorisation and normalisation, were estimated within cross-validation training folds and applied unchanged to held-out participants to avoid information 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. The interpretation is further supported by sensitivity analyses, in which discrimination persisted when restricted to counterbalanced task conditions, did not depend on the choice of pupil, and was not explained by medication exposure or diagnostic category within the chronic pain cohort.

Limitations remain. The groups differed in age, education, ethnicity, prior VR exposure, comorbidity burden and medication usage. Sensitivity analyses did not indicate that age alone accounted for the difference in classification performance, but the remaining demographic and clinical differences cannot be fully disentangled in a sample of this size. The single-session design precludes conclusions about within-person stability across days and fluctuating symptom states. The chronic pain cohort was pragmatic and clinically heterogeneous, including multiple diagnoses, comorbidities, and medication exposures that may influence cognition and ocular physiology. Potentially relevant exposures included use of opioids (45%), medications with anticholinergic effects (29%), antidepressants (31%), and reported ocular conditions (26%). Exploratory analyses did not identify associations between these factors and classification accuracy, but the study was not powered to exclude modest effects. The included chronic pain cohort should therefore not be interpreted as representing a single phenotype, and larger studies with stratified recruitment and richer phenotyping are needed. Pain severity and impact were generally low-to-moderate, further 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. Baseline and 1-Back were always presented first, meaning that the primary 1-Back versus 4-Back contrast may partly reflect order-related effects despite the counterbalanced 2-Back versus 4-Back sensitivity analysis. On-screen luminance was stable across task blocks, but therapeutic VR will involve greater visual and luminance variability that will need to be modelled explicitly to support future real-world closed-loop use. Ambient room lighting was not standardised, although an occlusive VR headset with individually adjusted fit should have limited external light exposure. Sample size was determined pragmatically rather than by formal power calculation, providing uncertainty related to estimates, particularly for subgroup analyses. Confidence intervals are therefore reported.

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. Feature attribution is also limited due to transformed and ensemble representations used by several classifiers. 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-available 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, although individual-level performance in chronic pain remains limited. Participant-calibrated, externally validated approaches are the next step towards 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 Register⁶¹ and the VIPA study¹⁵. Specific inclusion criteria for participants with chronic pain included current pain lasting ≥ 3 months at the time of recruitment⁶.

General inclusion criteria for both groups were: age ≥ 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.

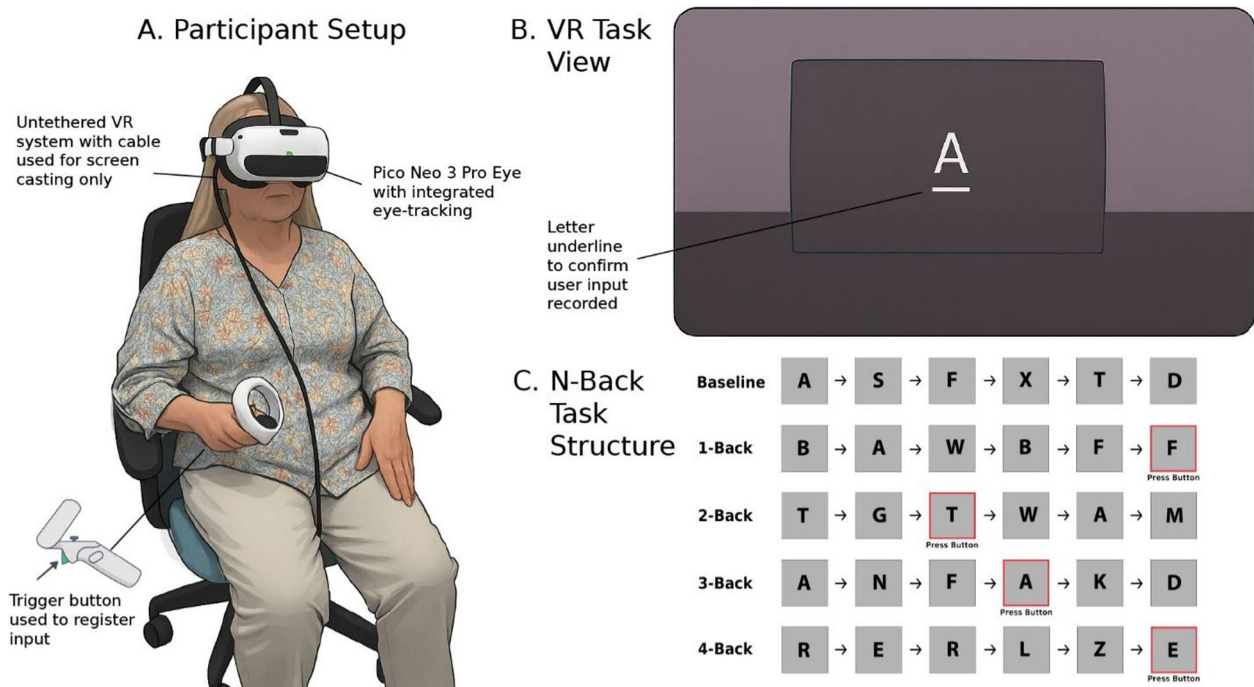


Fig. 6 Overview of VR-FOCUS experimental setup, in-headset task display, and N-back procedure. The baseline task acted as a control condition with minimal load, the 1-Back task was the least difficult active task and the 4-Back task was the most difficult active task.

Apparatus and virtual environment

The consumer-available Pico Neo 3 Pro Eye[®] VR system (Pico Interactive, San Francisco, USA)⁶² containing an integrated Tobii[®] binocular eye-tracking system (sampling rate 60–90 Hz) was used to deliver the VR N-back tasks. This included dual 1832 × 1920 pixel LCD display panels, 90 Hz 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[®] (v2021.3.45f1, Unity Technologies, San Francisco, USA)⁶³ with use of the Tobii Ocumen SDK⁶⁴ 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, and participants remained seated on a comfortable chair during VR use (Fig. 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. Calibration attempts and final calibration status were retained as data-quality audit measures. Where repeated calibration did not meet every predefined threshold, the calibration output was reviewed at acquisition and the participant proceeded only when the residual deviation was judged acceptable for the centrally presented task.

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 (Fig. 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^{65,66}, 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 (Fig. 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 ($Y = 0.2126R + 0.7152G + 0.0722B$)^{67,68}. 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 (task-level means 0.376–0.378; coefficient of variation <1%), and across participants ranged from 0.352 to 0.401 with a maximum within-block coefficient of variation of 2.8%. Ambient room lighting was not formally standardised, although testing was conducted in comparable rooms within a single facility. The head-mounted display used has an occlusive facial interface, and headset fit was individually checked for each participant, substantially limiting the impact of ambient light.

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 was underlined to provide input feedback (Fig. 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. Despite this, given the risk of order effect influencing 1-Back versus 4-Back analyses, we included condition order as a covariate in all mixed-effects models and conducted a sensitivity analysis restricted to the counterbalanced task levels (Supplementary Table 9).

After each task, participants completed brief VR questionnaires using the handheld controller. Following all tasks, participants removed the VR system 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)⁶⁹ and PROMIS Pain Interference (Short Form 8a)⁶⁹ instruments, providing structured assessments of pain severity and its disruption to day-to-day activity. They also completed the Pain Catastrophising Scale⁷⁰ 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^{71,72}. 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 Index⁷³, which includes six dimensions rated from 0 to 100, and an adapted version of the one-item PAAS scale⁷⁴, 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 Questionnaire⁷⁵. 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 study's exploratory nature and the use of machine learning analyses, a formal power calculation was not performed a priori. A target sample of 30–50 participants per group was selected to provide participant-level variability while remaining feasible for the VR-based protocol, with 42 participants ultimately recruited to each group. For classification, each cohort contributed 84 balanced epochs, with each leave-one-subject-out fold trained on 82 epochs from 41 participants. Contemporary benchmarks using aeon classifiers have demonstrated competitive performance on datasets with fewer than 100 training cases⁷⁶. Preprocessing and model fitting were restricted to training folds, maintaining the participant-level separation recommended for small-sample machine learning validation⁷⁷. The sample was considered adequate for exploratory feasibility assessment, but not definitive clinical validation or subgroup characterisation.

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 3⁷⁸ environment. Data were acquired at 60–90 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 latency⁷⁹.

Blink events were defined as gaps of 50–500ms⁸⁰ and reconstructed using a cubic-spline interpolation with linear fallback⁸¹; longer artefacts were removed. Processed eye-tracking quality was evaluated using structured audit outputs generated during preprocessing. Signal completeness was quantified for each participant as the proportion of task time reconstructed through interpolation of blink-related gaps lasting 50–500ms and the proportion represented by longer missing segments exceeding 500ms. Longer segments were not interpolated and were treated as missing during subsequent processing. Audit outputs were reviewed at participant and task-block level, and any exclusions attributable to inadequate eye-tracking quality were recorded.

Although binocular data were recorded, left-eye pupil diameter was selected consistently as the primary signal before inferential and classification analyses. This decision followed pre-analysis review of signal quality and was applied uniformly across participants rather than selecting the better-performing eye individually. Using one prespecified channel also avoided combining signals with different offsets and noise characteristics. All classification analyses were repeated using right-eye pupil diameter as a sensitivity analysis.

Unweighted NASA-TLX scores were used. The PAAS item was presented in VR on a 0–9 interface owing to a constraint within Unity. Recorded responses of 0 were recoded as 1, producing an analysed range of 1–9. Inferential analyses used these 1–9 scores throughout. For graphical presentation only, PAAS scores were linearly rescaled to 0–100 using $(\text{score}-1)/8 \times 100$. 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^{71,72}.

Statistical analysis

Conventional analyses were conducted in R (v4.4.3)⁸² and in Python (v3.12)⁸³. 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.

Within-participant differences across task levels were assessed using Friedman tests, with Kendall's W reported as the omnibus effect size. The number of hits, discrimination index (d'), response bias (c) and mean hit reaction time were compared across 1-Back to 4-Back; NASA-TLX and PAAS were compared across Baseline to 4-Back. Significant omnibus tests were followed by Wilcoxon signed-rank tests on adjacent levels only, with matched-pairs rank-biserial correlation as the pairwise effect size. Holm correction was applied separately within each measure across that measure's adjacent contrasts. No correction was applied across measures or across the six Friedman tests.

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 β -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

The primary predictive analysis was defined as binary classification of 1-Back versus 4-Back using left-eye pupil diameter. These conditions represented the widest separation in active N-back load. Models were fitted separately within the healthy control and chronic pain groups. The number of stimuli, stimulus duration and inter-trial interval were identical across task levels, so these timing parameters could not distinguish the class labels.

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 with parameters estimated on training folds only within each cross-validation split. 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)⁸³ using the time series machine learning toolkit *aeon* (v1.1)³⁵. 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 time-series input comprised left-eye pupil diameter alone. In a secondary analysis, a binary blink-indicator channel was added to the pupil-diameter channel to assess whether temporal blink information improved classification. Missing samples were imputed using the per-case, per-feature mean. Outliers were winsorised and data were normalised using per-channel z-scores. All preprocessing parameters were estimated within each cross-validation split using the training data only and applied unchanged to the held-out participant, so that no information from the test participant contributed to any transformation. The same procedure was applied to the summary features used in the logistic regression baseline. 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) were 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 study⁷⁶: Rotation Forest⁵⁴ is a standard classifier benchmark that trains an ensemble of decision trees on rotated feature spaces; ROCKET³⁶ is a pipeline classifier that uses a large set of random convolutional kernels to generate features for a linear classifier. QUANT⁵⁵ and DrCIF³⁸ 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 2.0³⁸ 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 (β) with corresponding two-tailed p-values were reported. Participants were then matched one-to-one between groups on age, with a maximum allowable within-pair difference of two years, yielding 14 matched pairs. Accuracy was first calculated for the matched participants by restricting the pooled out-of-fold predictions from the primary full-cohort models, thereby retaining the original training sets. Classifiers were then refitted separately within each matched subgroup using the same 1-Back versus 4-Back contrast, left-pupil input, leave-one-subject-out structure and foldwise preprocessing as the primary analysis. To distinguish any effect of age matching from the effect of reducing the training sample, 50 age-unrestricted random samples of 14 participants were drawn without replacement within each group. A separate ROCKET classifier was fitted in every sample using the same analysis pipeline. The resulting accuracy distributions were summarised by their means and 2.5th and 97.5th percentiles, and the matched-subgroup accuracy was located within the corresponding empirical distribution.

Sensitivity analyses

Five sensitivity analyses used the same leave-one-subject-out framework and fold-embedded preprocessing as the primary analysis.

Classification of 2-Back versus 4-Back assessed whether discrimination persisted when neither condition had a fixed presentation position.

Multi-class classification assessed discrimination across all four active N-back levels using exact accuracy, accuracy within one level and Spearman correlation (ρ). Chance accuracies were derived from the observed class distribution.

The primary analysis (1-Back versus 4-Back) was repeated using right-eye pupil diameter to assess sensitivity to the choice of eye.

Within the chronic pain group, participant-level accuracy was compared by medication exposure, ocular conditions and diagnosis using Mann-Whitney U tests, and correlated with comorbidity and medication counts using Spearman correlations (ρ). Accuracies from the full-cohort models were used to hold training-set size constant.

A label-permutation control independently exchanged each participant's paired N-back labels with probability 0.5 and repeated the complete pipeline across 100 permutations. This preserved the paired structure and class balance while breaking the association between load and label. One-sided permutation p-values were calculated as $(k+1)/(B+1)$, giving a minimum of 0.0099.

All analyses used ROCKET, the prespecified exemplar classifier; primary results were comparable across the five classifiers (Table 2).

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-026-74975-6>.

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.

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.

Data availability De-identified participant-level data are available to researchers under a data-sharing agreement, subject to approval by the study team. Managed rather than open access is required by the terms of participant consent, which permits sharing of anonymised data with research collaborators undertaking ethically approved research but does not extend to unrestricted public release.

Code availability All analysis code, including the eye-tracking preprocessing pipeline, the time-series classification and cross-validation scripts, the statistical analysis scripts and the figure-generation code, is openly available at [<https://doi.org/10.5281/zenodo.22133095>].

Declarations

Competing interests J.T. holds a 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.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

1. Gill, T. K. 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. *Lancet Rheumatol.* **5**, e670–e682. [https://doi.org/10.1016/S2665-9913\(23\)00232-1](https://doi.org/10.1016/S2665-9913(23)00232-1) (2023).
2. Lurie, J. M. & Javaid, A. Visualizing Global Chronic Pain. *Anesth. Analgesia.* **138**, 918–919. <https://doi.org/10.1213/ane.0000000000006564> (2024).
3. Gatchel, R. J. et al. The biopsychosocial approach to chronic pain: scientific advances and future directions. *Psychol. Bull.* **133**, 581–624. <https://doi.org/10.1037/0033-2909.133.4.581> (2007).
4. Hutchings, A. 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.* **34**, 2291–2300 (2007).
5. Rogers, A. H. et al. Impact of intraindividual pain variability on functional pain outcomes among adults with chronic pain: an ecological momentary assessment study. *J. Behav. Med.* **48**, 1035–1042. <https://doi.org/10.1007/s10865-025-00590-x> (2025).

6. National Institute for Health and Care Excellence. *Chronic pain (primary and secondary) in over 16s: assessment of all chronic pain and management of chronic primary pain* (National Institute for Health and Care Excellence (NICE), (2023).
7. Häuser, W. et al. Review of pharmacological therapies in fibromyalgia syndrome. *Arthritis Res. Ther.* **16**, 201. <https://doi.org/10.1186/ar4441> (2014).
8. Macfarlane, G. J. et al. EULAR revised recommendations for the management of fibromyalgia. *Ann. Rheum. Dis.* **76** (20160704), 318–328. <https://doi.org/10.1136/annrheumdis-2016-209724> (2017).
9. Rivera, J. et al. Opioids and fibromyalgia: frequency of use and factors associated with increased consumption in patients remitted to a tertiary care center. *BMC Musculoskelet. Disord.* **25**, 121. <https://doi.org/10.1186/s12891-024-07263-x> (2024).
10. Pretat, T., Koller, C. & Hügler, T. Virtual reality as a treatment for chronic musculoskeletal pain syndromes. *Joint Bone Spine.* **92**, 105769. <https://doi.org/10.1016/j.jbspin.2024.105769> (2025).
11. Tsigarides, J., Shenker, N. G. & MacGregor, A. Ready Patient One: the role of therapeutic virtual reality in the future management of chronic pain. *Rheumatol. (Oxford)*. **61**, 482–483. <https://doi.org/10.1093/rheumatology/keab594> (2022).
12. Eccleston, C. et al. A prospective, double-blind, pilot, randomized, controlled trial of an embodied virtual reality intervention for adults with low back pain. *Pain* **163**, 1700–1715. <https://doi.org/10.1097/j.pain.0000000000002617> (2022).
13. Garcia, L. M. 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.* **23**, 822–840. <https://doi.org/10.1016/j.jpain.2021.12.002> (2022).
14. Garcia, L. M. 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.* **23**, e26292. <https://doi.org/10.2196/26292> (2021).
15. Tsigarides, J. et al. Feasibility of attention-based virtual reality interventions in fibromyalgia syndrome: comparing systems, virtual environments and activities. *Br. J. Pain.* **20494637241310696**, 20250104. <https://doi.org/10.1177/20494637241310696> (2025).
16. Themelis, K. & Tang, N. K. Y. The Management of Chronic Pain: Re-Centring Person-Centred Care. *J. Clin. Med.* **12** (20231107). <https://doi.org/10.3390/jcm12226957> (2023).
17. Baker, C. & Fairclough, S. H. Chapter 9 - Adaptive virtual reality. In: (eds Fairclough, S. H. & Zander, T. O.) *Current Research in Neuroadaptive Technology*. Academic, 159–176. (2022).
18. Serrano, R. R. M. 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* (Harper-Collins, 1990).
20. Bantick, S. J. et al. Imaging how attention modulates pain in humans using functional MRI. *Brain* **125**, 310–319. <https://doi.org/10.1093/brain/awf022> (2002).
21. Beals, M. *Flow as a Tool to Manage Chronic Pain* (University of California, 2023).
22. Hoffman, H. G. et al. Using fMRI to study the neural correlates of virtual reality analgesia. *CNS Spectr.* **11**, 45–51. <https://doi.org/10.1017/s1092852900024202> (2006).
23. Adhanom, I. B., MacNeilage, P. & Folmer, E. Eye Tracking in Virtual Reality: a Broad Review of Applications and Challenges. *Virtual Real.* **27** (20230118), 1481–1505. <https://doi.org/10.1007/s10055-022-00738-z> (2023).
24. Schuetz, I. & Fiehler, K. Eye Tracking in Virtual Reality: Vive Pro Eye Spatial Accuracy, Precision, and Calibration Reliability. *J. Eye Mov. Res.* **15** (20220907). <https://doi.org/10.16910/jemr.15.3.3> (2022).
25. Beatty, J. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychol. Bull.* **91**, 276–292. <https://doi.org/10.1037/0033-2909.91.2.276> (1982).
26. Van Orden, K. F. et al. Eye Activity Correlates of Workload during a Visuospatial Memory Task. *Hum. Factors.* **43**, 111–121. <https://doi.org/10.1518/001872001775992570> (2001).
27. Yeung, M. K. et al. Prefrontal activation and pupil dilation during n-back task performance: A combined fNIRS and pupillometry study. *Neuropsychologia* **159**, 107954. <https://doi.org/10.1016/j.neuropsychologia.2021.107954> (2021).
28. Lee, J. Y. et al. Measuring Cognitive Load in Virtual Reality Training via Pupillometry. *IEEE Trans. Learn. Technol.* **17**, 704–710. <https://doi.org/10.1109/TLT.2023.3326473> (2024).
29. Wei, J. 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, F. H. F. et al. Pain-related attentional processes: A systematic review of eye-tracking research. *Clin. Psychol. Rev.* **80**, 101884. <https://doi.org/10.1016/j.cpr.2020.101884> (2020).
31. Appel, T. et al. Cross-subject workload classification using pupil-related measures. *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. Warsaw, Poland: Association for Computing Machinery, p. Article 4. (2018).
32. Ktistakis, E. et al. COLET: A dataset for COgnitive workLOAD estimation based on eye-tracking. *Comput. Methods Programs Biomed.* **224**, 106989. <https://doi.org/10.1016/j.cmpb.2022.106989> (2022).

33. Fridman, L. et al. Cognitive Load Estimation in the Wild. *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Montreal QC, Canada: Association for Computing Machinery, 2018, p. Paper 652. Montreal QC, Canada: Association for Computing Machinery, 2018, p. Paper 652. (2018).
34. Fuhl, W., Herrmann-Werner, A. & Nieselt, K. A temporally quantized distribution of pupil diameters as a new feature for cognitive load classification. *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*. Tübingen, Germany: Association for Computing Machinery, p. Article 33. (2023).
35. Middlehurst, M. et al. aeon: a Python Toolkit for Learning from Time Series. *J. Mach. Learn. Res.* **25**, 1–10 (2024).
36. Dempster, A., Petitjean, F. & Webb, G. I. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Min. Knowl. Disc.* **34**, 1454–1495. <https://doi.org/10.1007/s10618-020-00701-z> (2020).
37. Middlehurst, M., Large, J. & 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. pp.188–195. (2020) 2020.
38. Middlehurst, M. et al. HIVE-COTE 2.0: a new meta ensemble for time series classification. *Mach. Learn.* **110**, 3211–3243. <https://doi.org/10.1007/s10994-021-06057-9> (2021).
39. Johnson, E. L. et al. Task-evoked pupillometry provides a window into the development of short-term memory capacity. *Front. Psychol.* **5**, 218. <https://doi.org/10.3389/fpsyg.2014.00218> (2014).
40. van der Wel, P. & van Steenbergen, H. Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychon. Bull. Rev.* **25**, 2005–2015. <https://doi.org/10.3758/s13423-018-1432-y> (2018).
41. Bauer, R. et al. Pupillometry as a measure of cognitive load in mental rotation tasks with abstract and embodied figures. *Psychol. Res.* **86** (20210812), 1382–1396. <https://doi.org/10.1007/s00426-021-01568-5> (2022).
42. Beatty, J. & Lucero-Wagoner, B. The pupillary system. *Handbook of psychophysiology*. Cambridge University Press, xiii + 1039: 142–162. (2000).
43. 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, pp. 259–262. (2010).
44. Salari, M. et al. The effect of pupil size on data quality in head-mounted eye trackers. *Behav. Res. Methods.* **58**, 17. <http://doi.org/10.3758/s13428-025-02880-3> (2025).
45. Thomay, C. et al. Towards Cognitive Load-Based Decision Making in VR Training. In: *IEEE 2nd International Conference on Cognitive Aspects of Virtual Reality (CVR)* 26–27 Oct. 2023 2023, pp.000023–000028. 26–27 Oct. 2023 2023, pp.000023–000028. (2023).
46. Kim, J. et al. Change in Blink Rate in the Metaverse VR HMD and AR Glasses Environment. *Int. J. Environ. Res. Public Health.* **19**, 8551 (2022).
47. Stern, J. A., Walrath, L. C. & Goldstein, R. The endogenous eyeblink. *Psychophysiology* **21**, 22–33. <https://doi.org/10.1111/j.1469-8986.1984.tb02312.x> (1984).
48. Biondi, F. N. et al. Distracted worker: Using pupil size and blink rate to detect cognitive load during manufacturing tasks. *Appl. Ergon.* **106**, 103867. <https://doi.org/10.1016/j.apergo.2022.103867> (2023).
49. Magliacano, A. et al. Eye blink rate increases as a function of cognitive load during an auditory oddball paradigm. *Neurosci. Lett.* **736**, 135293. <https://doi.org/10.1016/j.neulet.2020.135293> (2020).
50. Okazaki, Y. & Yokoi, N. Time-course observation of tear film dynamics during VR headset use. *Sci. Rep.* **15**, 33082. <https://doi.org/10.1038/s41598-025-16634-w> (2025).
51. Chen, J., Wang, X. & Xu, Z. The Relationship Between Chronic Pain and Cognitive Impairment in the Elderly: A Review of Current Evidence. *J. Pain Res.* **16**, 2309–2319. <https://doi.org/10.2147/jpr.S416253> (2023).
52. Zhang, X. et al. Evidence for Cognitive Decline in Chronic Pain: A Systematic Review and Meta-Analysis. *Front. Neurosci.* **15**, 737874. <https://doi.org/10.3389/fnins.2021.737874> (2021).
53. Attridge, N. et al. The disruptive effects of pain on n-back task performance in a large general population sample. *Pain* **156**, 1885–1891. <https://doi.org/10.1097/j.pain.0000000000000245> (2015).
54. Rodriguez, J. J., Kuncheva, L. I. & Alonso, C. J. Rotation Forest: A New Classifier Ensemble Method. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 1619–1630. <https://doi.org/10.1109/TPAMI.2006.211> (2006).
55. Dempster, A., Schmidt, D. F. & Webb, G. I. quant: a minimalist interval method for time series classification. *Data Min. Knowl. Disc.* **38**, 2377–2402. <https://doi.org/10.1007/s10618-024-01036-9> (2024).
56. Dreesmann, N. J., Su, H. & Thompson, H. J. A Systematic Review of Virtual Reality Therapeutics for Acute Pain Management. *Pain Manag. Nurs.* **23**, 672–681. <https://doi.org/10.1016/j.pmn.2022.05.004> (2022).
57. Lier, E. J. et al. Effect modifiers of virtual reality in pain management: a systematic review and meta-regression analysis. *Pain* **164**, 1658–166520230322. <https://doi.org/10.1097/j.pain.0000000000002883> (2023).
58. Maddox, T. 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.* **9**, e1182. <https://doi.org/10.1097/pr9.0000000000001182> (2024).
59. Charles, R. L. & Nixon, J. Measuring mental workload using physiological measures: A systematic review. *Appl. Ergon.* **74**, 221–232. <https://doi.org/10.1016/j.apergo.2018.08.028> (2019).

60. Hogervorst, M. A., Brouwer, A. M. & van Erp, J. B. Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. *Front. Neurosci.* **8**, 322. <https://doi.org/10.3389/fnins.2014.00322> (2014).
61. Gwinnutt, J. M. 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. *Ann. Rheum. Dis.* **77**, 848–854. <https://doi.org/10.1136/annrheumdis-2017-212426> (2018).
62. Pico Interactive. *Pico Neo 3 Pro Eye virtual reality headset* (Pico Interactive, 2021).
63. Unity Technologies. *Unity Real-Time Development Platform, version 2021.3.45f1* (Unity Technologies, 2021).
64. Tobii, A. B., Tobii Ocumen, S. D. K. & Overview <https://developer.tobii.com/xr/ocumen/overview/> (2025). accessed 27/11/2025).
65. Byrne, E. M. et al. Evidence for separate backward recall and n-back working memory factors: a large-scale latent variable analysis. *Memory* **32**, 1182–1198. <https://doi.org/10.1080/09658211.2024.2393388> (2024).
66. Jaeggi, S. M. et al. The concurrent validity of the N-back task as a working memory measure. *Memory* **18**, 394–412. <https://doi.org/10.1080/09658211003702171> (2010).
67. International Telecommunication Union. *Parameter values for the HDTV standards for production and international programme exchange (ITU-R BT.709-6)* (International Telecommunication Union, 2015).
68. Mathôt, S. & Pupillometry Psychology, Physiology, and Function. *J. Cognition.* <https://doi.org/10.5334/joc.18> (2018).
69. Cella, D. et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *J. Clin. Epidemiol.* **63**, 1179–1194. <https://doi.org/10.1016/j.jclinepi.2010.04.011> (2010).
70. Sullivan, M. J. L., Bishop, S. R. & Pivik, J. The Pain Catastrophizing Scale: Development and validation. *Psychol. Assess.* **7**, 524–532. <https://doi.org/10.1037/1040-3590.7.4.524> (1995).
71. Haatveit, B. C. et al. The validity of d prime as a working memory index: Results from the Bergen n-back task. *J. Clin. Exp. Neuropsychol.* **32**, 871–880. <https://doi.org/10.1080/13803391003596421> (2010).
72. Kane, M. J. et al. Working memory, attention control, and the N-back task: a question of construct validity. *J. Exp. Psychol. Learn. Mem. Cogn.* **33**, 615–622. <https://doi.org/10.1037/0278-7393.33.3.615> (2007).
73. Hart, S. G. & Staveland, L. E. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In: (eds Hancock, P. A. & Meshkati, N.) *Advances in Psychology*. North-Holland, 139–183. (1988).
74. Paas, F. Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *J. Educ. Psychol.* **84**, 429–434 (1992).
75. Kim, H. K. et al. Virtual reality sickness questionnaire (VRSQ): Motion sickness measurement index in a virtual reality environment. *Appl. Ergon.* **69**, 66–73. <https://doi.org/10.1016/j.apergo.2017.12.016> (2018).
76. Middlehurst, M., Schäfer, P. & Bagnall, A. Bake off redux: a review and experimental evaluation of recent time series classification algorithms. *Data Min. Knowl. Disc.* **38**, 1958–2031. <https://doi.org/10.1007/s10618-024-01022-1> (2024).
77. Vabalas, A. et al. Machine learning algorithm validation with a limited sample size. *PLOS ONE*. **14**, e0224365. <https://doi.org/10.1371/journal.pone.0224365> (2019).
78. Anaconda Inc. *Anaconda Distribution, version 3* (Anaconda Inc., 2023).
79. Hoeks, B. & Levelt, W. J. M. Pupillary dilation as a measure of attention: a quantitative system analysis. *Behav. Res. Methods Instruments Computers.* **25**, 16–26. <https://doi.org/10.3758/BF03204445> (1993).
80. Nyström, M. et al. What is a blink? Classifying and characterizing blinks in eye openness signals. *Behav. Res. Methods.* **56**, 3280–3299. <https://doi.org/10.3758/s13428-023-02333-9> (2024).
81. Mathôt, S. et al. Safe and sensible preprocessing and baseline correction of pupil-size data. *Behav. Res. Methods.* **50**, 94–106. <https://doi.org/10.3758/s13428-017-1007-2> (2018).
82. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2025).
83. Python Software Foundation. *Python Language Reference, version 3.12*. Wilmington (Python Software Foundation, 2023).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Jordan Tsigarides^{1,3}  · Jennifer O. Bowler¹ · Jack Dainty¹ · Anthony Bagnall² · Alexander Macgregor¹ 

✉ Jordan Tsigarides
J.tsigarides@uea.ac.uk

¹ Norwich Medical School, Faculty of Medicine and Health Sciences, University of East Anglia, Norwich, UK

² Department of Computing Science, University of Southampton, Southampton, UK

³ Bob Champion Research and Education Building, University of East Anglia, Research Park, Rosalind Franklin Road, Norwich NR4 7UQ, UK