

Szczepaniak, Dominik (2026) *Shaping and evaluating cognitive training in virtual reality using artificial intelligence agents.* PhD thesis.

https://theses.gla.ac.uk/86336/

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses
https://theses.gla.ac.uk
research-enlighten@glasgow.ac.uk

Shaping and Evaluating
Cognitive Training in Virtual Reality Using Artificial Intelligence Agents

Dominik Szczepaniak

Doctor of Philosophy

The University of Glasgow



24th September 2026

Abstract

Cognitive training that targets sustained attention and working memory is useful in education, clinical rehabilitation and interactive human-technology systems, but its benefit depends on keeping task difficulty matched to the person. Virtual reality (VR) offers immersion and settings representative of everyday tasks, and allows difficulty to be changed in real time. Existing adaptive systems, however, often rely on simple performance heuristics, affect or estimate visual complexity rather than task-related cognitive load. Moreover, cognitive load is usually judged against the participant's own ratings, which are themselves an imperfect reference. The central hypothesis of this thesis is that eye-tracking and peripheral physiological signals recorded in VR can estimate how difficult a task feels to the person doing it, well enough for a model to adjust difficulty in real time with outcomes comparable to participant-driven control, and that the estimation part holds for older adults once age-related differences in the signal are taken into account.

We developed a bespoke VR sustained attention task and collected eye-tracking, photoplethysmography and galvanic skin response data from 54 younger (aged 18-58) and 19 older (aged 60+) adults at fixed difficulty levels. Models evaluated on unseen participants performed above chance (younger adults: accuracy 0.56, AUC 0.72), with eye-tracking carrying most of the signal and physiological features alone close to chance. Performance was lower in older adults (accuracy 0.49, AUC 0.66), and the most informative features appeared to differ between age groups. A bidirectional LSTM with self-attention reached a slightly higher macro-F1 than gradient-boosted trees on the same data (0.50 against 0.47), and performed better than a model given raw signals only.

We then deployed the LSTM in real time with 54 younger adults (51 analysed), comparing model-driven and participant-driven difficulty in single- and dual-task conditions. Per-trial agreement with participants' ratings was low (accuracy 36.7%, AUC 0.63), yet the two forms of control appeared to produce comparable outcomes. There were no significant differences in difficulty-weighted game performance, arithmetic errors or subjective workload, including frustration. Final difficulty was equivalent in single-task conditions

and slightly higher under model control in dual-task conditions, a difference that was borderline after correction. Just Right ratings were more frequent in the real-time cohort than in the offline cohort (57% and 39%).

Finally, we compared subjective ratings with performance-derived reference measures. Agreement between the two appeared fair and slightly lower in real time, where the steps between difficulty levels are less discrete. The same features seemed to predict performance-derived labels more accurately than subjective ratings in every matched comparison. This suggests that self-report may be a less stable reference for real-time adaptation.

Within a single session with younger adults, model-driven adaptation appears to have produced outcomes comparable to participant-driven control. The dataset (EyeDAPT) and the experimental software are publicly available.

Contents

Abstract **ii**

Acknowledgements **xxi**

Publications **xxii**

Declaration **xxiii**

Abbreviations **xxiv**

1 Introduction **1**

1.1 Motivation 1

1.2 Problem Statement 3

1.3 Thesis Statement 4

1.4 Research Questions and Hypotheses 6

1.5 Thesis Overview 8

2 Background and Related Work **11**

2.1 The Neuropsychological Foundation. Defining Cognition, Cognitive Decline, and Cognitive Training 11

2.1.1 Normal Cognitive Ageing and the Definition of Cognitive Decline . 12

2.1.2 Cognitive Training: Purpose, Scope, and Limitations 13

2.1.3 Virtual Reality as an Immersive Training Platform 13

2.1.4 Experimental Designs in Cognitive Training: Single vs. Dual-Task Paradigms 14

2.1.5 Sustained Attention: A Critical Component of Cognitive Function . 15

2.1.6 Understanding Cognitive Load Theory 16

2.1.7 Subjective Workload Assessment: The NASA Task Load Index . . 18
2.1.8 Reference Measures for Cognitive Load: Benefits and Drawbacks . 20
2.2 Educational and Motivational Frameworks for Adaptive Intervention . . . 21
2.2.1 The Zone of Proximal Development 21
2.2.2 The Flow State: Achieving Optimal Experience 22
2.2.3 Flow and the Zone of Proximal Development for Adaptive Designs 23
2.3 Real-Time Measurement and Physiological Feedback 24
2.3.1 Electroencephalography 24
2.3.2 Eye-Tracking and Gaze Analysis 26
2.3.3 Galvanic Skin Response 27
2.3.4 Heart Rate Variability and Associated Metrics 27
2.3.5 Multimodal Fusion and Dataset Benchmarks 28
2.4 Advanced Predictive Algorithms: Traditional Machine Learning and Deep Learning 30
2.4.1 Feature Engineering versus Feature Learning 31
2.4.2 Deep Learning and Representation Learning 31
2.4.3 Domain Specialisations: Matching the Architecture to the Data . . 32
2.4.4 The Optimisation Landscape and Training Strategies 33
2.4.5 Regularisation: Preventing Memorisation 34
2.4.6 Scaling Laws and Data Demands 34
2.5 Generalisation, Validation and Dataset Shift 35
2.6 Older Adults: Learning, Technology Acceptance, and the Suitability of Physiological Assessment 36
2.6.1 Virtual Reality for Cognitive Training in Older Adults 36
2.6.2 Physiological Baseline Differences with Age 37
2.6.3 Self-Assessment Reliability in Older Adults 38
2.6.4 Technology Acceptance and Usability in Older Adults 39
2.7 Brain-Computer Interfaces and Adaptive Environments 40
2.7.1 Adaptive Environments and Human-in-the-Loop Architecture . . . 41
2.7.2 Failure Modes in Physiologically-Driven Adaptive Systems 43
2.8 Designing Closed-Loop Adaptive Cognitive Training Environments 44

2.9 Summary: Gaps, Contributions, and Thesis Structure 45

3 Development of a Novel Physiologically-Driven Cognitive Load Model in VR 47

3.1 Introduction 47

3.2 Key contributions 48

3.3 Methodology 48

3.3.1 Experimental Design Overview 48

3.3.2 Participants 50

3.3.3 Task 50

3.3.4 Measure of Cognitive Workload 52

3.3.5 Materials 54

3.3.6 Eye-tracking and Physiological Signal Processing 55

3.3.7 Machine Learning Models for Cognitive Workload Prediction 61

3.3.8 Explainability of the Machine Learning Models 62

3.4 Results 62

3.4.1 Subjective Difficulty Distribution 63

3.4.2 Classification Performance 63

3.4.3 Physical versus Mental Demand 68

3.5 Discussion 68

3.5.1 The Dominance of Oculometric Features in VR 69

3.5.2 The Generalization Gap in Physiological Sensing 69

3.5.3 Methodological Implications of Validation Protocols 70

3.6 Summary 71

3.6.1 Inclusion of an Older Participant Cohort 72

3.6.2 Real-Time Adaptation (Closed-Loop VR) 72

3.6.3 Deep Learning Approaches 73

4 Bridging the Age Gap in Cognitive Workload Modeling 74

4.1 Introduction 74

4.1.1 Rationale for Including Older Adult Participants 75

4.1.2 Chapter Aim and Contributions 75

4.2 Methodology . . . 76
4.2.1 Experimental Design Overview . . . 76
4.2.2 Participants . . . 78
4.2.3 Task design . . . 80
4.2.4 Data Analysis Strategy . . . 80
4.3 Results . . . 82
4.3.1 Participant Game Performance Comparison . . . 82
4.3.2 Subjective Workload (NASA-TLX) . . . 83
4.3.3 Physiological Differences . . . 84
4.3.4 Subjective Perception of Difficulty . . . 86
4.3.5 Predictive Modelling Results . . . 88
4.4 Discussion . . . 92
4.4.1 Performance Limits and Neural Compensation . . . 92
4.4.2 Eye-tracking Differences: Attentional vs. Motoric Bottlenecks . . . 93
4.4.3 The Value of Mixed-Population Models for Inclusivity . . . 93
4.4.4 Implications of Sample Representativeness . . . 94
4.4.5 Implications for VR Design: From Cognitive to Motoric Scaffolding 95
4.5 Summary . . . 95

5 Development and Pre-Deployment Evaluation of a Real-Time Cognitive Load Estimation Model 97

5.1 Introduction . . . 97
5.1.1 Chapter Aim and Contributions . . . 100
5.2 Methodology . . . 101
5.2.1 Experimental Design Overview . . . 101
5.2.2 Training Data . . . 102
5.2.3 Feature Engineering . . . 103
5.2.4 Real-time Multi-modal Cognitive Load Estimation Model . . . 104
5.2.5 Model Training Configuration . . . 107
5.2.6 Cross-Validation and Generalization Assessment . . . 108
5.3 Results . . . 110

5.3.1 Model Evaluation (pre-deployment) . . . 110
5.3.2 Raw-Signal Baseline and Comparison with XGBoost . . . 112
5.3.3 Choice of the Deployed Model . . . 114
5.4 Summary . . . 115

6 Real-Time Deployment of an Adaptive Cognitive Training System in Virtual Reality 117

6.1 Introduction . . . 117
6.1.1 Rationale and Contributions . . . 118
6.2 Methodology . . . 118
6.2.1 Experimental Design Overview . . . 118
6.2.2 Participants . . . 120
6.2.3 Experimental Design - Framework for Real-time Model Adaptation 120
6.2.4 Materials . . . 124
6.2.5 Real-time Deployment for VR integration . . . 124
6.3 Results . . . 125
6.3.1 Evaluation of Physiologically-Informed Adaptive Difficulty Prediction in Virtual Reality (Analysis of the Classifier Performance During Deployment) . . . 126
6.3.2 Subjective Difficulty Analysis . . . 128
6.3.3 Game Performance Analysis: Influences of Control Mechanism and Task Complexity . . . 129
6.3.4 Adaptive Control Effectiveness: Difficulty Trends and Convergence 133
6.3.5 Divergence in Cognitive Load Judgments: Human vs. Model Thresholds 136
6.3.6 Subjective Workload Assessment with NASA TLX . . . 138
6.3.7 Learning and Fatigue Effects . . . 142
6.3.8 Individual Differences in Real-time Model Performance . . . 145
6.4 Discussion . . . 148
6.4.1 Methodological Contributions . . . 148
6.4.2 Key Findings . . . 149
6.4.3 Limitations and Future Directions . . . 153

6.5 Summary . . . 154

7 Subjective and Objective References for Cognitive Load in Adaptive Cognitive Training 155

7.1 Introduction . . . 155
7.1.1 The Physiology of Mental Effort . . . 157
7.1.2 Psychological Mechanisms of Perceptual Judgment and Anchoring . 158
7.1.3 Computational Challenges: Covariate and Domain Shift . . . 159
7.1.4 Practical Limitations of Performance-Driven Ground Truths . . . 160
7.1.5 Chapter Aim and Contributions . . . 161
7.2 Methodology . . . 162
7.2.1 Experimental Design Overview . . . 162
7.2.2 Participants, Task, Setup, Preprocessing . . . 163
7.2.3 Target Variable Formulation . . . 163
7.3 Analyses and Results . . . 164
7.3.1 Analysis 1: Target Distribution and Subjective-Objective Alignment 164
7.3.2 Analysis 2: Predictive Modelling of Objective Cognitive Load . . . 168
7.3.3 Analysis 3: Sample Size Sufficiency and Learning Curves . . . 176
7.3.4 Analysis 4: Visualising the Data through t-SNE Plotting . . . 180
7.4 Discussion . . . 184
7.5 Summary . . . 184

8 Discussion and Limitations 186

8.1 Introduction . . . 186
8.2 Outcome of the Hypotheses . . . 187
8.3 Real-Time Cognitive Load Estimation from Eye-Tracking and Physiology (RQ1) . . . 189
8.4 Extending the Model to Older Adults (RQ2, RQ3) . . . 191
8.5 Real-Time Deployment: Model- versus Participant-Driven Control (RQ4, RQ5, RQ6) . . . 192
8.6 Subjective and Objective References and the Choice of Ground Truth (RQ7) 194
8.7 Implications for Design Practice . . . 195

9 Conclusion and Future Work **198**

9.1 Thesis Summary 198

9.2 Contributions 200

9.3 Future Work 201

9.4 Concluding Remarks 204

10 Supplementary Material **205**

10.1 NASA-TLX Descriptive Statistics 205

10.1.1 Offline Experiment 205

10.1.2 Real-Time Experiment 206

List of Tables

3.1 Distribution of perceived difficulty ratings across all sessions. Counts come from the 58 participants with usable session data (1,384 ratings); the classification analyses use the 54 participants with complete physiological data (1,280 sub-sessions). 63

3.2 Pooled Cross-Validation Results (PCVP). This protocol, common in the literature, generally yields higher accuracies due to participant data leakage. . . . 65

3.3 Leave-One-Participant-Out (LOPO) Validation Results (Mean (SD)). This protocol tests generalizability to unseen users. Note that even with the full set of features, Physiological models struggle to generalize compared to Eye-Tracking. 67

4.1 Summary of the Age Comparison Results (Session Level). Significant differences were observed across performance, frustration, and physiological metrics, despite similarly reported mental and physical demand. 84

4.2 Age Classification Results (Physiological Features Only). The drop in performance for Within-Participant Z-Scoring suggests that age-related differences were driven by absolute baseline values rather than relative task reactivity. . . 86

4.3 Summary of XGBoost Classification Performance (LOPO Cross-Validation). Values represent Mean ± Standard Deviation across all participants. The Older Adult cohort shows reduced classification accuracy and AUC compared to the Younger and Mixed populations. 90

5.1 Overall Metrics Averaged Across LOPO-CV Folds (pre-deployment) 111

5.2 Models trained on the 73-participant offline dataset (1,736 task levels). Accuracy and macro-F1 for the two feature-based models are means across held-out participants; the raw-signal results are out-of-fold. 114

6.1 Overall Metrics Averaged Across All Real-Time Predictions (During Deployment) . . . 127
6.2 Cognitive Load Judgment Agreement Analysis for Raw Game Performance and Difficulty-Weighted Game Performance by Task Type and Perceived Difficulty Category - . **Bold values indicate statistical significance.** *** p < 0.001, ** p < 0.01, * p < 0.05. Positive t-statistics indicate higher self-control game performance relative to model-control. Cohen's d effect sizes: small (0.2), medium (0.5), large (0.8). . . . 136
6.3 Repeated-Measures ANOVA Results for Single-Task Conditions . . . 142
6.4 Repeated-Measures ANOVA Results for Dual-Task Conditions . . . 142

7.1 The 15 XGBoost analyses performed to examine the difference in predictive performance between weighted performance and subjective judgments. Training-cohort participant counts are given for reference; row-level sample sizes (number of sub-session/level observations) vary by cohort in the same proportions as Chapters 3-6 and are not restated per row here. . . . 169
7.2 Machine Learning Performance Metrics across 15 Experimental Configurations 174

8.1 Outcome of each hypothesis, with the chapter that tested it. . . . 188

10.1 NASA-TLX descriptive statistics, offline experiment (mean (SD); median in brackets) . . . 206
10.2 NASA-TLX descriptive statistics, real-time experiment, overall ($N = 216$) . . . 206
10.3 NASA-TLX Total and Mental Demand by condition, real-time experiment (mean (SD), $n = 54$ per condition) . . . 207

List of Figures

1.1 Graphical overview of the thesis structure, showing how each experimental chapter addresses a distinct subset of the research questions introduced in this chapter. 9

3.1 Example views of the experiment progressing in difficulty from an easier subsession on the left (only one cube of each colour present at a time) to a more difficult one on the right (several cubes, some moving). 52

3.2 Macro-F1 as a function of the number of SHAP-ranked features retained, for the combined younger and older dataset ($N = 73$), with fixed hyperparameters and participant-grouped 5-fold cross-validation repeated ten times. The line shows the mean across repeats, the band ± 1 SD and the points the individual repeats. Performance peaks at $N = 30$ (macro-F1 $= 0.519$) and remains within 0.02 of this peak from $N = 20$ to $N = 85$ (shaded). 60

3.3 Performance and interpretability analysis of the XGBoost model (Combined Feature Set) under the Within-Participant (Pooled Cross-Validation) protocol for the younger population. (a) **Confusion Matrix:** The model shows the highest sensitivity for the *Too Easy* class (70.06%), with *Just Right* (53.01%) and *Too Difficult* (47.72%) more frequently confused with one another. (b) **Multiclass ROC Curve:** The Macro-average AUC of 0.75 indicates above-chance classification performance, with the *Too Easy* class the most discriminable (AUC = 0.80). (c) **SHAP Feature Importance:** Oculometric features dominate the ranking, led by Fixation Duration SD, Blink Count and Saccade Velocity SD, while physiological features such as GSR Slope contribute comparatively little. 64

3.4 Performance and interpretability analysis of the XGBoost model (Combined Feature Set) under the Leave-One-Participant-Out (LOPO) protocol for the younger population. (a) **Confusion Matrix:** The model demonstrates high sensitivity for the *Too Easy* class (67.28%) but shows confusion between *Just Right* and *Too Difficult*, reflecting the subjective overlap between these states. (b) **Multiclass ROC Curve:** The Macro-average AUC of 0.72 indicates classification performance above chance level, with the 'Too Easy' class achieving the highest discriminative power (AUC = 0.77). (c) **SHAP Feature Importance:** Eye-tracking features, particularly Fixation Duration Standard Deviation (SD) and Mean Fixation Duration, emerge as the top predictors. Physiological features like Heart Rate (BPM) and GSR Slope appear lower in the ranking, suggesting a secondary role in this population. 66

4.1 Performance differences between age groups. Left: Older adults show significantly lower Weighted Performance scores ($p < 0.001, d = 1.51$). Right: Comparison of Raw Task Performance across difficulty levels (Spawn Rates 3.1 to 1.6), showing a steeper decline in older adults as task difficulty increases. 83

4.2 Comparison of the Age Prediction models. The Global model (left) effectively distinguishes age groups, whereas removing baseline information (Within-Participant Z, right) eliminates predictive capability. 86

4.3 Distribution of the Subjective Difficulty Ratings by Age Group. A significant distributional shift ($p < 0.001$) is observed, with older adults significantly shifting towards *Just Right* and *Too Difficult* ratings while younger adults predominantly selected *Too Easy*. 87

4.4 Performance comparison between Mixed Population (Top) and Older Adults (Bottom). Panels (a) and (c) show confusion matrices, highlighting the higher misclassification rate in the older cohort, particularly between *Too Easy* and *Just Right*. Panels (b) and (d) display Multiclass ROC curves, revealing a drop in Macro-Average AUC from 0.72 to 0.64 in the older population. 89

4.5 Feature Importance Comparison (SHAP Summary Plots). Left: Younger Adults. Right: Older Adults. The Younger model is dominated by Fixation Stability (Fixation Dur. SD), whereas the Older Adult model relies heavily on Oculomotor Mechanics (Saccade Amplitude and Acceleration/Deceleration Ratios). 91

5.1 Diagram of the LSTM classifier architecture. 104
5.2 Confusion matrix, ROC curves and feature importance rating for the offline (pre-deployment) model. 112
5.3 Confusion matrix of the balanced raw data model (F1 = 0.4) 113

6.1 In-game view at different difficulty levels including mental arithmetic menu (middle right) and perceived difficulty selection menu (top right). 121
6.2 The flowchart depicting the experimental flow of the adaptive experimental design compared to the original flow used in Chapters 3 and 4. 123
6.3 Complete model evaluation across all experimental conditions. (a) Overall confusion matrix showing prediction accuracy with percentages and counts, (b) Model performance by condition showing ROC AUC scores for each difficulty class, (c) ROC curves for overall performance, (d) ROC AUC comparison by control type, (e) ROC AUC comparison by task type, (f) Macro-average ROC AUC by individual condition. 126
6.4 Distribution of Subjective Difficulty Ratings. The chart compares the original dataset (baseline) with the current study's Single and Dual Task conditions. . 129

6.5 Integrated performance and error analysis across experimental conditions based on 2×2 repeated measures ANOVA. (a) Raw game performance scores by condition showing significant Control × Task Complexity interaction ($p = 0.008$) with equivalent model-self game performance in single tasks but self-control advantage in dual tasks; (b) Difficulty-weighted game performance revealing no significant effect of control type ($p = 0.219$) and borderline effects of task complexity and of the Control × Task Complexity interaction (both $p = 0.050$); (c) Condition-specific performance comparisons highlighting the interaction pattern, showing how model-self differences vary by task complexity; (d) Raw game performance interaction plot demonstrating significant crossing pattern ($p = 0.008$) where control type effectiveness depends on task complexity; (e) Weighted game performance interaction plot (Control × Task Complexity interaction borderline, $p = 0.050$); (f) Arithmetic error rates in dual-task conditions showing numerically higher errors for model-control (18.6%) versus self-control (15.0%, $p = 0.168$, ns). Statistical analyses used repeated measures ANOVA for factorial comparisons and paired t-tests for specific contrasts. Box plots show median, quartiles, and outliers; interaction plots display condition means with statistical annotations. 130

6.6 Final difficulty analysis revealing control type effectiveness and task complexity interactions. (a) Control type × task type interaction plot demonstrating equivalent performance in single-task conditions but divergence in dual-task conditions, with model control achieving harder final difficulties. Statistical comparisons show non-significant differences in single-task conditions (p=0.714) but marginally significant differences in dual-task conditions (p=0.025, corrected p=0.05), with model-based control reaching harder final difficulties than self-assessment control when cognitive resources were divided between VR task performance and arithmetic processing; (b) Final difficulty distributions across all four experimental conditions, showing the magnitude of differences between control types within each task complexity level; (c) Difficulty trajectories over level sequence in single-task conditions and (d) in dual-task conditions, with linear fits for model-driven (red) and participant-driven (blue) control. Lower difficulty values indicate harder task settings. Box plots show median, quartiles, and outliers; error bars in the interaction plot represent standard error of the mean. 134

6.7 Game performance analysis showing the relationship between difficulty judgments/predictions and actual game performance. The 2×3 grid compares raw performance scores (top row) and weighted game performance scores (bottom row) for both model and self-driven conditions, including an additional breakdown for single (left column) and dual tasks (middle column) as well as both conditions pooled (right column). Boxplots display the distribution of mean performance scores aggregated by participant and difficulty (*Too Easy, Just Right and Too Difficult*). This visualization presents the mapping of raw and weighted game performance across difficulty estimations. 137

6.8 NASA TLX workload analysis revealing dual-task cognitive burden and performance relationships. (a) Dual-task workload increases showing significant elevations in mental demand ($p < 0.001$, $d = 1.41$), effort ($p < 0.001$, $d = 0.78$), and frustration ($p < 0.001$, $d = 0.69$) when comparing dual-task to single-task conditions within participants; (b) TLX-performance correlation demonstrating a significant negative relationship ($r = -0.323$, $p < 0.001$) between subjective workload and spawn-destroy task performance across all conditions; (c) Total workload distribution confirming significantly higher TLX scores in dual-task compared to single-task conditions ($p < 0.001$, $d = 0.87$); (d) TLX subscale profiles for mental demand, physical demand, and temporal demand across the four experimental conditions, with Model Dual-task showing the highest mental demand ratings; (e) TLX subscale profiles for effort, frustration, and total workload, demonstrating consistent patterns where dual-task conditions (Model Dual-task and Self Dual-task) exceed single-task conditions across all measures; (f) Effect size analysis quantifying the magnitude of dual-task workload increases, with mental demand showing the largest effect ($d = 1.41$), followed by total workload ($d = 0.87$), effort ($d = 0.78$), and frustration ($d = 0.69$), while physical and temporal demands showed smaller but significant effects ($d = 0.46$ and $d = 0.36$, respectively). Box plots display median, quartiles, and outliers; bar charts show condition means with individual values labeled; effect size reference lines indicate Cohen's conventions for small (0.2), medium (0.5), and large (0.8) effects. 141

6.9 Linear regression of Weighted Performance trajectories over time. (a) Single Task condition (left) and (b) Dual Task condition (right). Red lines represent Model Control; Blue lines represent Self Control. The significant positive slope across all conditions ($p < .001$) indicates a consistent learning effect. Shaded regions represent 95% confidence intervals. 144

6.10 Significant learning and fatigue effects across experimental conditions. The figure displays five key findings: (a) Weighted game performance improvements in both single-task (Δ=0.224, $p<0.001$) and dual-task conditions (Δ=0.146, $p<0.001$); (b) Raw game performance improvement in dual-task conditions only (Δ=0.024, p=0.016); (c) Final difficulty level reduction in single-task conditions only (Δ=-0.162, $p<0.001$), indicating progression to harder tasks; (d) Overall workload reduction in dual-task conditions (Δ=-8.5, $p<0.001$); and (e) Specific TLX subscale reductions, with single-task conditions showing improvements in mental demand (Δ=-1.5, p=0.024), performance perception (Δ=-1.2, p=0.049), and effort (Δ=-2.0, p=0.007), while dual-task conditions demonstrated broader improvements across mental demand (Δ=-2.4, $p<0.001$), physical demand (Δ=-1.0, p=0.026), performance perception (Δ=-1.2, p=0.007), and effort (Δ=-1.9, p=0.003) subscales (all $p<0.05$). Box plots show median, quartiles, and outliers; bar charts display mean differences with statistical significance levels indicated. 146

7.1 Distribution of discretized objective cognitive load classes across the three experimental cohorts. The normalization process successfully maintained comparable class balances regardless of the specific task paradigm. 166

7.2 The Subjective-Objective Mismatch. Stacked bar charts illustrating the true objective performance state of participants conditional upon their self-reported cognitive load. 166

7.3 Performance baselines for the Real-time and Offline predictive models. The top row displays the One-vs-Rest Receiver Operating Characteristic (ROC) curves, highlighting moderate discrimination for extreme cognitive states. The bottom row displays the normalized confusion matrices, indicating that misclassifications primarily occur between adjacent classes. 172

7.4 SHAP (SHapley Additive exPlanations) summary plots for the top 10 predictive features of both baseline models. Despite the differences in VR environments, both models overwhelmingly relied on identical ocular metrics, specifically pupil variance and fixation duration variance, to make their predictions. 172

7.5 A 3x3 matrix of predictive model learning curves demonstrating sample size sufficiency across all experimental contexts. The x-axis represents the number of participants available to the algorithm, while the y-axis shows the resulting LOOCV Global Macro AUC (with shaded regions representing ± 1 standard deviation). Across pure, combined, and age-isolated cohorts, all models exhibit a clear performance plateau, suggesting diminishing returns from further participant recruitment. 177

7.6 t-SNE embeddings of the 112 level-level features (28 signals $\times$ 4 statistics), one embedding per cohort (rows). The coordinates in the left and right columns are identical. The left column is coloured by the participant's subjective rating (*Too Easy*, *Just Right* and *Too Difficult* shown as Low, Medium and High Load) and the right column by the objective label derived from difficulty-weighted task performance. 182

Acknowledgements

I would like to begin by thanking my outstanding supervisory team - Dr Fani Deligianni and Prof Monika Harvey. I hope I have learned as much as I could have from you both, yet I know I barely scratched the surface. My thanks extend to the entire Social AI CDT team who made this journey possible in ways I am not even fully aware of. This work would not have been possible if it wasn't for the 130+ participants who volunteered their time and provided feedback on the experience and the design of the experiments - thank you all.

I would like to thank my parents, Aneta and Waldemar, for always being supportive and encouraging even if they are not always entirely sure of what I am up to. Thank you for instilling in me the curiosity, discipline and patience I needed to complete this thesis. To my grandma Renia, thank you for all the care packages and warm welcomes but most importantly your endless support and belief in me.

I dedicate this thesis to my grandpa Walenty, who, from a very young age, taught me the value of patient scientific pursuit and inspired me with a relentless work ethic combined with ceaseless awe at the machinations of the world. I am grateful we got to discuss this thesis in its early stages and that you saw its potential even then.

This PhD journey would mean much less if not for the great friendships I made along the way. To Cristina, Nicole and Michal, thank you for being equally supportive during the good as well as bad times.

Finally to Juliane, the person I am the most grateful for. Thank you for every single one of the many moments we shared together. Being by your side is the most exhilarating adventure, and I am so relieved to be able to close this first chapter while looking forward to the future full of excitement.

Publications

Earlier versions of this work, relating in particular to Chapter 3, were published at:
ACM ETRA: Predictive modelling of cognitive workload in VR: An eye-tracking approach. In Proceedings of the 2024 Symposium on Eye Tracking Research and Applications, D. Szczepaniak, M. Harvey and F. Deligianni, Association for Computing Machinery, Art 46, pp. 1-3, 2024 (2024) https://dl.acm.org/doi/10.1145/3649902.3655642
IEEE ISMAR: ML-driven cognitive workload estimation in a VR-based sustained attention task, D. Szczepaniak, M. Harvey and F. Deligianni, IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), pp. 557-560, 2024 (2024) https://ieeexplore.ieee.org/document/10765330

A paper related to Chapter 6 is pending a journal review and is currently available on arXiv:
D. Szczepaniak, M. Harvey, and F. Deligianni, "Your Eyes Controlled the Game: Real-Time Cognitive Training Adaptation based on Eye-Tracking and Physiological Data in Virtual Reality," arXiv preprint arXiv:2512.17882, 2026. https://arxiv.org/abs/2512.17882

The full dataset has been published on Open Science Framework:
Szczepaniak, D., Harvey, M., and Deligianni, F. (2026, March 16). The EyeDAPT Dataset: Multimodal Eye-Tracking and Physiology for Real-Time Cognitive Load Adaptation in VR. https://doi.org/10.17605/OSF.IO/VJSNF

The full experimental software as well as relevant code was made publicly available on Github: https://github.com/EyeDAPT

Declaration

I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution

Dominik Szczepaniak

Abbreviations

- `VR` - Virtual Reality
- `HMD` - Head-Mounted Display
- `ML` - Machine Learning
- `ET` - Eye-tracking
- `PPG` - Photoplethysmography
- `GSR` - Galvanic Skin Response
- `CT` - Cognitive Training
- `CL` - Cognitive Load
- `MCI` - Mild Cognitive Impairment
- `ZPD` - Zone of Proximal Development
- `MKO` - More Knowledgeable Other
- `EEG` - Electroencephalography
- `EDA` - Electrodermal Activity
- `HRV` - Heart Rate Variability
- `SDNN` - Standard Deviation of Normal-to-Normal intervals
- `RMSSD` - Root Mean Square of Successive Differences
- `pNN50` - Percentage of successive NN intervals that differ by more than 50 milliseconds

- `CNN` - Convolutional Neural Networks
- `RNN` - Recurrent Neural Networks
- `LSTM` - Long Short Term Memory
- `SGD` - Stochastic Gradient Descent
- `CV` - Cross-validation
- `LOSO` - Leave-One-Subject-Out
- `LOPO` - Leave-One-Participant-Out
- `PCVP` - Pooled Cross-Validation Protocol
- `SHAP` - SHapley Additive exPlanations
- `AUC` - Area Under the ROC Curve
- `UFOV` - Useful Field of View
- `CRUNCH` - Compensation-Related Utilisation of Neural Circuits Hypothesis
- `BCI` - Brain-Computer Interface
- `HITL` - Human In the Loop
- `CCT` - Computerized Cognitive Training
- `DDA` - Dynamic Difficulty Adjustment
- `PD` - Pupil Dilation
- `LC-NE` - Locus Coeruleus-Norepinephrine

Chapter 1

Introduction

1.1 Motivation

Some cognitive decline is common in healthy ageing [124]. It can affect a wide range of mental functions, including memory, executive function and sustained attention. Its effects extend beyond the laboratory. Momentary lapses in attention are associated with everyday errors and accidents. Studies using self-report questionnaires, such as the Cognitive Failures Questionnaire, have linked these lapses to errors and accidents at work [141]. The proportion of older adults in the population continues to grow, and with it the prevalence of age-related cognitive decline. The practical stakes of maintaining sustained attention and related executive functions across the lifespan are therefore increasing rather than diminishing [32]. Cognitive training is intended to slow these declines. Structured, repeated practice targeting specific cognitive domains has been shown in some studies to produce improvements that persist for years after the intervention has ended [182, 137]. Whether such training is effective, however, may depend, at least in part, on how well the difficulty of the task is matched to the individual's evolving capacity over time, rather than on repetition alone [52]. It is this adaptive, personalised premise, that the rest of this work is built around.

The effectiveness of cognitive training is not settled. Differences in methodology, population, and outcome measures across studies have left the field with a contested evidence base. Calls for a more mechanistic account of when and why training transfers to real-world benefit have become more prominent [153, 100]. One of the central levers available to a training system, and the one this thesis is concerned with, is cognitive load. Keeping a participant's cognitive load within an appropriate range, neither so low that the task becomes tedious nor so high that it becomes overwhelming, is widely regarded as important for both engagement and learning [161, 43]. Adaptive systems that manage cognitive load automatically and in real time, rather than relying on fixed difficulty schedules, can therefore be seen as an extension of cognitive training rather than a separate technology. Virtual reality offers a suitable delivery platform for this kind of adaptive training, since it combines naturalistic settings representative of everyday tasks with a degree of experimental control that is difficult to achieve outside a laboratory [183, 187], a point developed further in Chapter 2.

Building an adaptive system of this kind requires the ability to measure cognitive load while the user is still performing the task, not after the fact. This is where the practical difficulty begins. Electroencephalography (EEG) records neural activity directly and, although it remains an approximation of cognitive load, would in principle be a strong candidate proxy for it in this thesis. In practice, however, EEG is difficult to use in interactive, head-mounted VR. Head movement and natural interaction, which are central to an immersive VR task, introduce motion and electromyographic artifacts that degrade signal quality. The electromagnetic noise generated by head-mounted display (HMD) hardware, together with the combined weight and long setup time, further limit its practicality outside tightly controlled laboratory conditions [165, 64, 118, 95]. This thesis therefore does not use EEG, the modality that would otherwise be preferred, because it is incompatible with the interactive, mobile and potentially at-home context this work targets. Eye-tracking and peripheral physiological sensing, discussed in detail in Chapters 2 and 3, are adopted instead as a more practically viable, if individually less direct, alternative.

This substitution raises a deeper problem than simply choosing between available sensors. Cognitive load has no single, directly observable ground truth. Subjective self-report, most commonly operationalised through instruments such as the NASA Task Load Index, is the measure most often treated as capturing the construct of interest, perceived mental effort, directly. Yet subjective report is typically collected retrospectively. It can be shaped by recall bias and by the most demanding or most recent moments of a task, rather than by a stable average across it [131]. It is therefore difficult to reconcile with a system that must act while the task is still under way [66]. Physiological and behavioural measures avoid this timing problem but introduce a different one, since they are indirect proxies validated against subjective report or against performance, neither of which is itself an uncontested ground truth. This tension between objective and subjective assessment, and between what each can offer a real-time system, is re-examined throughout the thesis. It returns whenever the population or the deployment conditions change enough to call the assumed relationship between these measures into question. Perhaps the clearest example will be examined in Chapter 4 when testing on older adults, whose physiological baselines and self-reporting behaviour can diverge from those of younger populations [51].

This thesis focuses on healthy adults across the lifespan rather than on clinical populations. Conditions such as Mild Cognitive Impairment or dementia raise their own methodological and ethical considerations that fall outside the present scope. The underlying measurement problem may be even more acute in those groups, and addressing it there is left to future work rather than treated here.

1.2 Problem Statement

Despite a substantial literature on estimating cognitive load from eye-tracking and physiological data, several methodological gaps limit how far current findings can be trusted or deployed. First, cognitive load is commonly manipulated through task-stacking, in which additional concurrent tasks are layered onto a base task to represent increasing load. This risks training models to recognise the number or type of tasks being performed

rather than a genuine modulation of cognitive load within a single, consistent task [178, 25]. Second, evidence that models generalise is limited. Model evaluation frequently relies on pooled-subject cross-validation. This can allow models to learn participant-specific physiological signatures rather than generalisable markers of cognitive state, which may inflate reported accuracy relative to what a genuinely new user would experience [25, 113]. It is also less common to examine how models behave when the conditions of use differ from those under which they were trained (dataset shift) [159]. Third, many of the existing systems reviewed in Chapter 2, including those that estimate cognitive load with reasonable accuracy offline, have not been deployed as real-time, closed-loop adaptive systems, meaning their practical utility for dynamic difficulty adjustment remains largely untested [93]. Fourth, the suitability of these approaches for older adults, who are a primary intended beneficiary of adaptive cognitive training, has rarely been established directly, given that many physiological computing models are developed and validated on younger cohorts [152]. Finally, even where subjective workload report is used as ground truth for training or validating these models, its own suitability as a stable reference for a real-time adaptive system has been comparatively under-examined, despite a growing methodological literature questioning this convention [14].

1.3 Thesis Statement

The central hypothesis of this thesis is that eye-tracking and peripheral physiological signals recorded in virtual reality can estimate how difficult a task feels to the person doing it, well enough for a model to adjust difficulty in real time with outcomes comparable to those obtained when participants adjust difficulty themselves. The estimation part is expected to hold for older adults once age-related differences in the signal are taken into account.

The relevance of this hypothesis follows from how cognitive training works and from the state of cognitive load estimation. Cognitive training aims to improve working memory and executive functions such as attention or multitasking through targeted exercises and activities [153], although its efficacy remains contested [100]. Cognitive load manipulation is one of its key elements, since it allows the intensity of the intervention to be matched to the participant's ability level and kept enjoyable, neither too frustrating nor too boring [163].

Cognitive load estimation remains methodologically inconsistent, for the reasons set out in the Problem Statement above. These include pooled-subject cross-validation and task-stacking designs, in which low load is often a single task, medium load a dual task and high load a triple task. Such designs can also be constructed entirely from visual components, which risks affecting eye movements differently across conditions and therefore the predictive ability of a model built on eye-tracking data. A model can then learn to distinguish between different tasks rather than detecting a genuine modulation of cognitive load within a single task, which is arguably a more realistic representation of real-life demands.

To test this hypothesis, the thesis leverages the virtual reality environment to develop and evaluate a physiologically driven approach to adapting difficulty in cognitive training. The task used across the experimental work is a sustained attention task designed explicitly for VR for the purpose of this thesis. Using the eye-tracker integrated within the headset, the aim is to estimate cognitive load through a combination of eye-tracking metrics and physiological data, specifically galvanic skin response and heart-rate derived metrics. These are used to develop an online, AI-driven adaptive system, related in architecture to a Brain-Computer Interface but based solely on eye-tracking and physiological data, in which task difficulty is adjusted throughout the experimental session according to the participant's estimated cognitive load, derived from eye-tracking and physiological data. Once developed, the model is deployed in real time to assess how participant behaviour differs between a condition in which participants control their own difficulty and one in which the model controls it for them. Supplementary analyses assess how well

the model generalises between single-task and dual-task designs. They also examine the challenges of designing adaptive VR interventions for older participants. A further analysis investigates whether subjective self-report remains a trustworthy reference once the system moves from offline validation to real-time, low-contrast deployment (contrast here meaning the size of the step between adjacent difficulty levels: 0.5 s in the offline experiments, 0.1 to 0.3 s in the real-time system). The research questions and hypotheses that test the central hypothesis are set out in the next section.

1.4 Research Questions and Hypotheses

The research questions below test the central hypothesis stated in the Thesis Statement. The seven research questions and their hypotheses (H1 to H7) divide it into three parts. H1 to H3 concern whether cognitive load can be estimated from these signals in a within-task VR paradigm, and for older adults. H4 to H6 concern whether a model-driven controller can replace participant-driven adjustment once deployed. H7 concerns whether the participant's own rating is a stable enough reference for judging either. Each hypothesis is revisited with its outcome in Chapter 8.

1. **RQ1.** Is eye-tracking and physiological data fusion a suitable proxy for predicting cognitive load in a dynamic, single-task setting where load changes are subtler than in task-stacking designs? (Chapter 3)
   H1: A model combining eye-tracking and physiological features will predict subjectively-reported cognitive load above chance under leave-one-participant-out validation, despite the absence of the discrete, task-stacked conditions typically relied upon in the literature.

2. **RQ2.** Is eye-tracking suitable to reliably estimate cognitive load in older participants in VR? (Chapter 4)

   H2: A model built on eye-tracking data will retain meaningful predictive validity for cognitive load in older adults, but only once age-related shifts in baseline oculomotor physiology, such as senile miosis, are accounted for rather than treated as noise.

3. **RQ3.** Are there design considerations that need addressing when training models between younger and older populations? (Chapter 4)

   H3: The physiological and oculomotor features most predictive of cognitive load will differ systematically between younger and older cohorts, reflecting genuine differences in underlying physiology rather than a weaker or absent cognitive load signal in older adults.

4. **RQ4.** Does a model trained on a single-task design generalise well to a dual-task design? (Chapter 6)

   H4: A model trained on single-task data will show reduced, but non-trivial, predictive performance when applied to a dual-task paradigm, reflecting partial rather than complete generalisation across task structures.

5. **RQ5.** Is a model-driven adaptive setting a suitable replacement for participant-driven designs? (Chapter 6)

   H5: A model-driven adaptive system will maintain engagement and task performance comparable to a participant-driven system, without a corresponding increase in subjective frustration.

6. **RQ6.** Does real-time adjustment affect participant difficulty judgments compared to a fixed task design? (Chapter 6)

   H6: Real-time adaptation, whether model-driven or participant-driven, will alter how participants judge task difficulty relative to a static, fixed-difficulty design, particularly as task complexity increases.

7. **RQ7.** Are subjective workload assessments a reliable ground truth in real-time adaptive designs? (Chapter 7)

   H7: As task environments shift from discrete, high-contrast difficulty conditions to continuous, low-contrast real-time adaptation, the correspondence between subjective difficulty judgments and objective, physiologically-indexed cognitive load will weaken.

RQ7 differs from the others in that it questions the reference against which much of the thesis is judged. The models in Chapters 3 to 6 are trained and scored against participants' own difficulty ratings. The concern, developed in Chapter 2, is that such ratings are relative to the range of difficulty a person has recently experienced. This matters more when adaptation moves from large steps between difficulty levels (0.5-1.5 s in the offline experiments, referred to here as high contrast) to small ones (0.1 to 0.3 s in the real-time system, referred to as low contrast). H7 states the expected consequence: the correspondence between subjective judgments and objective, physiologically-indexed load weakens as contrast falls.

The outcome of each hypothesis is summarised in Table 8.1, and the central hypothesis is assessed against those outcomes in Chapters 8 and 9.

1.5 Thesis Overview

Figure 1.1 summarises how the research questions above map onto the remaining chapters of this thesis, and how each chapter's contribution builds on the one before it.

Chapter 2 provides the theoretical and empirical background for the thesis. Chapter 3 develops the offline cognitive load estimation model that underpins the remaining chapters, addressing RQ1. Chapter 4 extends this model to an older adult cohort, addressing RQ2 and RQ3. Chapter 5 builds and offline-validates a model architecturally suited to real-time use. Chapter 6 deploys that model in real time within a closed-loop adaptive

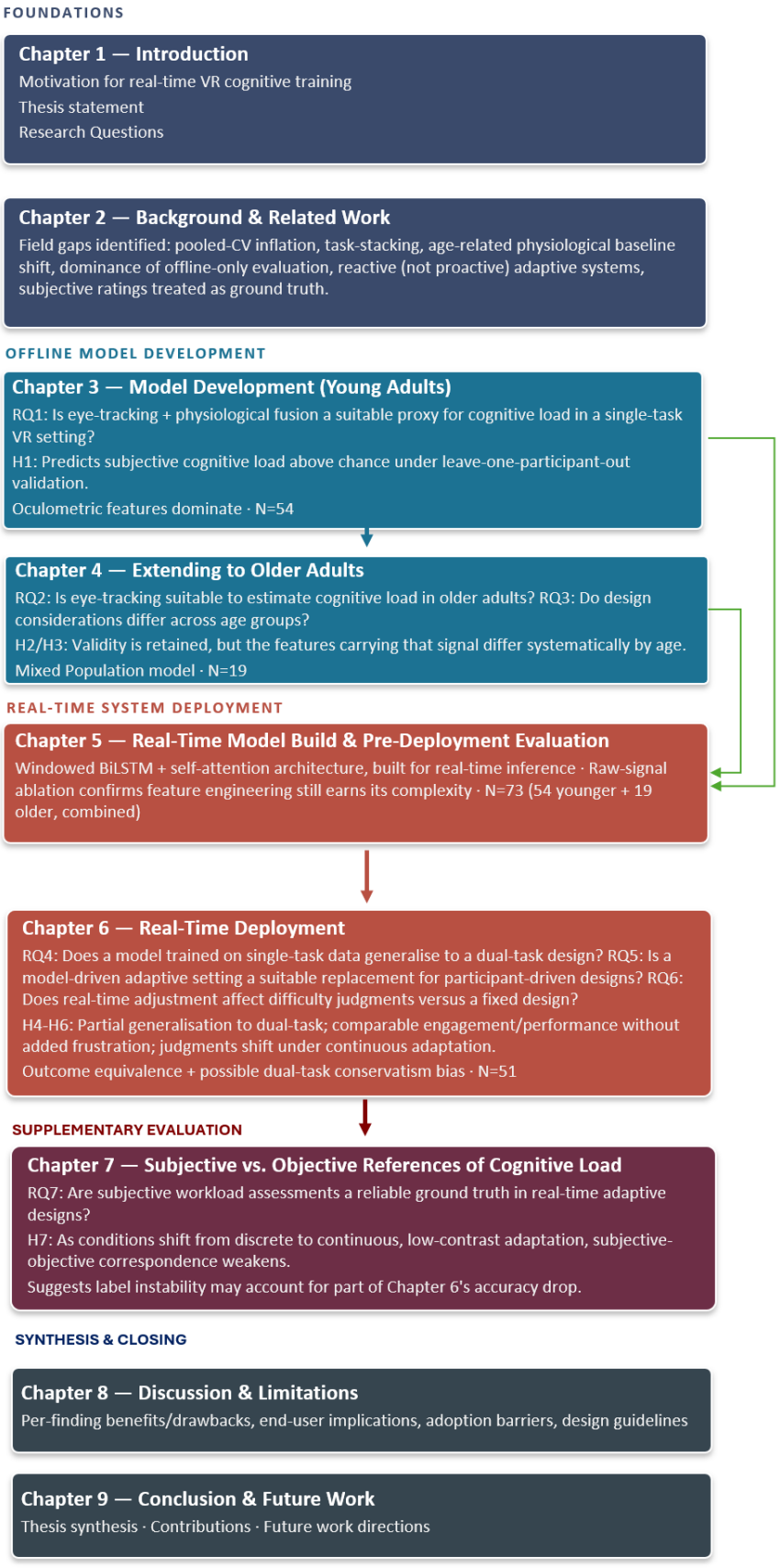


Figure 1.1: Graphical overview of the thesis structure, showing how each experimental chapter addresses a distinct subset of the research questions introduced in this chapter.

system, addressing RQ4, RQ5, and RQ6. Chapter 7 examines the relationship between subjective and objective cognitive load assessment under the real-time, low-contrast conditions introduced in Chapter 6, addressing RQ7. Chapter 8 discusses the findings and their limitations, and Chapter 9 concludes the thesis.

Chapter 2

Background and Related Work

This chapter provides the theoretical and empirical foundation for the thesis. It begins with the neuropsychological, educational and computational concepts required to understand adaptive cognitive training. It then turns to a detailed review of related work, organised around the gaps identified in Chapter 1 and the research questions they motivate. Rather than treating background and related work as separate concerns, many subsections close by relating the literature discussed to the specific contribution made in the corresponding experimental chapter.

2.1 The Neuropsychological Foundation. Defining Cognition, Cognitive Decline, and Cognitive Training

With age, some mental capacities decline gradually across a number of domains, such as the ability to multitask, attention and memory [124]. The development of advanced cognitive training systems is aimed at alleviating or slowing that decline so that it affects the activities of daily living less [17]. This section explains the core concepts related to cognitive interventions.

2.1.1 Normal Cognitive Ageing and the Definition of Cognitive Decline

Different cognitive abilities, which include thinking, reasoning, and memory, peak at different ages, with many fluid abilities peaking in early adulthood and declining gradually thereafter [67]. This age-related cognitive decline is considered a normal part of the ageing process and commonly manifests as an overall slowness in thinking, difficulties with multitasking, challenges in holding information in mind, and occasional word-finding difficulties [139]. These changes are linked to common aspects of ageing, such as decreased volumes in key brain structures, including the hippocampus, frontal lobe, and temporal lobe [136]. It is important to note that not all abilities decline, as vocabulary, reading, and verbal reasoning often remain stable or may even improve with age [139].

These mild, subtle changes must be distinguished from abnormal ageing conditions like Mild Cognitive Impairment (MCI), where decline exceeds what is expected for age while daily function is largely preserved, or dementia, where decline is severe and impacts daily function, such as navigation or problem-solving [126]. As noted in Chapter 1, this thesis is concerned with healthy ageing rather than these clinical populations, though the underlying measurement questions it raises may matter even more once cognitive impairment is present. Regardless of age, users with reduced processing speed or attentional capacity, whether due to normal ageing, early cognitive decline, or individual variability, are particularly vulnerable to training designs that fail to adapt [71]. Their limited resources can be overwhelmed, which motivates a difficulty control system that automatically and subtly manages the challenge level to reduce the risk of mental overload, creating a system that adapts to its user, further referred to as an adaptive system.

2.1.2 Cognitive Training: Purpose, Scope, and Limitations

Cognitive training is a structured and repeated practice of challenging tasks designed to enhance specific cognitive domains such as memory, executive function, or attention. While cognitive training offers benefits, traditional methods commonly face two challenges: ensuring the training is sufficiently personalised to the individual's exact needs, and maximising the transferability of the learned skills to real-world situations [182]. Adaptive systems aim to address the first of these by personalising instruction in real time, continuously adjusting task difficulty to match the user's evolving ability. Virtual reality platforms immerse users in simulated, naturalistic environments that can closely resemble everyday tasks, which may make the training more representative of everyday tasks and increase the likelihood that skills transfer to real-world situations, compared to traditional screen- or paper-based methods [187]. Moreover, VR environments allow experimenters to present highly controlled, standardised stimuli while limiting external distractions, offering a level of experimental control and replicability that is difficult to achieve in unstructured, naturalistic real-world settings [183].

2.1.3 Virtual Reality as an Immersive Training Platform

Building on this, VR's value for cognitive training extends beyond its representativeness of everyday tasks and its experimental control, since its immersive nature may enhance user engagement in its own right [125]. For instance, a person recovering from a stroke can navigate a virtual supermarket to re-learn skills involving decision-making, spatial orientation, and memory [125]. This contextual learning environment may make it more likely that the cognitive gains achieved during training transfer to real-world situations, supporting independence.

The application of VR with older adults specifically requires careful consideration of hardware and software design, a point developed in Section 2.6 of this chapter, and returned to empirically in Chapter 4.

2.1.4 Experimental Designs in Cognitive Training: Single vs. Dual-Task Paradigms

In the design of cognitive training protocols, tasks are generally categorised into single-task or dual-task paradigms. A single-task design focuses on one specific cognitive domain, such as responding to a visual target, requiring the participant to allocate their attentional resources to a single stream of information. It is often used to establish either baseline performance or to train a specific isolated skill.

Conversely, dual-task or multi-task designs require the participant to process two or more streams of information simultaneously [111]. For example, a user might be asked to solve arithmetic problems while simultaneously navigating an environment or responding to visual cues. In a research context, dual-task paradigms are frequently employed to simulate real-world complexity and to artificially increase cognitive load, as they tax executive functions and force the division of limited attentional resources.

Because they more closely approximate the concurrent demands of everyday activities, such as conversing while walking or monitoring surroundings while performing a manual task, dual-task designs are also often considered more representative of everyday tasks than single-task paradigms. Dual-task training has also been reported to improve real-world cognitive and physical function in older adults with cognitive impairment [4]. Whether a model trained on single-task data generalises to a dual-task context is examined directly in Chapter 6.

2.1.5 Sustained Attention: A Critical Component of Cognitive Function

Sustained attention is among the abilities reported to change with age, with declines emerging gradually across adulthood rather than at any specific age threshold. The precise nature of this relationship remains debated, however, as it is shaped by task demands, individual differences in strategy, and compensatory neural mechanisms rather than age alone [169, 24, 71].

Sustained attention is the fundamental ability to maintain focus on a specific task or stimulus over an extended period [107]. It is a crucial element of daily activities such as driving, working, reading or studying, and lapses in sustained attention have documented real-world consequences, as introduced in Chapter 1 [141]. Because effective cognitive training demands consistent mental engagement, if a user's sustained attention falters, the training stimulus becomes less effective, which may lead to frustration and reduced productivity [76].

Two constraints are important when designing sustained attention tasks specifically: capacity, the volume of information being attended to, and selectivity, the extent to which unattended information is still processed [37]. When the intrinsic, task-specific, load diminishes, the influence of distractors grows because additional processing capacity becomes accessible [135, 96]. Two main risks in managing cognitive load within a sustained attention task therefore are that the design might be overly simplistic, permitting distractors to divert attention from the task, or that the intrinsic load could be excessively elevated, surpassing the participant's attentional capacity and thereby impairing performance. A typical experimental setup in this domain involves reacting to chosen stimuli during extended intervals of inactivity, such as observing a sequence of letters and responding solely to a rare target [62]. Owing to these prolonged inactive periods, this approach is susceptible to generating boredom, which in turn can lead participants to withdraw from the task and direct greater attention toward distractions. In contrast, ongoing assessment of cognitive load facilitates real-time adjustments to the assigned

task load, ideally resulting in a task that is neither overly demanding nor insufficiently stimulating [103]. Mental arithmetic can serve as a secondary task for cognitive load manipulation in this context, since it allows controlled cognitive load conditions to be created, and arithmetic operations are commonly encountered in daily life [111].

2.1.6 Understanding Cognitive Load Theory

A core concept governing effective learning is Cognitive Load Theory, pioneered by Sweller [161]. Effective learning is generally understood as the formation of durable, transferable changes in long-term memory, knowledge and skills. These changes persist and can be applied beyond the immediate training context, whereas improved performance during instruction itself may not last [154]. Cognitive Load Theory explains the total mental strain placed upon the learner's limited working memory capacity during instruction. It was developed for instructional design, and its application here to a sustained attention task is an extension of that framework. Cognitive load is traditionally partitioned into three types, although the status of germane load in particular has been debated [46, 162].

1. **Intrinsic Load.** The inherent difficulty of the material itself, determined by its element interactivity, the number of interdependent pieces of information that must be processed simultaneously, relative to the learner's prior knowledge [163]. Intrinsic load is not simply minimised or maximised. Instead, it must be carefully managed, since some minimal level is necessary and productive for learning, but excessive levels exceed working memory capacity. Techniques such as segmenting complex material or providing pre-training on isolated elements can reduce the effective intrinsic load without altering the material's underlying complexity [163].

2. **Extraneous Load.** Load imposed by poor instructional design, confusing interfaces, or irrelevant information [161]. This should generally be minimised, as it does not contribute to learning. However, what constitutes extraneous load is not fixed, since instructional guidance that reduces extraneous load for novices, such as worked examples or integrated explanations, can become redundant, and therefore

extraneous, once a learner has developed sufficient expertise. This phenomenon is known as the expertise reversal effect [81]. Effective adaptive systems must therefore adjust the level of guidance as user competence changes, rather than applying a fixed minimisation strategy.

3. **Germane Load.** The mental effort dedicated to constructive learning, such as integrating new information into long-term memory structures, or schema [161]. While generally maximised within the learner's available capacity, germane load can also be deliberately increased through instructional techniques such as worked-example variability, which encourages deeper schema construction [171]. This is most beneficial once extraneous load has first been minimised, since increasing germane load without addressing extraneous load risks overwhelming working memory altogether.

The theory suggests that when instructional materials are well designed and readily available, they reduce extraneous cognitive load, thus allowing the learner to focus their resources on meaningful learning tasks, or germane load [161]. Since populations experiencing cognitive decline may possess reduced attentional and processing resources, they may be particularly vulnerable to high extraneous cognitive load. One study found that older adults benefited more than younger learners from instructional formats that minimise extraneous load, such as worked examples, achieving equivalent or superior performance while investing less subjective mental effort [171]. Applied to an adaptive system, this suggests that its success depends in part on its ability to dynamically manage task difficulty and, where relevant, instructional presentation, so that capacity remains available for germane load and extraneous load does not overwhelm the learner. This makes cognitive load a key real-time marker for an adaptive system to monitor.

2.1.7 Subjective Workload Assessment: The NASA Task Load Index

While Cognitive Load Theory provides the theoretical framework for understanding mental capacity, measuring this load in experimental settings requires standardised validation tools. The NASA Task Load Index, developed by Hart and Staveland [66], is the most widely utilised instrument for assessing subjective workload across cognitive load research [58]. Its dominance stems less from clear psychometric superiority than from practical factors: it is freely available, quick to administer, validated across a very wide range of domains, and its long history of use has produced an extensive normative dataset against which new results can be benchmarked [58].

Other multidimensional subjective workload instruments exist, most notably the Subjective Workload Assessment Technique and the Workload Profile. One comparison of the three found that while all three show high convergent validity with one another, the NASA Task Load Index correlates most strongly with task performance, whereas the Workload Profile demonstrates superior diagnostic power in distinguishing between different sources of workload, such as perceptual versus cognitive demand [142].

The instrument rates perceived workload on six distinct subscales: mental demand, physical demand, temporal demand, performance, effort, and frustration. In a typical administration, participants rate each dimension on a scale immediately following a task block. While the original method involves a weighting procedure to account for individual biases, many researchers employ the unweighted Raw version, since it has been reported to have comparable sensitivity and is simpler to administer [65]. However, as with any self-report instrument, its reliability is bounded by the limits of introspection. Ratings can be influenced by recall bias and by individual differences in how respondents interpret the scale. Retrospective judgements also tend to be shaped by the most demanding or

the most recent moments of a task, rather than reflecting a stable average across its duration. A further limitation is its post-hoc nature, since it provides a summary of workload after the event or task, rather than the higher-frequency resolution needed for dynamic difficulty adjustment [66].

In this thesis the NASA Task Load Index is used to compare workload between conditions and between populations, from questionnaires completed after each session or condition. It is not used to assess participant state at the frequent intervals a real-time adaptive system requires, which is how subjective workload is often used as ground truth in studies with longer task periods. Instead, the perceived-difficulty measure that serves as the model target throughout the experimental work is a purpose-built, higher-frequency instrument. It is administered after every sixty-second sub-session rather than at the end of a task block. It is grounded directly in Flow theory rather than in the general workload constructs underlying the NASA Task Load Index. The instrument, its categories, and its administration protocol are described in detail in Chapter 3. It is introduced here to motivate the design choices behind this alternative, given the limitations of the NASA Task Load Index as a frequent ground truth for a real-time adaptive system.

For adaptive systems research more broadly, subjective workload report of this kind serves as one important source of evidence when validating physiological or behavioural workload proxies, rather than as an infallible ground truth. Physiological signals, such as pupil diameter or heart rate variability, and subjective self-report each capture different, only partially overlapping aspects of workload. Subjective and performance-based measures of workload have been reported to dissociate under a range of conditions [173, 186], with subjective ratings driven more strongly by the amount of effort a person feels they have invested, while performance and physiological measures are more sensitive to the actual competition for limited cognitive resources [186]. Subjective report is often treated as ground truth because it is the measure in this pipeline most often treated as capturing the construct of interest, perceived cognitive load, directly, whereas physiological signals are, by definition, indirect proxies with no ground truth of their own. Recent methodological critiques have explicitly challenged this convention, arguing that the psychometric

weaknesses of instruments such as the NASA Task Load Index make them a questionable target for prediction and calling for more rigorous validation of what physiological models are actually learning to predict [14]. As such, if a machine learning model predicts high load from eye-tracking data while a user rates their mental demand and effort as low, this discrepancy does not necessarily indicate the model is wrong. It may instead reflect a genuine dissociation between subjective experience and underlying resource allocation, a possibility examined directly in Chapter 7.

Whatever the instrument, a rating also depends on the frame of reference of the person giving it. Judgments of magnitude and category are relative rather than absolute: people tend to divide the range of stimuli they have recently experienced into categories and to use those categories with similar frequency, which is the core of range-frequency theory [122, 123, 112]. A rating of a task as *Too Difficult* or *Just Right* may therefore depend on the range of difficulty experienced as well as on the difficulty of the task at hand. People are also imperfect judges of their own performance, and their metacognitive sensitivity falls as a task becomes noisier [53, 133]. These are drawbacks of subjective labels that apply whenever the rating is collected. Whether they matter in the present design, where the steps between difficulty levels shrink from 0.5-1.5 s in the offline experiments to 0.1-0.3 s in the real-time system, is tested in Chapter 7.

2.1.8 Reference Measures for Cognitive Load: Benefits and Drawbacks

The measures reviewed above serve different roles in a cognitive load model, and none is an uncontested reference. Timing separates the options distinctively. An end-of-block questionnaire is retrospective and takes a relatively long time. A per-level rating is available every minute but interrupts play. Performance is continuous and free. Physiological signals are continuous but need a reference to be interpreted. Older adults complicate each family in a different way. Self-report can dissociate from performance and physiology. Performance can reflect motor and sensory limits, or be maintained through compensat-

ory effort. Physiological baselines shift with age. This thesis uses all four and treats the choice of reference as a question in its own right (RQ7).NASA-TLX is used to compare workload between conditions and populations, per-level rating is used a target label for the models trained in chapters 3 to 6. Task performance is used as an outcome measure in chapter 6 and an alternative model target in chapter 7. Physiological and oculomotor signals are used as model input in chapters 3 to 6.

2.2 Educational and Motivational Frameworks for Adaptive Intervention

The reasoning behind adaptive interventions is grounded in educational and psychological frameworks. This section explains the mechanisms behind human learning and motivation, key to understanding why managing cognitive load in real time is important for designing effective adaptive systems.

2.2.1 The Zone of Proximal Development

The Zone of Proximal Development is a foundational concept used in education, first developed by Lev Vygotsky during the early 1930s [175]. It defines the theoretical space between what a learner can achieve independently and what they can achieve with the guidance of a more knowledgeable other, who might be an adult or a peer [175] or, in this context, an intelligent machine.

Vygotsky argued that genuine development occurs when educational experiences are provided specifically within the learner's Zone of Proximal Development, meaning tasks too difficult to complete unaided but achievable with appropriate guidance [175]. Learning occurs through social interaction, dialogue, and sense-making with the more knowledgeable other, gradually allowing the learner to internalise the skills needed to solve

problems independently [175]. A major implication of this framework is that traditional assessments, which only measure what a student knows or can do on their own, overlook the student's capacity to solve problems with assistance. The Zone of Proximal Development is not the sole determinant of learning. Motivational state, prior experience and the usability of the instructional interface also shape outcomes. A full treatment of these factors is beyond the scope of this section. It is emphasised here specifically because it speaks directly to task difficulty, which is the dimension an adaptive system is best positioned to manipulate in real time. For adaptive training systems, the technology can be viewed as taking a similar role to the more knowledgeable other. It adjusts the difficulty, and where relevant the guidance or support, needed to keep the user operating at the challenging but achievable edge of their ability. In this thesis the system adjusts difficulty only and provides no explicit guidance.

2.2.2 The Flow State: Achieving Optimal Experience

While the Zone of Proximal Development defines the ideal level of instruction, the Flow state describes the ideal experience of learning. The concept was proposed by psychologist Mihaly Csikszentmihályi, notably in his book *Flow: The Psychology of Optimal Experience* [43]. Flow is described as the optimal experience, commonly referred to in everyday language as being in the zone, characterised by deep enjoyment, creativity, and engagement [43]. During Flow, the individual's attention is largely focused on the task at hand [43].

To achieve Flow, three conditions must be met. The activity must provide clear goals and immediate feedback. Most importantly, there must be a good balance between the perceived challenges of the task and one's perceived skills [43]. If the challenges are too low relative to skills, the user drifts into boredom or apathy. On the contrary, if the challenges are too high, the user experiences anxiety or frustration [43]. Cognitive training, especially

in clinical settings, can be tedious, and adherence varies widely among older adults [166], with misaligned challenge levels being one plausible contributor. Designing the training with the principles of Flow in mind may provide intrinsic motivation and gratification, increasing the likelihood of sustained engagement and adherence [43].

This three-way structure, boredom below the optimal band, Flow within it, and frustration above it, is the theoretical basis for the perceived-difficulty instrument used throughout this thesis and introduced briefly in Section 2.1.7. Rather than asking participants to rate several independent dimensions of workload, as the NASA Task Load Index does, the instrument used here asks participants to locate their own experience directly within this three-category structure after every sub-session. The full instrument and its administration are described in Chapter 3.

2.2.3 Flow and the Zone of Proximal Development for Adaptive Designs

The Zone of Proximal Development and the Flow state are closely linked theoretical concepts for adaptive design. The former defines the optimal area of difficulty for instruction, while the latter defines the optimal experiential state achieved when operating in that difficulty area. A technological system that successfully maintains the user in the narrow overlapping band where perceived skill equals perceived challenge, and where guidance maximises potential, would be expected to function as an effective adaptive training environment (this synthesis of the two frameworks is the thesis's own).

To maintain this delicate balance, the system benefits from frequent, objective measurement of the user's cognitive and emotional state. If the system were to detect high anxiety/frustration, a negative physiological state indicative of high challenge relative to skill, it would mean it has pushed the user outside the optimal channel. This transforms

the adaptive system from a simple trainer into a dynamic control mechanism, where the continuously informed algorithm can take over part of the role of the knowledgeable human instructor, helping instruction manage cognitive load and reduce drift into states of boredom or anxiety [161, 43].

2.3 Real-Time Measurement and Physiological Feedback

To execute the dynamic control required by the Zone of Proximal Development and Flow, the adaptive system needs immediate, objective data that approximate the user's internal emotional and effortful state. This feedback often bypasses subjective self-report, providing raw, continuous data streams. As set out in Chapter 1, none of the modalities discussed below is an uncontested ground truth. Each is evaluated here on its own merits and practical constraints, and the reasoning for excluding or including each one in this thesis is made explicit.

2.3.1 Electroencephalography

Electroencephalography is a non-invasive method of monitoring the brain's electrical activity. It involves placing electrodes along the scalp to measure voltage fluctuations resulting from ionic current within the neurons of the brain [165]. Data is typically analysed by examining spectral power within specific frequency bands, such as Delta, associated with deep sleep, Theta, associated with drowsiness or memory processing, Alpha, associated with relaxation, and Beta, associated with active thinking and focus. Electroencephalography is widely used to approximate cognitive state, valued for its high temporal resolution and its capacity to provide a continuous measure [6]. It has shown

promise in predicting cognitive load in VR, with controlled binary classification tasks reporting accuracies as high as 81% [165]. In principle, it is a strong candidate proxy for cognitive load among the modalities considered in this thesis, although it remains an approximation of it.

In practice, however, its application in interactive, immersive VR is constrained by significant gaps between laboratory validation and real-world deployment. Many successful electroencephalography-based cognitive load models rely on highly controlled laboratory protocols designed to minimise signal artifacts [165]. These constraints often require participants to restrict physical movement to prevent myogenic interference [31] or suppress natural ocular behaviours, such as blinking, which generate electrical shifts that obscure low-voltage neural activity [189]. In the context of immersive VR, where head movement and interaction are fundamental to the experience, this reliance on physical stillness creates a tension, since the more naturalistic the interaction, the more artifacts are introduced [118]. Head-mounted display hardware compounds this further, introducing substantial electromagnetic noise, particularly around its refresh rate, which can obscure low-voltage neural signals [95]. Beyond signal integrity, the logistical burden of electrode placement and impedance checking can increase participant fatigue and pre-experimental stress [64]. Integrated guidelines for VR-EEG synchronisation report that this preparation can take thirty to forty-five minutes [64]. The combined weight of research-grade electroencephalography caps and head-mounted displays can also exceed 1.5 kg, which can lead to neck strain and a reduced sense of presence [95]. These practical and technical barriers have led researchers to argue that electroencephalography-based results are often not representative of everyday interactive use [118]. They are the reason this thesis, as stated in Chapter 1, does not adopt electroencephalography despite its theoretical advantages.

2.3.2 Eye-Tracking and Gaze Analysis

Eye-tracking technology captures metrics such as gaze direction, fixation duration and pupil dilation. It has been reported to be a sensitive and robust indicator of cognitive load in VR, often proving less susceptible to environmental noise than electroencephalography [12]. Fixation patterns reveal how long the user focuses on objects of interest, with longer fixation times having been interpreted both as the user being more in control of the task and as the user withdrawing and no longer trying to keep up with its pace [49]. Saccades, the rapid eye movements that occur between fixations, and their frequency, duration, and speed have been linked to mental demand in three-dimensional environments [49, 26, 158]. Recent research has identified saccade duration and peak velocity specifically as important predictors, with longer, slower saccades associated with lower cognitive load [104, 174]. Pupil dilation, along with related metrics such as the rate of pupil diameter change, is a widely used indicator of instantaneous cognitive workload [49], thought to reflect the mobilisation of effort through the locus coeruleus-norepinephrine system [11, 179]. Unlike electroencephalography, which requires strict environmental control to avoid visual-evoked potential interference, eye-tracking within a head-mounted display allows for the simulation of varying environmental conditions. By leveraging localised luminance masking or baseline-relative pupil dilation metrics, the contribution of the pupillary light reflex can be partly separated from cognitive load, keeping the VR experience naturalistic while limiting the loss of data quality [50].

A further practical consideration is eye-tracking frequency. Consumer VR headsets that allow control of lighting conditions typically sample gaze at around 90 Hz, considerably lower than the 1000 Hz used by high-grade research eye-trackers [15]. The implications of this lower sampling rate for the present work are discussed in Chapter 3. By measuring how the user allocates their external attention and processing intensity, eye-tracking provides insight into cognitive load fluctuations and, as reported in Chapter 3, was the most informative individual modality in this thesis.

2.3.3 Galvanic Skin Response

Galvanic skin response, also referred to as electrodermal activity, provides a measure of emotional arousal. It captures the skin's electrical conductance, which fluctuates in response to sweat gland activity controlled by the autonomic nervous system [28]. Because it reacts automatically and unconsciously to emotions such as stress, fear or excitement, it offers an objective, continuous indicator of a person's arousal. It has a latency of a few seconds and is sensitive to non-cognitive influences such as temperature and movement [28]. In the context of adaptive cognitive training, a sudden spike in this signal may indicate acute emotional stress or frustration, signalling that the task may have induced anxiety, deviating from the ideal Flow state [43, 28]. This feedback may allow a system to react to an early sign of emotional strain, rather than waiting for an explicit performance error.

2.3.4 Heart Rate Variability and Associated Metrics

Heart Rate Variability measures the fluctuation in the time intervals between successive heartbeats. This metric reflects the balance of the autonomic nervous system and is associated with cognitive workload and performance [3]. Research indicates that lower Heart Rate Variability is typically associated with higher cognitive workload [128], while higher variability is generally associated with better cognitive performance [3]. It is therefore commonly used as an indirect, objective proxy for mental effort, allowing a system to gauge whether the cognitive load imposed by a task is optimal or overwhelming [161, 128].

To quantify these fluctuations, researchers rely on specific time-domain metrics derived from the raw heartbeat intervals. The Standard Deviation of NN intervals represents the overall variability of the heart rate and reflects the combined input of both the sympathetic and parasympathetic branches of the nervous system, typically decreasing as cognitive demand increases [128]. The Root Mean Square of Successive Differences, calculated from

the difference between adjacent heartbeats, is the primary measure for estimating parasympathetic activity, which decreases as cognitive load increases [170]. The percentage of successive intervals differing by more than fifty milliseconds, commonly abbreviated pNN50, behaves similarly, tending to decrease as mental stress or workload increases [3].

2.3.5 Multimodal Fusion and Dataset Benchmarks

When eye-tracking metrics are fused with heart-rate derived measures and galvanic skin response, multimodal models have been reported to capture the autonomic signature of workload more effectively than single modalities [68, 25]. One driving simulator study reached 97.8% accuracy on a three-class task detection problem (no task, 1-back, 2-back), using a within-participant split in which each participant's own data appeared in both training and test [68]. The CLARE benchmark reports lower binary accuracies, with up to 85.6% under 10-fold cross-validation and 72.7% under leave-one-subject-out cross-validation [25]. As the field has matured, several dedicated public datasets have emerged to benchmark multimodal cognitive load estimation, though each carries distinct limitations. The COLET dataset provides eye-tracking data from 47 participants solving visual search puzzles under varying time pressure and multi-tasking conditions, relying entirely on subjective NASA Task Load Index scores for ground truth [90]. Addressing the need for peripheral physiological data, ADABase offers a comprehensive multimodal repository, including eye-tracking, electrodermal activity, and photoplethysmography, from 51 subjects in a driving simulator, though it induces workload primarily by stacking secondary working memory tasks onto a primary driving task [121]. Other recent contributions include GAZELOAD, which captures eye-tracking via smart glasses during industrial human-robot collaboration [83]. EM-COGLOAD explores cognitive load and ageing across 75 participants using videos of eye movements, although it relies on a two-dimensional, screen-based dual-task paradigm rather than an immersive VR environment [114]. A large-scale industry dataset from HP Labs, drawn from a broader sample of 738 participants, fuses pupillometry and photoplethysmography to predict load without user-specific calibration and is, of the datasets discussed here, the only one collected in VR

[178]. However, to modulate the difficulty, which closely maps onto their participants' cognitive load ratings, they stack three visual tasks. Arguably, this approach also has the potential to bias the participants' ratings, who will rank their ratings based on the explicit number of tasks, rather than own perception of the experience.

All of these datasets are offline and observational. As described in their papers, they capture physiological responses to predefined task scripts without any closed-loop, real-time adaptation of the environment based on the user's cognitive state, a gap this thesis addresses directly in Chapters 5 and 6. As discussed, the existing literature reveals two recurring methodological limitations. First, many models rely on task-stacking, layering concurrent tasks to induce load, which risks training algorithms to recognise specific task structures rather than a pure modulation of cognitive intensity [178]. Recent research advocates instead for within-task difficulty modulation, which helps separate intrinsic cognitive load by varying the complexity of a single task while maintaining context, and may provide a cleaner mapping of mental effort [13, 35, 27]. This is the approach adopted throughout this thesis, and its consequences for model training and generalisation are examined in Chapter 3. Second, there is a frequent reliance on pooled-subject cross-validation, discussed further in Section 2.5, which often results in inflated accuracy scores by allowing models to learn individual physiological signatures rather than generalisable cognitive states [25, 113].

While the studies discussed above support eye-tracking and physiological fusion as a credible proxy for cognitive load, almost all of them, with the exception of the HP Labs dataset, were conducted in desktop-based, simulator-based, or real-world settings rather than immersive VR. This thesis instead develops and validates its model within a fully immersive, gamified VR sustained attention task, designed explicitly to avoid task-stacking. Chapter 3 argues that this task is more representative of everyday tasks than comparable desktop paradigms, such as the n-back task [165]. It is also well suited to an adaptive, real-time deployment where the difficulty can be modulated incrementally, based on machine learning predictions.

2.4 Advanced Predictive Algorithms: Traditional Machine Learning and Deep Learning

The transition from traditional machine learning to deep learning marks a shift in computational statistics and artificial intelligence. For much of its history, the field of AI was constrained by limited computational power and small datasets [56]. During this period, traditional machine learning models were carefully refined to extract maximum value from limited data.

Today, to a casual observer, traditional machine learning and deep learning appear to perform the exact same function. They ingest historical data, identify underlying patterns and use those patterns to make predictions about unseen future events. Rather than treating them as opposing paradigms, this section considers both under a single umbrella of advanced predictive algorithms. They are distinguished less by their end goal than by how raw input data is turned into something a model can learn from: through features hand-crafted by a human, or through features the model learns directly from data [97, 23].

Traditional machine learning encompasses a suite of effective algorithms grounded in statistical theory, such as Support Vector Machines and decision tree ensembles [41]. Deep learning, conversely, utilises multi-layered artificial neural networks and frequently achieves strong performance across complex data domains, at the cost of far greater computational and data demands [97]. No paradigm is universally superior. The relevant question is which inductive biases best match the structure of a given dataset. This thesis returns to that question directly, since its features are engineered from physiological signals but retain a temporal, sequential structure that sits between the tabular and time-series cases discussed below.

2.4.1 Feature Engineering versus Feature Learning

The most defining distinction between classical machine learning and deep learning is how raw data is transformed into a mathematical format that an algorithm can process. This step strongly influences the upper limits of a model's performance and establishes the amount of human intervention required. This distinction is best understood not as a matter of whether the underlying data is structured or unstructured, but of whether the relevant features are hand-crafted by a human or learned automatically by the model.

Traditional machine learning algorithms are highly dependent on the format and engineering of their input features. They generally lack the intrinsic architectural ability to process raw, unstructured data such as a dense array of image pixels or a raw audio waveform and extract high-level semantic meaning directly [23, 56]. Consequently, they typically require a manual, labour-intensive process known as feature engineering, in which domain experts manually construct the relevant data points for the model to analyse. In the eye-tracking domain relevant to this thesis, information such as gaze direction can be used to compute the number and duration of fixations, which can in turn be used to engineer more informative features, such as the trend in fixations over a sixty-second period. A key limitation of this paradigm is that the algorithm cannot learn from patterns the human engineer has not explicitly defined. If an informative variable is omitted, the model cannot recover it [23].

2.4.2 Deep Learning and Representation Learning

Deep learning reduces reliance on manual feature engineering through a paradigm known as representation learning [97, 23]. Instead of relying on human experts to define exactly what to look for, a deep neural network takes raw, unprocessed data and automatically discovers representations useful for detection or classification, which typically emerge in hierarchical form as a by-product of this layered process. A neural network is built from layers, successive groups of interconnected artificial neurons, each of which takes the

output of the previous layer, applies a learned mathematical transformation, and passes the result forward. As raw data passes through successive layers, the network constructs a compositional hierarchy of features of increasing abstraction. No human programs this extraction process. It is optimised dynamically during training as the network learns which features minimise its predictive error.

2.4.3 Domain Specialisations: Matching the Architecture to the Data

Because of their differing mathematical assumptions, known as inductive biases, these two paradigms often excel in fundamentally different environments. It is rarely a matter of one approach being universally superior but of selecting the appropriate tool for the specific data topology.

While deep learning has become highly prominent, traditional machine learning, specifically tree-based ensemble models like Random Forests and XGBoost, often remains the preferred, highly efficient choice for structured, tabular data [30, 34, 59]. Tabular datasets such as those found in financial records or medical databases typically contain highly heterogeneous features, combining continuous numbers, categorical labels, and boolean values within a single dataset. Tree-based models divide data using rigid boundaries aligned strictly to individual axes and do not assume that these distinct columns can be mathematically blended together, making them well suited to handling the irregular patterns and disparate feature types that characterise spreadsheets [59]. This is directly relevant to the present work. The engineered eye-tracking and physiological features used throughout this thesis are heterogeneous and tabular in form, which makes tree-based models a natural comparison point against sequential deep learning approaches. The underlying signals also retain a temporal structure, which motivates the recurrent architectures discussed next.

Temporal and sequential data, such as spoken language transcripts or financial market prices, presents a different challenge, since the sequential order of the data points matters greatly. Traditional time-series models often rely on rigid statistical properties and manually defined time windows [56]. Recurrent Neural Networks address this by introducing an internal mathematical memory state. At each sequential time step, such a network processes the current input alongside a hidden state generated from the previous time step, allowing it to maintain a running memory of the sequence [74, 23]. However, standard recurrent networks struggle with long-term memory due to a mathematical issue where learning signals shrink exponentially over time. Long Short-Term Memory networks were engineered specifically to solve this [74]. They are built around a central cell state that acts as a protected informational highway. Access to this state is regulated by gates that analyse the current input dynamically. The gates decide which historical information should be discarded, which new information is crucial and what should be output at the current step [74, 56]. A recent review of adaptive VR systems highlights Long Short-Term Memory networks as promising for real-time cognitive load estimation, given their ability to process the time-series nature of physiological data, while also noting that their application in real-time VR settings remains comparatively underexplored [115]. This architecture, extended with a self-attention mechanism, is the one built and offline-validated in Chapter 5 and deployed in real time in Chapter 6.

2.4.4 The Optimisation Landscape and Training Strategies

At its core, most machine learning is an optimisation problem where the ultimate goal is to find the specific set of parameters that minimises a loss function, which mathematically measures the difference between a model's predictions and reality.

Many traditional machine learning algorithms rely on convex optimisation, in which any local minimum found is guaranteed to be a global minimum, making training stable and predictable [56, 44]. Deep learning, in contrast, is governed by non-convex optimisation [56, 44], in which the landscape is filled with countless local minima, flat plateaus,

and steep ridges, so that finding the global minimum is computationally unmanageable. Finding it is generally unnecessary, since landing in a high-quality local minimum is often sufficient for good predictive performance [44]. Deep learning therefore relies on strategies such as Stochastic Gradient Descent, whose noise can help the algorithm escape sub-optimal regions [56]. Adaptive optimisers such as Adam tailor the learning rate for every parameter [87, 56]. Batch Normalisation further stabilises training by standardising the inputs to each layer [78].

2.4.5 Regularisation: Preventing Memorisation

Because deep neural networks can contain millions of parameters, they can memorise the training data, which leads to overfitting: the model performs very well on its training data but fails to generalise. Regularisation prevents this. Weight decay, a widely used technique in both traditional and deep learning, penalises large weights so that the model prefers simpler solutions [192, 73]. Dropout, a cornerstone regularisation strategy developed for neural networks, randomly removes neurons during training so that each must learn features that remain useful regardless of which others are active [156, 73, 85, 192, 160, 72].

2.4.6 Scaling Laws and Data Demands

The strengths of deep learning come at a substantial cost in data, since these models are notoriously data-hungry [160, 72]. Traditional algorithms such as Support Vector Machines often reach near-optimal performance with a modest amount of structured data, with diminishing returns beyond that point, which makes them well suited to specialised, data-poor domains. Deep networks follow empirical scaling trends [82, 72]: performance

has, in many domains, increased predictably with data and compute. They therefore often show their advantage over traditional algorithms only with large volumes of data. On a very small dataset, a deep network can be outperformed by a simple linear model [160].

2.5 Generalisation, Validation and Dataset Shift

A trustworthy cognitive load model has to generalise beyond the data it was trained on. It can fail to do so in two distinct ways: for a new person, and in a new context. Choosing an appropriate algorithm and securing sufficient data are necessary but not sufficient to prevent either, and the reported performance of a model is only as meaningful as the evaluation used to obtain it.

Generalising to new people is what cross-validation is meant to test, and how the data is partitioned matters particularly in physiological computing, where inter-individual differences are large. In pooled-subject cross-validation, data from all participants is split into training and testing sets without regard for participant identity. This yields higher accuracy but risks data leakage. The model can learn individual-specific signatures, such as a participant's baseline pupil size or resting heart rate, rather than generalisable cognitive states. The results can then be optimistic and fail to hold for a new person [25, 151, 113]. Leave-One-Subject-Out cross-validation instead tests on a participant withheld from training, and recent benchmarks commonly show a drop in performance relative to pooled protocols, reflecting the difficulty of inter-subject physiological modelling [25, 16]. Because real-time, user-independent deployment is itself a leave-one-out scenario by construction, this protocol is adopted throughout the model development described in Chapter 3, and the resulting drop in accuracy is treated as an expected cost of building a transferable system.

Generalising to a new context is a separate question. Even a model that works for new participants may not do so when the conditions of use differ from those under which it was trained. In machine learning terms, the distribution of the input features can change between training and use, which is called covariate shift [159, 148]. The relationship between the features and the label can change as well, for example when the task, the sensors or the way labels are collected differ between the two settings. Validation on unseen participants from the same data collection does not address this. Deploying a model trained offline in a live adaptive system, as in Chapter 6, is such a change, and Chapter 7 examines whether one aspect of it, the smaller steps between difficulty levels, changes how reliable the participant's own rating is as a reference.

2.6 Older Adults: Learning, Technology Acceptance, and the Suitability of Physiological Assessment

Much of the physiological computing literature reviewed so far, including many of the models it produces, appears to be developed and validated on younger cohorts, often university students, a pattern documented for behavioural research more broadly [70]. EM-COGLOAD is an exception that includes ageing. Transferring these models directly to older adults, one of the primary intended beneficiaries of adaptive cognitive training, is not something that can simply be assumed to work, for reasons that are both practical and psychophysiological.

2.6.1 Virtual Reality for Cognitive Training in Older Adults

Research applying VR to older adults generally falls into three categories. Testing validates VR against traditional metrics. Screening aims to distinguish healthy ageing from cognitive impairment. Training uses VR to improve cognitive skills and activities of daily living [152]. The present thesis sits within the third category. Longitudinal designs have

been used to assess the efficacy of VR training in this category. Szczepocka et al. conducted a twelve-week intervention using a VR-based task in healthy older adults aged 65 to 85. The task was structured around focused attention, task switching and working memory. They found improvements over time in visual memory, sustained attention and working memory, which suggests that VR is a promising cognitive intervention for this population [164]. This is consistent with a systematic review that narrowed its search to ten key studies of VR-based training for cognitive disorders. The review found a generally medium positive effect on symptom improvement, with gamified interventions showing a stronger effect than traditional programs. VR also improved adherence to treatment for recipients, through increased engagement, and for providers, through automated protocol delivery and objective progress monitoring [117]. That evidence concerns people with cognitive disorders rather than healthy older adults.

2.6.2 Physiological Baseline Differences with Age

A significant barrier to applying standard physiological computing models to older adults is a shift in biological baselines. Ageing introduces physiological constraints that could be misread by algorithms trained on younger users as indicators of cognitive state rather than as constants of ageing itself. Eye-tracking, which is the most informative individual modality in Chapter 3, is also susceptible to ageing effects. With age, the pupil becomes smaller and more rigid thought to be due to atrophy of the iris dilator muscle, a condition known as senile miosis [184]. A standard algorithm might therefore interpret a younger adult's dilated pupil as a sign of cognitive effort, whereas an older adult's pupil may be physically limited in such dilation, capping at a smaller diameter even under high load. Because baseline pupil size is naturally smaller in older adults, models may also misinterpret this resting state as low arousal, when it is simply a biological constant of ageing [130]. Similar shifts occur in physiological measures more broadly. Heart Rate Variability, a standard metric for mental workload and stress, decreases with age [167], which may compress the dynamic range available to reflect task engagement. Without

accounting for this baseline shift, high-demand VR tasks may yield physiological data that looks superficially unresponsive in older adults, potentially masking the cognitive cost they are experiencing. Whether these baseline shifts can be disentangled from genuine differences in cognitive load reactivity is examined directly in Chapter 4.

2.6.3 Self-Assessment Reliability in Older Adults

Compounding this physiological variance is the potential unreliability of self-report in older cohorts. While younger adults often report workload that correlates moderately well with objective difficulty, older adults have been reported to exhibit a subjective-objective dissociation [51]. Older adults may report higher frustration or subjective workload even when performing well [1]. Conversely, they may mask difficulties to avoid appearing impaired [55]. This dissociation can render standard subjective feedback loops, such as dynamic difficulty adjustments based solely on self-report, noisy and potentially misleading for this population [180]. Compensation is a further reason for divergence: according to the Compensation-Related Utilisation of Neural Circuits Hypothesis (CRUNCH), older adults recruit additional neural resources to maintain performance on tasks of low to moderate difficulty [138], so similar performance or similar reported effort can conceal a higher underlying cost. This is the older-adult-specific instance of the broader ground truth problem introduced in Chapter 1 and it is more pronounced than the general case. Self-report becomes less reliable with age, and the physiological baselines that might otherwise substitute for it are simultaneously shifting for reasons unrelated to cognitive load. Validating a VR cognitive training task for this demographic therefore requires determining whether objective physiological markers can serve as a more dependable reference than self-report once these baseline shifts are properly accounted for, a question taken up directly in Chapter 4.

2.6.4 Technology Acceptance and Usability in Older Adults

A further consideration, distinct from measurement reliability, is whether older adults are willing and able to use the technology an adaptive system depends on at all. The Technology Acceptance Model [45] and its successors, most notably the Unified Theory of Acceptance and Use of Technology [172], explain adoption through constructs such as perceived usefulness, perceived ease of use and behavioural intention. They have been applied specifically to VR and wearable sensing technologies aimed at older users. A study examining acceptance of a VR headset designed for fall prevention found that older adults' intention to use the device was positively predicted by perceived usefulness, perceived enjoyment, and perceived ease of use, supporting the applicability of this framework to head-mounted VR hardware specifically in this population [110]. Acceptance modelling of wearable health-monitoring devices for older adults has similarly identified perceived usefulness, compatibility, facilitating conditions, and self-reported health status as predictors of intention to use [99]. Beyond acceptance and willingness to use a device, older adults' capacity to learn a novel interactive task is itself not fixed. The enrichment hypothesis suggests that engagement with cognitively demanding, novel activities can help preserve functional capacity in older adults [71], which is consistent with the view that a well-designed adaptive VR task is not merely tolerated by older users but potentially beneficial to them independent of its cognitive training content. Cybersickness, or visually induced motion sickness, can skew physiological data and degrade cognitive performance if not addressed through appropriate hardware and software design [88]. Evidence on age and susceptibility is mixed: a meta-analysis found lower total Simulator Sickness Questionnaire scores in older than in younger samples, although few of the included studies involved older users [147]. Kourtesis et al. argued that VR cognitive training should adhere to technical guidelines, such as high frame rates and ergonomic interactions, to be valid in any population, and particularly in older adults. The aim is to ensure that the

measured cognitive load results from task difficulty rather than from the struggle to use the interface. Increased physiological arousal caused by simulator sickness can mimic the signals of high cognitive load, potentially leading to false positives in adaptive systems [88].

Taken together, the acceptance, usability, and baseline-physiology literature reviewed in this section suggests that older adults are neither an edge case to be accommodated after the fact, nor a population for whom existing younger-cohort models can simply be assumed to transfer. Chapter 4 treats this as a central empirical question rather than a limitation to be noted and set aside.

2.7 Brain-Computer Interfaces and Adaptive Environments

Having reviewed how physiological and behavioural signals can be modelled and validated, the remainder of this chapter turns to how such models can be embedded within a complete, real-time system. The system proposed in this thesis is closely related in architecture to a Brain-Computer Interface, although it uses eye-tracking and peripheral physiology rather than direct neural recording.

A Brain-Computer Interface facilitates direct communication between the brain and an external device [185]. It consists of three main components. The first is a measurement device, often a headset or cap with sensors to record signals such as electroencephalography. The second is a computer that uses specialised software and algorithms to interpret the brain activity and determine the user's intent or state. The third is an application or device that can be controlled [185]. A Brain-Computer Interface measures neural activity directly, offering data on neural engagement and intent, which is a founda-

tional component of a neuroadaptive environment [185]. The system in this thesis follows the same measure, interpret and adapt structure but replaces neural recording with eye-tracking and peripheral physiology. It is therefore better described as a physiological computing system than as a Brain-Computer Interface in the strict sense.

2.7.1 Adaptive Environments and Human-in-the-Loop Architecture

Adaptive environments represent an approach in human-computer interaction, offering a personalised user experience that evolves dynamically based on individual needs and context [33]. These environments typically comprise three elements, user modelling, detecting changes in state via sensors, and an adaptation engine, the algorithmic core that modifies the interface, content, or difficulty [33].

The domain of adaptive systems in VR has been reviewed by Zahabi et al., who proposed frameworks for training environments, and by Aranha et al., who focused specifically on affective computing [188, 8]. Despite the breadth of this research, many of the existing systems rely on rule-based heuristics that trigger adaptations based on performance metrics, such as error rates or reaction times [9], or on specific user behaviours, such as gaze duration. Drey et al. used gaze tracking to generate adaptive hints in training environments, but this adaptation was triggered by a simple timeout heuristic, looking at an object for too long, rather than a continuous estimation of cognitive state [48]. Performance-based triggers of this kind are typically reactive, modifying the system only after a user has already failed or stalled, rather than proactively managing load. A recent systematic review categorises these performance-based approaches as indirect assessment methods, noting that while they are computationally efficient, they often fail to capture the internal cost of performance [115]. A user may maintain high accuracy while operating at the very edge of their cognitive capacity, a state of high cognitive load that may be unsustainable and prone to sudden collapse [115].

A notable share of adaptive research instead targets affective states or contextual interface optimisation rather than cognitive load specifically. Lindlbauer et al. proposed an optimisation-based approach for mixed reality interfaces that adapts placement based on context and task priority, not the user's mental resources [101], while other recent advances in physiological computing have reported high accuracy in detecting affect using electroencephalography and behavioural data [129, 132]. Affect, valence and arousal, is distinct from cognitive load, however, since an angry or happy user does not necessarily imply an overloaded one. Closer to the domain of workload, Chiossi et al. successfully used Electrodermal Activity to adapt a VR environment, but their adaptation strategy modulated visual complexity, reducing visual clutter, rather than the intrinsic difficulty of the task itself [36]. While reducing extraneous load in this way is beneficial, it does not address the modulation of intrinsic cognitive load required for effective training progression. Simple rule-based physiological loops, such as adjusting a meditative experience based on thresholds in Electrodermal Activity [144], were similarly not designed to handle complex cognitive tasks. This gap lies between systems that react to performance or affect and systems that estimate and respond to intrinsic cognitive load itself, using eye-tracking and physiological data alone. This thesis addresses it in Chapters 5 and 6.

A common component of developing AI-driven systems, particularly those relying on sensitive physiological inputs, is Human-in-the-Loop architecture, in which a human acts as a supervisor involved in the decision-making process through reviewing results, correcting errors, and guiding the model, especially when handling exceptions [143, 116]. This need arises directly from the nature of the predictive models discussed throughout this chapter, since they are largely opaque function approximators, and when trained on noisy, individually variable bio-signals, an erroneous or unrepresentative reading can be classified with high confidence despite being wrong. Given this complexity, Human-in-the-Loop oversight serves as a clinical and ethical safety valve, allowing a human supervisor to verify edge cases or unusual physiological readings and preventing the automated system from making maladaptive training adjustments based on misclassified or noisy sensor data [143]. In a high-stakes example, such as a clinician-supervised cognitive rehabilitation programme for a stroke or dementia patient, an unchecked misclassification of

cognitive load could push a vulnerable patient into a distressing or even harmful level of task difficulty, making human oversight of the model's adaptations advisable. The system evaluated in this thesis, by contrast, operates in a comparatively low-stakes setting, and Human-in-the-Loop oversight is therefore not implemented directly within the adaptive loop. Instead, the model's predictions are evaluated against participants' own self-driven difficulty assessments, examining how closely the model's automated judgments approximate the adjustments a participant would make for themselves, as described in Chapter 6.

2.7.2 Failure Modes in Physiologically-Driven Adaptive Systems

The fully realised adaptive system envisaged throughout this chapter functions as an automated dynamic mechanism, continuously modulating the difficulty and support offered within the VR environment. This mechanism is consistent with Vygotsky's emphasis on the dynamic assessment of a learner's potential [175]. However, because such a mechanism measures and reacts to subconscious physiological and emotional states in real time, it is also prone to error, particularly since many of the states discussed above are to some degree subjective. For home-based interventions, where a human supervisor is not continuously present, the system would instead benefit from a safety fallback. This would keep its adaptations clinically appropriate, especially when dealing with the highly individualised trajectories of cognitive decline [139]. The automation would then continue to serve the user's documented therapeutic needs rather than optimising purely for an algorithmic metric.

2.8 Designing Closed-Loop Adaptive Cognitive Training Environments

Combined, the concepts introduced in this chapter, cognitive training, sensors capturing physiological data, the algorithms making sense of that data, and the theory behind closed-loop adaptive environments, make possible a closed-loop system designed to provide personalised cognitive training. This section outlines the general architecture of such an environment before the specific architecture implemented and evaluated in this thesis is presented in full detail in Chapters 5 and 6.

The core objective of a closed-loop adaptive architecture is to dynamically modulate task challenge within the VR environment to keep the user within the Zone of Proximal Development and the Flow channel. The system operates in a closed loop. Physiological sensors and neural or oculomotor measurement tools continuously record the user's cognitive and emotional state [28, 128, 185]. A predictive model analyses the incoming data streams to extract features that approximate cognitive states such as load, arousal, and attention [97]. An adaptation engine compares the identified cognitive state against the optimal target state defined by the Zone of Proximal Development and Flow theories. If the load is judged too high or too low, adaptation is deemed necessary [175, 43]. Finally, the VR platform dynamically adjusts the parameters of the training task, for example by slowing down the pace, reducing visual complexity, or providing automated scaffolding [125, 33].

This loop repeats continuously throughout a training session, rather than executing once. The underlying assumption is that a user's learning outcomes depend on time spent within the optimal zone, so that the more consistently the task is kept within the narrow band where challenge matches skill, the greater the cumulative cognitive benefit is expected to be. The system is therefore designed to continuously steer the user toward a level of performance that is neither too challenging nor too boring, rather than to optimise any single adaptation in isolation. If a predictive model examines the incoming data and

judges it to represent high cognitive load, the system interprets this as the user drifting toward challenge exceeding skill. It then reduces task difficulty to bring the user back into the Zone of Proximal Development, so that mental capacity can be dedicated to *germane load* rather than extraneous stress [161]. Conversely, if the model judges the data to represent low cognitive load alongside faltering sustained attention, the system may interpret this as boredom [107, 43] and increase task complexity to re-engage the user and push them back toward the Zone of Proximal Development threshold.

2.9 Summary: Gaps, Contributions, and Thesis Structure

This chapter has reviewed the theoretical foundations of cognitive training and cognitive load, the physiological and behavioural modalities available for real-time load assessment, the machine learning approaches used to model them, and the evidence base specific to older adults. It then turned to adaptive systems and the architecture that unifies these components. Across this review, five gaps recur and motivate the research questions introduced in Chapter 1.

First, cognitive load is most commonly manipulated through task-stacking, which risks training models to distinguish task structure rather than genuine modulation of cognitive intensity. Second, evidence that models generalise is limited: the field's reliance on pooled-subject cross-validation risks inflating reported accuracy relative to what a genuinely new user would experience, and models are less often examined for how they behave when the conditions of use differ from those of training (dataset shift). Third, the multimodal cognitive load datasets and models reviewed here are almost entirely offline and observational, with no closed-loop, real-time adaptation of the environment based on the user's estimated cognitive state. Where adaptive systems do exist, they typically rely on performance heuristics or target affect and extraneous load rather than intrinsic cognitive load directly. They also rarely compare a model-driven adaptive design against

the participant-driven design it is intended to improve upon. Fourth, the majority of the physiological computing models for cognitive load reviewed here are developed and validated on younger cohorts. Their suitability for older adults, who stand to benefit most from adaptive cognitive training, is therefore comparatively unexamined. This matters because older adults present both shifted physiological baselines and a documented subjective-objective dissociation in self-report. Fifth, subjective workload report, most commonly operationalised through instruments such as the NASA Task Load Index, is widely treated as ground truth for training and validating these models despite well-documented limitations. Its adequacy as a reference for real-time, low-contrast deployment has been comparatively under-examined (contrast being the size of the step between adjacent difficulty levels: 0.5 to 1.5 s in the offline experiments, 0.1 to 0.3 s in the real-time system).

Chapter 3 addresses the first gap and the across-people generalisability part of the second. It develops and validates a cognitive load estimation model on a novel, within-task VR paradigm designed to be representative of everyday tasks, evaluated under leave-one-participant-out and pooled cross-validation. Chapter 4 addresses the fourth gap, extending this model to an older adult cohort and examining the baseline-physiology and self-report considerations discussed in Section 2.6 directly. Chapters 5 and 6 address the third gap. They build a model architecturally suited to real-time use and deploy it within a real-time, closed-loop adaptive system that compares model-driven against participant-driven difficulty control across single-task and dual-task designs. Chapter 7 addresses the fifth gap and the across-context generalisability part of the second, examining whether the correspondence between subjective difficulty judgments and objective, physiologically-indexed load holds once the task environment moves from the discrete, high-contrast conditions used to train the model to the continuous, low-contrast conditions of real-time deployment.

Chapter 3

Development of a Novel Physiologically-Driven Cognitive Load Model in VR

3.1 Introduction

This chapter addresses RQ1, introduced in Chapter 1: whether eye-tracking and physiological data fusion is a suitable proxy for predicting cognitive load in a dynamic, single-task VR setting, where load changes are subtler than those typically induced by task-stacking designs. Chapter 2 established two points. Eye-tracking is, in principle, the most promising non-EEG modality for this purpose, and Leave-One-Participant-Out (LOPO) validation is a suitable protocol for testing this credibly. Neither claim had yet been tested directly in an interactive, single-task VR paradigm. This chapter closes that gap.

To do so, a novel sustained-attention task was designed explicitly for VR, built around a single, continuously present source of cognitive load rather than the task-stacking manipulations that dominate the existing literature. This design choice is deliberate: task-stacking is useful for producing clearly separable difficulty levels under laboratory conditions, but it sits at odds with the more subtle, continuous fluctuations in demand that a real-time adaptive system would actually need to detect. The task is also argued to be more representative of everyday tasks than an established alternative such as the n-back [165], and its gamified format, together with the absence of task-stacking, gives it a

clearer route to applied or commercial deployment. Framed this way, the task serves as a deliberately conservative testbed: if eye-tracking and physiological fusion cannot predict subjective load here, the case for pursuing it in more complex or naturalistic VR settings would be considerably weaker.

3.2 Key contributions

This chapter makes three contributions. First, it introduces a novel, single-task, VR-native sustained-attention paradigm that avoids the task-stacking manipulation common in prior work. The task is fully open-sourced for reproducibility. Second, it integrates eye-tracking and peripheral physiological sensing into a consumer VR headset to establish the feasibility of real-time, personalised biofeedback for cognitive training. This extends prior work largely confined to desktop or mixed-reality settings [178]. It also builds on VR's established capacity to increase the immersiveness of cognitive training tasks [7]. Third, it directly compares Pooled Cross-Validation and Leave-One-Participant-Out protocols on this task, using machine learning rather than post-hoc ANOVA [79] to support the eventual development of real-time, personalised training.

3.3 Methodology

3.3.1 Experimental Design Overview

This chapter tests RQ1/H1 (Chapter 1): whether eye-tracking and physiological data fusion predicts subjective cognitive load above chance in a single-task VR setting, under Leave-One-Participant-Out (LOPO) validation. The design is within-subjects and repeated-measures: every participant experienced all four difficulty levels, each level twice per session, across three repeated sessions.

Independent variable. Objective task difficulty, manipulated in four levels via the inter-cube spawn interval (1.6s, 2.1s, 2.6s, 3.1s; lower values indicate greater difficulty), presented in a randomised order within each session.

Dependent variable (model target). Subjective difficulty rating, a three-level categorical judgment (*Too Easy*, *Just Right*, *Too Difficult*) reported by the participant via the difficulty prompt described in the Measure of Cognitive Workload subsection below, after each 60-second sub-session.

Measures used to predict the dependent variable. Eye-tracking features (fixation count and duration mean/SD, saccade count/amplitude/velocity and acceleration profile, pupil diameter and its dynamics) and peripheral physiological features (GSR mean/SD/slope; heart-rate-derived HRV metrics: BPM, SDNN, RMSSD, PNN50), aggregated per sub-session (full feature list in Section 3.3.6 below). Two validation protocols, Pooled Cross-Validation (PCVP) and LOPO, were compared as an analysis-level factor to assess how much of any reported accuracy depends on the model having seen a given participant's data during training. The NASA-TLX was collected after each session to compare workload between conditions and populations, and is not the model target. Its Mental and Physical Demand subscales were used here as a manipulation check, to confirm the task was predominantly mentally, rather than physically, demanding, independent of the classification analysis.

The Results section below follows this structure. Classification performance against the subjective difficulty target is reported first, separately for each validation protocol and feature set (eye-tracking only, physiological only, combined). Feature importance (SHAP) is then examined to indicate which of the measures listed above contributed most to that prediction. The NASA-TLX manipulation check is reported last, to confirm that the task was predominantly mentally demanding rather than physically demanding.

3.3.2 Participants

A sample of 62 participants aged 18-58 years ($M = 29.94, SD = 8.91$, 30 female, 6 left-handed) was recruited for the experiment. Due to a technical error, the data of 4 participants were not stored and a further 4 participants were missing large portions of the physiological data due to sensor failure. All these were removed from the analyses, thus leaving a final sample size of 54 participants aged 18-58 years ($M = 29.82, SD = 9.5$, 27 female, 6 left-handed). After each of the three sessions, participants completed a NASA-TLX questionnaire [66] (6 subscales per administration; $58 \times 3 = 174$ administrations as delivered). This uses the 58 participants with usable session data, rather than the further-reduced $N = 54$ used in the classification analyses below. The additional 4 exclusions there were for missing physiological data and did not affect NASA-TLX completeness. At the end of the experiment participants also completed a Simulation Sickness Questionnaire (SSQ) [84] once each ($N = 62$ administrations, i.e. the full recruited sample, since the SSQ was not itself a basis for any of the exclusions described above). Ethical approval was granted by the University of Glasgow, College of Science and Engineering, School of Computing Science (application number 300230006). Out of all 62 participants, 19 reported moderate symptoms on the SSQ. Those included eye strain (8), fatigue (7), difficulty focusing (8), blurred vision (2), sweating (2), nausea (2), stomach awareness (1), dizziness (3), vertigo (1). No participant reported symptoms in the high-discomfort range that would have warranted exclusion, and no participants were excluded on the basis of SSQ responses; these figures are reported descriptively to characterise the sample's experience of the VR task, separately from the technical/data-quality exclusions described above.

3.3.3 Task

Consistent with the single, continuous source of load established above, difficulty here was manipulated within one ongoing task rather than by layering a second, concurrent one. The experiment incorporated a novel sustained attention VR task, derived from continuous performance tasks [38] in which participants were required to respond or withhold

responses to stimuli across periods of 8 minutes of active gameplay (about 9 minutes including the 10-second intervals between sub-sessions). Within the task, blue and yellow cubes appeared at random in the space before the participant. The participant held a controller in each hand, one projecting a blue pointer and the other a yellow pointer. The objective of the task was to eliminate the cubes materialising in front of the participant by employing the corresponding pointer (the blue pointer exclusively destroyed blue cubes, and the yellow pointer exclusively destroyed yellow cubes) (see Figure 3.1 for reference). Four objective difficulty levels were established by regulating the intervals at which cubes appeared (with lower values indicating greater difficulty, as more cubes spawned within each interval) – 1.6s, 2.1s, 2.6s, 3.1s. As an example, a 2.1s difficulty meant a 2.1s interval between two blue cubes spawning and the same 2.1s between two yellow cubes. However, there is an offset between a blue and yellow cube appearing so that they do not appear simultaneously. The difficulty levels were determined through a pilot study involving 6 participants. To exemplify, at the most challenging difficulty level, a subsequent cube of the same color spawned 1.6s after the appearance of a blue or yellow cube. If cubes remained on the screen without being destroyed, the likelihood of another cube spawning in the same location increased, which led to both cubes starting to move across the screen in a motion that was both translational and rotational, further escalating the difficulty. This escalation raises intrinsic task load directly, by increasing the number of cubes competing for attention within the same continuous task, rather than through an added distractor stream, consistent with the capacity/selectivity distinction introduced in Chapter 2.

The experiment consisted of three sessions, interspersed with 5-minute breaks. The three sessions were identical as the aim was to increase the amount of data collected. During the breaks, participants completed a NASA-TLX questionnaire to evaluate their subjective cognitive load pertaining to the prior session. Within each session, every difficulty level was iterated twice, resulting in eight 60-second sub-sessions, separated by 10-second intervals. The sequence of difficulty levels was randomized for each session.

The pivotal aspect of this design lies in its capability to adjust difficulty over brief intervals and solicit participant feedback without taking off the headset or the completion of an extensive questionnaire, thereby avoiding disruption to the experimental flow. This aspect is of particular importance for the potential real-time adaptation of this design.

This task, the VR application it runs in, and the accompanying data collection pipeline were designed and developed by the author specifically for this thesis. The full experimental software, along with the real-time infrastructure described in later chapters, is made publicly available on GitHub (https://github.com/EyeDAPT), and the resulting multimodal dataset has been deposited on the Open Science Framework: Szczepaniak, D., Harvey, M., and Deligianni, F. (2026, March 16). The EyeDAPT Dataset: Multimodal Eye-Tracking and Physiology for Real-Time Cognitive Load Adaptation in VR. https://doi.org/10.17605/OSF.IO/VJSNF.

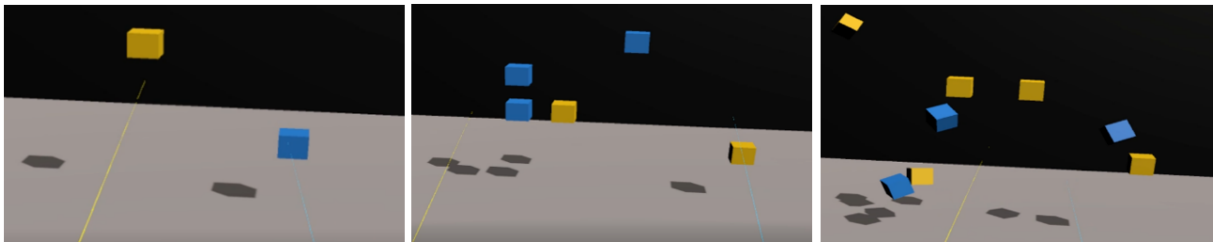


Figure 3.1: Example views of the experiment progressing in difficulty from an easier subsession on the left (only one cube of each colour present at a time) to a more difficult one on the right (several cubes, some moving).

3.3.4 Measure of Cognitive Workload

This subsection makes explicit how the dependent variable defined above, the subjective difficulty rating, was actually collected. Feedback provided by participants after every level served as this study's target label for cognitive workload. It captured how difficult each participant experienced each sub-session in the game. This perceived difficulty was the feedback provided verbally by participants after every sub-session (60 seconds), im-

mediately following it and before the next sub-session began, so that the reported label corresponds to that specific sub-session rather than the session as a whole. Participants were instructed to provide feedback according to the following guidelines with regards to mental demand:

- ***Too Easy*** - 'You don't find the task challenging and don't need to focus on it to perform well. You begin to disengage from the task and pay more attention to distracting thoughts, starting to feel boredom.'
- ***Just Right*** - 'You find the task moderately challenging and need to keep concentrating on it to perform well. You find the task enjoyable and engaging.'
- ***Too Difficult*** - 'You feel that the task is impossible to be completed at that level of difficulty and think it is outside of your reach. You begin to feel frustrated and would enjoy it more if it was easier.'

This three-level judgment, rather than the objective difficulty level that produced it, is the label predicted by all models reported in this chapter.

As Chapter 2 argued, no measure of cognitive load is an uncontested ground truth, so this choice needs a reason. Perceived difficulty was used as the target because it reflects how the task was experienced by the person doing it. It can also be collected after every 60-second sub-session, which a questionnaire completed at the end of a session cannot do at the frequency an adaptive system needs. Its drawbacks are that it remains a subjective judgment, open to anchoring and to differences in how individuals use the scale. The NASA-TLX was therefore collected after each session as a separate measure for comparing conditions and not as the label, and how far the perceived-difficulty label can be trusted is examined again in Chapters 4 and 7.

3.3.5 Materials

The experiment was conducted using an HTC VIVE PRO EYE headset with an integrated eye-tracker connected to a desktop with NVIDIA RTX 2070 SUPER graphics card, AMD Ryzen 5 3600 processor and 16GB of RAM. The headset collected raw eye-tracking metrics at 90 Hz using a free Unity plugin. These included eye openness, pupil diameter, sensor position, gaze origin and direction, and convergence distance. Physiological data included photoplethysmography (PPG) and galvanic skin response (GSR). It was collected using a wrist-based Shimmer sensor at 51Hz and processed using the pyshimmer [109] library in Python. A data collection pipeline was established to timestamp-sychronise eye-tracking data, physiological data, and in-task events using LabStreamingLayer [157]. After each session participants completed a NASA-TLX questionnaire [66].

Eye-tracking Frequency Justification

Eye-trackers differ widely in sampling rate, and the difference matters mainly for how finely fast eye movements are resolved. A research-grade tracker sampling at 1000 Hz records a new gaze sample roughly every millisecond [15], whereas a headset-integrated tracker such as the one in the HTC VIVE PRO EYE samples at 90 Hz, roughly every 11 ms. For slow events the two behave alike. A fixation lasts hundreds of milliseconds, so even at 90 Hz it is covered by around twenty or more samples. A comparison of trackers with different sampling rates found similar numbers of detected fixations, with fixation durations differing at the lowest rate tested, 60 Hz [98]. Saccades, the ballistic eye movements between fixations, are the difficult case. They are brief (roughly 20 to 50 ms) and can reach peak velocities of several hundred degrees per second. At 1000 Hz a saccade is covered by tens of samples and its velocity profile is traced closely, whereas at 90 Hz only about two to four samples fall within a single saccade. Detection of saccades and estimation of their duration improve as the sampling rate rises from 60 to 120 Hz [98]. Analyses of high-rate recordings indicate that frequency components up to roughly 75 Hz carry the signal needed to preserve saccades [134]. This is above the 45 Hz Nyquist

frequency of a 90 Hz tracker. It follows that peak velocity and acceleration measured at 90 Hz are likely to be biased downward: the peak lasts only a few milliseconds, velocity estimated from consecutive samples is an average over the sampling interval, and the samples are unlikely to land on the peak. The direction of this bias is therefore systematic, while its size varies from one saccade to the next. Averaging over more data reduces the random temporal uncertainty introduced by a low sampling rate [5], but it cannot remove a systematic downward offset. Because the same tracker was used for every participant and condition in this thesis, the offset is shared. Comparisons between conditions and groups therefore remain meaningful, although absolute velocity values should not be compared with those obtained from higher-rate trackers. This thesis therefore relies on features aggregated across a window or session, such as fixation duration variability and mean saccade amplitude. It uses saccadic velocity only as an aggregate across the saccades within a window, without interpreting the peak velocity of any individual saccade. The limits of this choice are discussed in Chapter 8, and higher-rate hardware is listed as future work in Chapter 9.

3.3.6 Eye-tracking and Physiological Signal Processing

Eye-tracking Feature Extraction Pipeline

Eye-tracking data was processed using a comprehensive pipeline designed to extract robust gaze behaviour and pupil dynamics metrics. Raw eye-tracking data was collected at 90 Hz from the integrated HTC Vive Pro Eye tracker and processed through several stages following established protocols for VR-based cognitive load assessment [95].

Preprocessing and Noise Reduction: Raw gaze direction vectors were computed by averaging left and right eye 3D gaze directions and normalizing to unit vectors. Direction signals were smoothed using Savitzky-Golay filtering (window length = 11, polynomial order = 2) to reduce high-frequency noise while preserving important signal characteristics. This approach preserves peak information while filtering signal artifacts, following established practices for noise reduction in eye-tracking applications [134].

Blink Detection and Interpolation: Blinks were detected using adaptive thresholding based on pupil diameter fluctuations, with eye openness values below 70% classified as blinks. Rolling statistics (100ms windows, 9 samples) were used to establish dynamic thresholds that account for individual variations in eye anatomy and blink patterns [120]. Blink segments were extended by 50ms margins on each side to account for partial blinks, and missing data during blinks was interpolated using linear interpolation following the Identification-Artifact Correction (I-AC) protocol.

Velocity and Acceleration Computation: Angular velocity was calculated from consecutive gaze direction samples using the formula: $velocity = \sqrt{(\Delta x)^2 + (\Delta y)^2} \times \text{sampling_rate}$, where Δx and Δy represent frame-to-frame changes in normalized gaze direction. Acceleration was computed as the temporal derivative of velocity, providing input for advanced saccade characterization.

Fixation and Saccade Classification: A Hidden Markov Model (HMM) with two Gaussian states was employed to classify each sample as either fixation (state 0) or saccade (state 1) [86]. The HMM was fitted using velocity profiles with 2D Gaussian emission densities, providing more robust classification than simple velocity thresholding methods. This is a well-established approach: the original Identification-by-HMM (I-HMM) algorithm, which pairs a velocity-based threshold with HMM reclassification, was introduced by Salvucci and Goldberg [145] and remains in wide use, with more recent unsupervised

velocity-based HMM classifiers such as gazeHMM [105] further validating the approach. This approach has been reported to identify saccade patterns better than traditional threshold-based methods [42] and is of particular importance in real-time designs where all the calculations need to be performed during the experiment.

Advanced Saccade Metrics: Beyond basic saccade counting, advanced metrics were extracted including peak velocity, mean acceleration/deceleration phases, velocity variability, and acceleration-deceleration ratios [47]. Peak saccadic velocity was extracted as the highest velocity during individual saccades, with main sequence relationship normalization for amplitude effects. These metrics provide insights into saccade dynamics that correlate with cognitive load and demonstrate sensitivity sufficient to distinguish among cognitive demand levels [15].

Pupil Dynamics Analysis: Pupil diameter was processed to extract both static measures (mean, standard deviation, range) and dynamic measures including linear trend (slope), constriction velocity (rate of pupil size decrease), and dilation velocity (rate of pupil size increase) with a 2nd degree Butterworth filter applied for preprocessing [89].

Physiological Signal Processing

Photoplethysmography (PPG) Processing: PPG signals were collected at 51 Hz using a wrist-mounted Shimmer sensor and processed using the HeartPy library [170]. Signal preprocessing included 2nd-order Butterworth low-pass filtering (cutoff = 3 Hz, order = 3) to minimize PPG waveform distortion [94], followed by peak enhancement using two iterations of linear transformation. Heart rate variability metrics were computed using segmentwise analysis to extract beat-per-minute (BPM), standard deviation of normal-to-normal intervals (SDNN), root mean square of successive differences (RMSSD) and percentage of successive RR intervals differing by more than 50ms (pNN50) [155].

Galvanic Skin Response (GSR) Processing: GSR signals were processed to extract three primary metrics: mean conductance level, conductance variability (standard deviation), and linear trend (slope) [191]. Missing values were handled using forward-fill followed by backward-fill interpolation to maintain signal continuity. Baseline-relative averaging was applied following established protocols for VR stress response measurement.

Temporal Windowing and Feature Aggregation

To capture temporal dynamics while maintaining computational efficiency for real-time processing, data from each 60-second experimental level was divided into four sequential 15-second windows (1,325 raw samples per window at 90 Hz for eye-tracking and 765 samples per window for physiological data at 51Hz) [190]. This windowing approach enables the model to learn temporal patterns in cognitive load changes throughout each task level while providing sufficient temporal resolution for real-time applications [146].

In windowed analyses of physiological signals, the window length affects the estimates obtained and is best matched to the signal of interest [22]. A 15-second window was chosen here as a compromise between temporal resolution and the amount of data available to compute the eye-tracking and heart rate variability features. It yields four windows per 60-second level.

For each window, 21 eye-tracking features and 7 physiological features were extracted, resulting in 28 features per window. The complete feature set included:

Eye-tracking features per window: blink count, blink rate, fixation count, mean/std fixation duration, saccade count, mean/std saccade amplitude, saccade-fixation ratio, advanced saccade metrics (peak velocity, acceleration/deceleration characteristics), and comprehensive pupil measures (mean, std, range, slope, constriction/dilation velocities).

Physiological features per window: GSR metrics (mean, std, slope) and heart rate variability metrics (BPM, SDNN, RMSSD, PNN50).

For each metric, 4 aggregate metrics were computed to represent the metric across each level (4 windows) - mean, standard deviation, slope, range for a total of 112 features.

Feature Selection and Dimensionality Reduction

To optimise model performance and mitigate the curse of dimensionality, a feature selection protocol based on SHAP (SHapley Additive exPlanations) values was applied. SHAP was used for selection because it attributes each prediction to individual features for every class, so that one ranking can be formed across the three classes and the same values reused to interpret the model (see Explainability of the Machine Learning Models below). Features were ranked once by mean absolute SHAP value on the combined younger and older dataset (Section 3.3.6), and the number of top-ranked features retained was set for each input modality as described below.

Justification of Feature Count

Feature selection was performed once, before model evaluation, using SHAP-based importance. It was computed on the combined dataset of younger and older adults ($N = 73$ participants; 1,736 sub-session rows, the dataset assembled across this chapter and Chapter 4), not on the younger sample alone. The aim was a single feature set that is representative of the overall dataset, so that performance can be compared across the younger, older and combined populations without differences in the inputs affecting the comparison. This chapter reports results for the younger adults only, and Chapter 4 reports results for the older adults and the combined population using the same features. The selection is described here because this is the chapter in which the features are first used.

An XGBoost model was fitted to all 112 candidate eye-tracking and physiological features, and the features were ranked by mean absolute SHAP value summed across the three difficulty classes. To decide how many features to retain, the number of top-ranked features was varied from 5 to all 112, with the model hyperparameters held fixed. Each setting was scored by macro-F1 under participant-grouped five-fold cross-validation, repeated ten times with different fold assignments, so no participant contributed data to both a training and a validation fold. As Figure 3.2 shows, performance rises steeply up to about 20 features and peaks at 30 (macro-F1 $= 0.519$). It then stays within 0.02 of that peak up to 85 features, so the results are not sensitive to the exact cutoff. We therefore adopted 30 features, a compact set at the empirical optimum. It comprises 24 eye-tracking and 6 physiological features. The 24 eye-tracking features define the eye-tracking-only subset analysed below, and the physiological-only model retains all 28 physiological features.

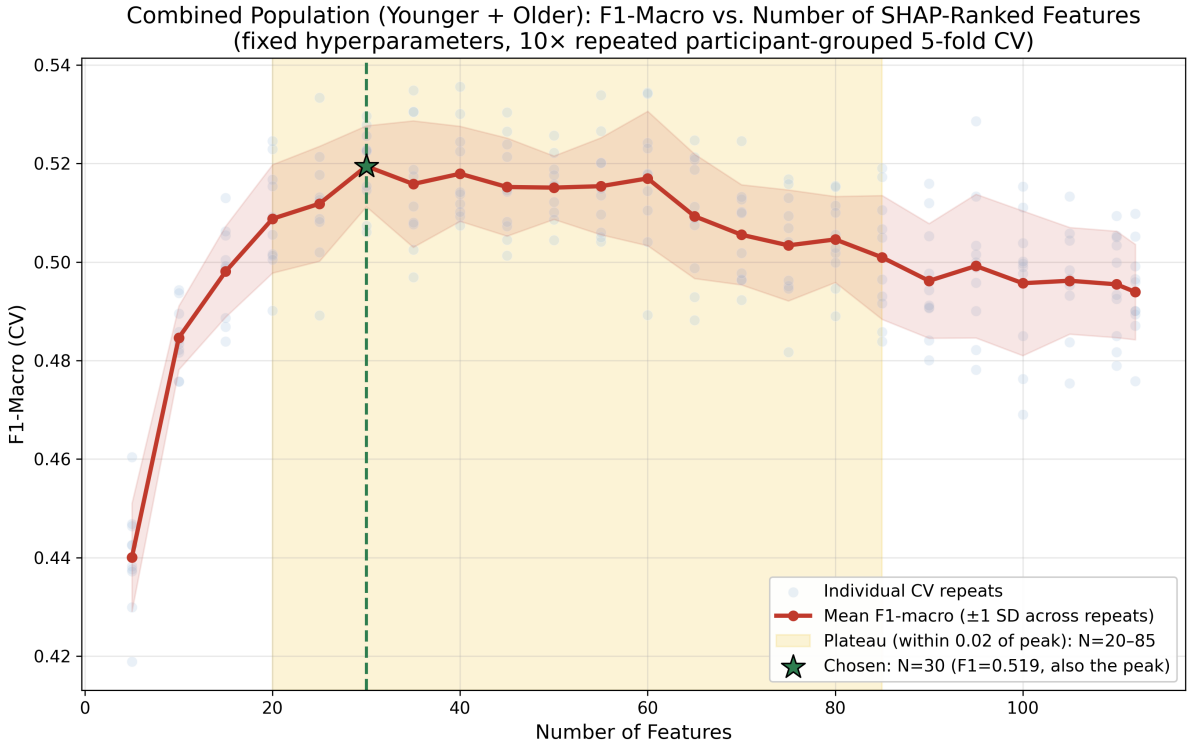


Figure 3.2: Macro-F1 as a function of the number of SHAP-ranked features retained, for the combined younger and older dataset ($N = 73$), with fixed hyperparameters and participant-grouped 5-fold cross-validation repeated ten times. The line shows the mean across repeats, the band ± 1 SD and the points the individual repeats. Performance peaks at $N = 30$ (macro-F1 $= 0.519$) and remains within 0.02 of this peak from $N = 20$ to $N = 85$ (shaded).

Because the ranking was computed on the full combined sample, including participants who are held out in the analyses reported here, we checked that it is unlikely to drive the reported performance. Every model reported in this chapter is fitted only on its training participants and evaluated only on participants withheld from fitting. The only thing carried over from the full sample is the identity of the 30 input columns. Recomputing the ranking on the training partition alone of each cross-validation fold recovered on average 25 of the 30 selected features (83%) and 14 of the top 16. It returned the same top-ranked feature in all five folds. A participant-level bootstrap (100 resamples, drawing whole participants with replacement) kept the 15 highest-ranked features within the top 30 in at least 85% of resamples. The selected set therefore appears to be reasonably stable across train-test partitions. It is held fixed across all models and evaluation schemes in this chapter.

This selection procedure was performed once, on the combined dataset, and the resulting 30-feature list is used unchanged for every model and population in this thesis. This is deliberate. Re-deriving a separate feature count and ranking for each population would make it harder to attribute differences in performance between populations to the populations themselves. One consequence is that the ranking reflects data from all participants, including the older adults who do not appear in this chapter's results. Only the identity of the input columns is carried over, and every model is fit on its training participants alone, so the stability checks above are a relevant safeguard.

3.3.7 Machine Learning Models for Cognitive Workload Prediction

Training set composition. All models in this chapter were trained on the younger-adult dataset only ($N = 54$ participants after exclusions). Aggregated at the level of one row per 60-second sub-session, this yields 1,280 samples drawn from 5,216 raw 15-second task-level windows (Section 3.3.6). Candidate predictors were drawn from a pool of 112 engineered

eye-tracking and physiological columns, ranked on the combined dataset (Section 3.3.6). The final Combined model used the top 30 of these, and the Eye-Tracking-only model used the top 24. The Physiological-only model retained all 28 physiological columns, since preliminary analysis found no redundant physiological features to prune.

Three classifiers were trained to predict each of the target variables – Support Vector Machines, Random Forest and XGBoost. For every analysis, a separate model was trained on eye-tracking data only, physiological data only (heart rate and galvanic skin response), and both combined.

3.3.8 Explainability of the Machine Learning Models

SHAP (SHapley Additive exPlanations) is a versatile approach used to interpret the results of machine learning models. It calculates the contribution of each feature to the prediction of a specific instance, regardless of the model used. This method leverages Shapley values from cooperative game theory to distribute the overall value generated by a group of players fairly among individual players. By assigning SHAP values to each feature, one can effectively rank the features based on their importance in the prediction process [106]. As opposed to traditional explainability approaches which estimate feature importance by shuffling one feature at a time, SHAP is relatively resilient to correlated variables.

3.4 Results

Results are reported against the independent and dependent variables set out in the Experimental Design Overview above. The manipulated objective difficulty (IV) is examined first against the subjective-difficulty target (DV) to check that the manipulation behaved as intended. The chapter then turns to how well the eye-tracking and physiological measures predict that target under each validation protocol.

3.4.1 Subjective Difficulty Distribution

Table 3.1 presents the relationship between the objective difficulty levels (1–4) and the participants' subjective perception (*Too Easy*, *Just Right*, *Too Difficult*). Consistent with the intended design, perceived difficulty generally aligned with objective difficulty. However, individual differences were evident; some participants rated the most difficult sessions as *Just Right*, while others found the easiest sessions *Too Difficult*. This discrepancy supports training models on subjective labels rather than objective difficulty levels.

Table 3.1: Distribution of perceived difficulty ratings across all sessions. Counts come from the 58 participants with usable session data (1,384 ratings); the classification analyses use the 54 participants with complete physiological data (1,280 sub-sessions).

Objective Level	**Too Easy**	**Just Right**	**Too Difficult**
Level 1 (Easiest)	287	55	4
Level 2	194	141	11
Level 3	90	204	52
Level 4 (Hardest)	5	141	200
Total Frequency	**41.6%**	**39.1%**	**19.3%**

3.4.2 Classification Performance

We evaluated the predictive models using two distinct validation protocols to contextualize our findings within the broader literature and assess real-world viability.

Pooled Cross-Validation Protocol (PCVP)

First, we employed a stratified 5-fold cross-validation approach, pooling data from all participants. This method, widely utilised in physiological computing literature, typically establishes an optimistic performance baseline as the model is exposed to data from the test subjects during training.

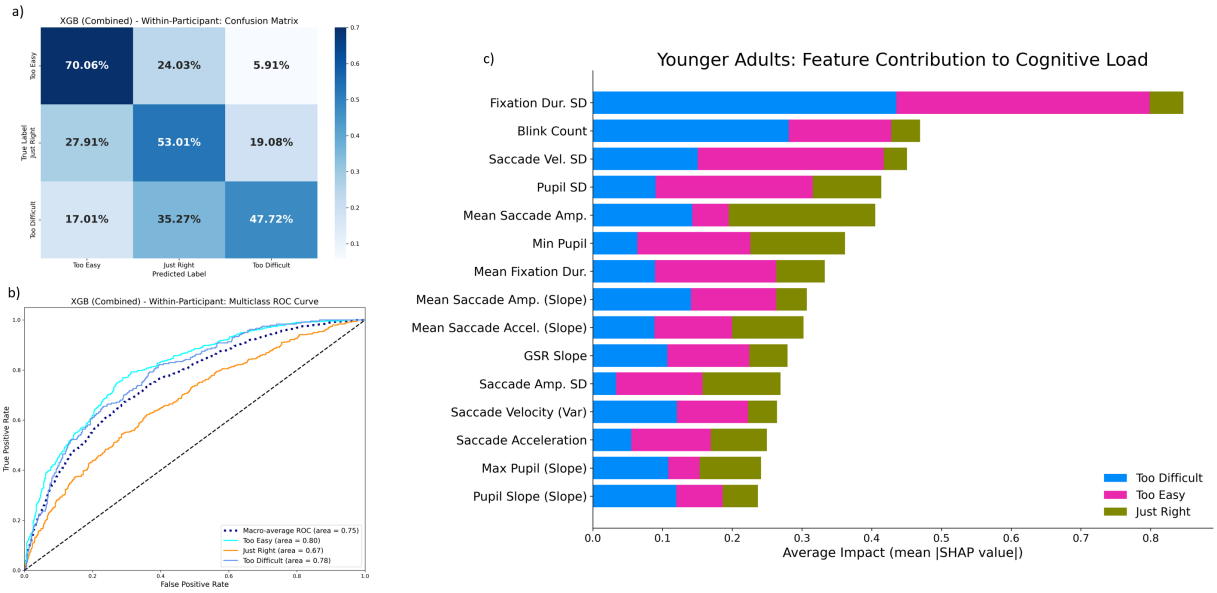


Figure 3.3: Performance and interpretability analysis of the XGBoost model (Combined Feature Set) under the Within-Participant (Pooled Cross-Validation) protocol for the younger population. (a) Confusion Matrix: The model shows the highest sensitivity for the Too Easy class (70.06%), with Just Right (53.01%) and Too Difficult (47.72%) more frequently confused with one another. (b) Multiclass ROC Curve: The Macro-average AUC of 0.75 indicates above-chance classification performance, with the Too Easy class the most discriminable (AUC = 0.80). (c) SHAP Feature Importance: Oculometric features dominate the ranking, led by Fixation Duration SD, Blink Count and Saccade Velocity SD, while physiological features such as GSR Slope contribute comparatively little.

Table 3.2 presents these results. The Combined XGBoost model achieved the highest performance (Accuracy: 0.59, F1-Score: 0.57). Consistent with literature trends, this protocol yielded higher accuracy and F1-scores than the leave-one-participant-out approach (Table 3.3). Notably, even in this favorable training scenario, the addition of physiological data provided only a marginal gain over Eye-Tracking alone (Combined F1: 0.57 vs. Eye-Tracking F1: 0.56), suggesting that oculometric features are the dominant signal for this task. The confusion matrix (Figure 3.3a) shows the same pattern seen under LOPO, with *Too Easy* sub-sessions identified most reliably and *Just Right* and *Too Difficult* more often confused with each other. The SHAP analysis (Figure 3.3c) likewise mirrors the LOPO ranking, with oculometric features driving the prediction and physiological metrics contributing only marginally, consistent with the modest gain from adding physiological data noted above.

Table 3.2: Pooled Cross-Validation Results (PCVP). This protocol, common in the literature, generally yields higher accuracies due to participant data leakage.

Feature Set	Model	Accuracy	Precision	Recall	F1-Score	AUC
Combined	**XGBoost**	**.59**	**.57**	**.57**	**.57**	**.75**
Combined	Random Forest	.57	.55	.54	.54	.74
Combined	SVM	.57	.54	.55	.54	.73
Eye-Tracking	XGBoost	.58	.56	.56	.56	.74
Eye-Tracking	Random Forest	.58	.56	.54	.54	.73
Eye-Tracking	SVM	.55	.53	.54	.53	.73
Physiological	XGBoost	.41	.41	.42	.40	.57
Physiological	Random Forest	.44	.42	.42	.42	.58
Physiological	SVM	.42	.41	.42	.41	.57

Leave-One-Participant-Out (LOPO)

To test generalisability to unseen users, we employed a Leave-One-Participant-Out (LOPO) protocol. In this setup, the model is trained on $N-1$ participants and tested on the remaining participant, simulating a "plug-and-play" scenario without prior calibration. Performance dropped compared to the pooled baseline for all feature sets. The drop was clearest in macro-F1, which fell by 0.06 to 0.11 across models (for the Combined XGBoost model, from 0.57 to 0.49), against 0.02 to 0.06 in accuracy and up to 0.06 in AUC. This

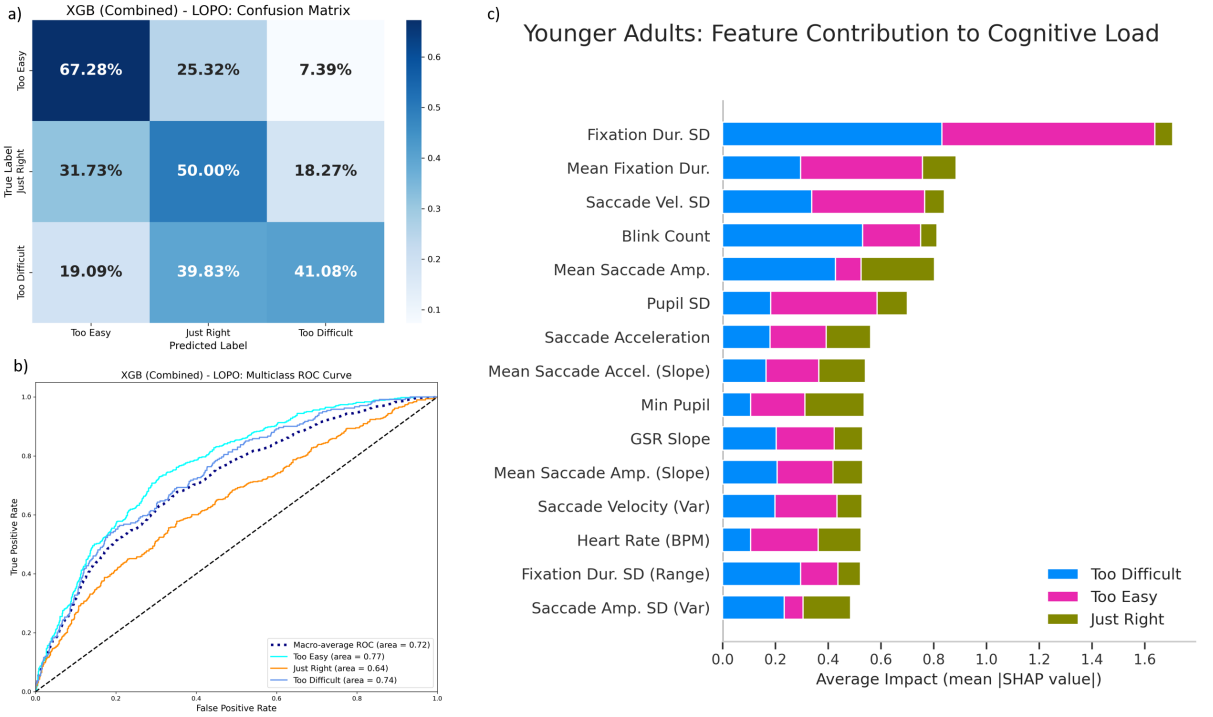


Figure 3.4: Performance and interpretability analysis of the XGBoost model (Combined Feature Set) under the Leave-One-Participant-Out (LOPO) protocol for the younger population. (a) Confusion Matrix: The model demonstrates high sensitivity for the Too Easy class (67.28%) but shows confusion between Just Right and Too Difficult, reflecting the subjective overlap between these states. (b) Multiclass ROC Curve: The Macro-average AUC of 0.72 indicates classification performance above chance level, with the 'Too Easy' class achieving the highest discriminative power (AUC = 0.77). (c) SHAP Feature Importance: Eye-tracking features, particularly Fixation Duration Standard Deviation (SD) and Mean Fixation Duration, emerge as the top predictors. Physiological features like Heart Rate (BPM) and GSR Slope appear lower in the ranking, suggesting a secondary role in this population.

is a reasonably substantial gap, in line with the drop reported in the literature when models are tested on participants they have not seen [25, 113]. Physiological-only models, however, started from a much lower level and remained close to chance, consistent with the greater inter-subject variability of that modality.

Table 3.3 details the performance across feature sets (Mean ± SD). The Combined XGBoost model achieved the highest overall performance, with an Accuracy of 0.56 ± 0.15 and an AUC of 0.72 ± 0.13. The Eye-Tracking models performed similarly, with the Random Forest model trained solely on oculometric features achieving an F1-score of 0.48, close to the best Combined model (0.49). This suggests that eye-movement dynamics are the primary drivers of generalisation in this task. Models trained solely on physiological data suffered a marked performance drop compared to eye-tracking. The best physiological model (Random Forest) achieved an F1-score of 0.34 and an AUC of 0.55, which is only marginally above the random chance baseline. The confusion matrix (Figure 3.4a) reveals that the model is particularly effective at identifying low cognitive load (*Too Easy*), achieving a recall of 67.28%. However, distinguishing between optimal load (*Just Right*) and overload (*Too Difficult*) proved more challenging, likely due to the fine granularity of subjective reporting in these states. The SHAP analysis (Figure 3.4c) corroborates the findings from the modality comparison, identifying oculometric features such as *Fixation Duration SD* as the primary drivers of the model's decision-making process, while physiological metrics contributed minimally to the prediction.

Table 3.3: Leave-One-Participant-Out (LOPO) Validation Results (Mean (SD)). This protocol tests generalizability to unseen users. Note that even with the full set of features, Physiological models struggle to generalize compared to Eye-Tracking.

Feature Set	Model	Accuracy	Precision	Recall	F1-Score	AUC
Combined	**XGBoost**	**0.56 (0.15)**	**0.53 (0.15)**	**0.53 (0.16)**	**0.49 (0.14)**	**0.72 (0.13)**
Combined	SVM	0.52 (0.16)	0.48 (0.16)	0.49 (0.16)	0.45 (0.16)	0.71 (0.14)
Combined	Random Forest	0.53 (0.16)	0.50 (0.17)	0.50 (0.17)	0.46 (0.16)	0.71 (0.13)
Eye-Tracking	XGBoost	0.52 (0.15)	0.48 (0.16)	0.49 (0.15)	0.45 (0.15)	0.71 (0.12)
Eye-Tracking	SVM	0.51 (0.14)	0.47 (0.16)	0.48 (0.14)	0.44 (0.14)	0.71 (0.13)
Eye-Tracking	**Random Forest**	**0.54 (0.16)**	**0.51 (0.17)**	**0.52 (0.15)**	**0.48 (0.17)**	**0.72 (0.13)**
Physiological (All)	Random Forest	0.41 (0.12)	0.39 (0.15)	0.38 (0.12)	0.34 (0.12)	0.55 (0.11)
Physiological (All)	XGBoost	0.39 (0.15)	0.37 (0.15)	0.38 (0.15)	0.34 (0.14)	0.55 (0.11)
Physiological (All)	SVM	0.38 (0.12)	0.34 (0.11)	0.32 (0.10)	0.30 (0.10)	0.51 (0.10)

3.4.3 Physical versus Mental Demand

To check that the task was predominantly mentally rather than physically demanding, we analyzed the NASA-TLX questionnaires completed after each 9-minute session. The mean physical demand score was 7.38 (SD=4.60) compared to a mean mental demand score of 10.23 (SD=4.54). A Wilcoxon signed-rank test confirmed a statistically significant difference (W=1133,p<0.001), indicating that the task was predominantly mentally demanding. Full NASA-TLX descriptive statistics for all subscales, including those not analysed here, are reported in Appendix 10, Section 10.1.

We computed Pearson correlations separately for performance against physical and mental demand. Physical demand showed no correlation with performance (r=.007,p=.92), whereas mental demand showed a stronger, approaching significance trend (r=.14,p=.06). This is consistent with the task being predominantly mentally demanding, although these analyses do not show what the physiological features themselves captured.

3.5 Discussion

The aim of the study was to validate automatic cognitive workload detection in a Virtual Reality setting using both eye-tracking and physiological data for future use in real-time adjustable cognitive training. In order to achieve that, the task has been adapted with the potential for dynamic modulation, and is designed to be closer to a home-based administration of the task than to a highly controlled laboratory environment with an unwieldy eye-tracker. Our results demonstrate that a gradient boosting model (XGBoost) trained on a combined feature set of eye-tracking and physiological data can classify subjective cognitive load with an AUC of 0.72 in a strict Leave-One-Participant-Out (LOPO) setting. However, the analysis reveals a distinct hierarchy in feature modality performance, with oculometric metrics driving predictive power and physiological signals struggling to generalise when applied to unseen users.

3.5.1 The Dominance of Oculometric Features in VR

Consistent with recent findings in the literature, our results identify eye-tracking as the most robust single modality for cognitive load estimation in HMD-based environments [12, 64]. The Random Forest model trained solely on eye-tracking features achieved performance similar to the combined multimodal model (F1-score 0.48 vs. 0.49).

The SHAP analysis provides granular insight into this dominance, identifying *Fixation Duration Standard Deviation* and *Mean Fixation Duration* as the top predictors. This aligns with cognitive load theory, which suggests that as task difficulty increases, gaze behaviour becomes more fixated and less exploratory to optimize visual processing [158]. Furthermore, the robust performance of saccadic metrics supports the work of Lopez et al. [104], who reported a link between saccade velocity profiles and mental effort. The success of these metrics in a LOPO setting suggests that oculomotor responses to load share characteristics across individuals, making them promising candidates for plug-and-play VR systems that do not require extensive user calibration.

3.5.2 The Generalization Gap in Physiological Sensing

A notable finding of this study concerns the comparatively weak generalisation of physiological signals relative to eye-tracking. Physiological features stayed close to chance under both protocols. The best physiological model (Random Forest) reached an F1-score of 0.42 (AUC 0.58) under pooled cross-validation and 0.34 (AUC 0.55) under LOPO. The gap between modalities is the more informative one: even under LOPO, where the eye-tracking and combined models retained an AUC of around 0.72, physiological-only models remained close to chance (AUC 0.55). This indicates that physiological signals contribute little unique, generalisable information regardless of validation protocol.

This discrepancy highlights a fundamental challenge in physiological computing: the high inter-subject variability of autonomic responses [25, 15]. Unlike eye movements, which are mechanically constrained, heart rate variability (HRV) and galvanic skin response (GSR) are influenced by a multitude of individual factors including baseline fitness, circadian rhythms, and emotional lability. The failure of the physiological-only feature set to improve LOPO performance suggests that per-user calibration may matter more than further feature selection. While large-scale studies like the HP Omnicept dataset [178] have achieved success with physiology without calibration, they utilised a dataset an order of magnitude larger (N=738), potentially allowing the model to learn a wider distribution of physiological baselines. For smaller cohorts typical of VR research (N=54), our results suggest that physiological signals may add little in user-independent models unless individual baselines are normalized via a calibration phase.

3.5.3 Methodological Implications of Validation Protocols

The contrast between our Pooled (Subject-Dependent) and LOPO (Subject-Independent) results is reasonably substantial when measured by macro-F1. For the combined model, F1 falls from 0.57 to 0.49, and for the XGBoost eye-tracking model from 0.56 to 0.45. Accuracy and AUC fall by less (from 0.59 to 0.56 and from 0.75 to 0.72 for the combined model), so they understate the gap. This is the direction and rough size of gap that the literature leads us to expect when moving from subject-dependent to subject-independent validation [25, 113]. The above-chance result holds under both protocols, but the LOPO figures are the more realistic estimate of what a new user would experience, which is why they are used as the headline throughout. The contrast is also informative for physiological signals specifically. Pooled cross-validation is common in much of the affective computing literature [151]. Here it would still have supported only a mediocre estimate of what physiological features can do, since even the pooled F1-score for physiological-only models (0.42) was already low.

By employing LOPO, we provide a more realistic estimate of how such a system would perform when a new user puts on the headset for the first time. Performance above chance under LOPO suggests that oculomotor markers of cognitive load generalise reasonably well across individuals even without user-specific calibration, although less well than pooled validation would suggest. Physiological-only models remained close to chance under both protocols, consistent with reports elsewhere in the literature that physiological signals are more idiosyncratic and calibration-dependent than eye-tracking [25, 178].

3.6 Summary

This chapter set out to test RQ1/H1: whether an eye-tracking and physiological fusion model, trained on a single-task VR paradigm without task-stacking, could predict subjective cognitive load above chance under LOPO validation. It could, with an AUC of 0.72, although the result is carried mainly by eye-tracking once the two modalities are considered separately. Eye-tracking carried nearly all of the predictive signal on its own. Physiological features added little once genuine inter-subject generalisation was required. The fusion model's success is therefore best understood as an eye-tracking result. It does not show that physiology and eye-tracking combine to outperform either alone. This also has a bearing on the methodological point made in Chapter 2. The gap between pooled and LOPO validation is often treated in the literature as evidence that pooled validation is optimistic. The gap here is reasonably substantial in macro-F1 (0.57 to 0.49 for the combined model), in line with that expectation. The choice of modality still does most of the work in determining what a model can achieve, and the choice of validation protocol changes the headline figure enough to matter.

Two open questions follow directly from this baseline, and each is taken up in a subsequent chapter. A third limitation, noted at the end of this section, is only partly addressed in this thesis.

3.6.1 Inclusion of an Older Participant Cohort

Chapter 4 takes up the first of these directly. The current study utilized a sample of younger adults (M=29.82 years). However, cognitive training interventions in VR are frequently targeted at older adults to mitigate age-related cognitive decline [7]. Ageing is widely reported to affect both the autonomic nervous system and the oculomotor system. Older adults typically exhibit reduced heart rate variability, slower saccadic velocities and prolonged fixation durations even at baseline levels of load.

Consequently, a model trained exclusively on a younger population is unlikely to generalise to an older cohort without adjustment. Chapter 4 replicates the present experimental design with participants aged 60 and over. It assesses whether the features identified here, such as fixation duration SD, remain predictive in older adults or whether the SHAP feature importance shifts. It also assesses whether a model trained on both populations offers a more suitable generalised solution than specialised, population-specific models.

3.6.2 Real-Time Adaptation (Closed-Loop VR)

Chapters 5 and 6 take up the second. The ultimate goal of this research is to create adaptive VR environments, and the processing pipeline developed here, which aggregates features in 15-second windows, is computationally lightweight enough to run in real time. Chapter 6 closes the loop, using the model's output to dynamically adjust the spawn rate of the cubes in the sustained attention task. The aim is to keep the user within a state of optimal challenge, the Zone of Proximal Development, for a greater proportion of a training session than a static difficulty progression would allow.

3.6.3 Deep Learning Approaches

While this study focused on interpretable feature engineering, the raw time-series data from eye-tracking and physiological sensors contains temporal dependencies that aggregate statistics (mean, std) may obscure. Chapter 5 takes a first step in this direction with a small bidirectional LSTM with self-attention. This is a modest application of deep learning compared with what could be done, for example with larger architectures, pre-training on other physiological datasets, or 1D Convolutional Neural Networks (CNNs) that learn latent features directly from the raw signal streams. Given the limited sample size (N=54), such approaches would require strong regularization to prevent overfitting, and the XGBoost results presented here provide a baseline for performance comparison.

Chapter 4

Bridging the Age Gap in Cognitive Workload Modeling

4.1 Introduction

This chapter addresses RQ2 and RQ3, introduced in Chapter 1. RQ2 asks whether eye-tracking is a reliable proxy for cognitive load in older adults. RQ3 asks what design considerations are needed when comparing models trained on younger versus older populations. Chapter 2 gave reason to doubt that the model developed in Chapter 3 would transfer unmodified. Physiological baselines shift with age in ways that can be mistaken by an algorithm for changes in cognitive state. Self-report becomes less reliable as a reference point in older cohorts. VR-based cognitive training is also increasingly popular for exactly this population, despite most physiological computing models being developed and validated on younger cohorts [152]. This chapter tests these claims by comparing 19 older adults (aged 60-87; 20 were recruited and one was excluded for missing data) against the younger cohort from Chapter 3 ($N = 54$, aged 18-58).

4.1.1 Rationale for Including Older Adult Participants

Older adults are one of the primary intended beneficiaries of the adaptive cognitive training system this thesis develops, since sustained attention and related executive functions decline with age and are precisely the targets that structured cognitive training aims to support (Chapter 1). Yet the model validated in Chapter 3 was built and tested exclusively on adults aged 18 to 58. Nothing in that validation establishes whether the same eye-tracking and physiological features, or the same trained model, remain valid once applied to the population the intervention is ultimately intended for. Including an older adult cohort in this thesis is therefore not an incidental extension but a test of whether the thesis's central technical contribution transfers to the population it is meant to serve, and, if it does not transfer without modification, what specifically needs to change.

The implications of this test run in both directions. If the younger-cohort model generalised to older adults without adjustment, this would support a single, age-agnostic sensing and modelling pipeline, considerably simplifying deployment. If it did not, as Chapter 2 gives reason to expect, this has direct design consequences. An adaptive system intended for older users would need age-specific calibration, age-aware feature engineering, or training data spanning the age range it is meant to serve. Accuracy figures reported from younger-only samples elsewhere in the literature should also not be assumed to hold once a system is deployed to older users.

4.1.2 Chapter Aim and Contributions

The aim of this chapter is to validate the experimental task with an older cohort and compare their results directly against the younger population from Chapter 3. By quantifying the divergences between these groups, the chapter examines the baseline physiological and behavioural constraints that may need to be accommodated to make VR systems more inclusive and adaptive for older adults.

Key Contributions:

- **Disentanglement of Baseline Physiology from Task Reactivity:** We investigate the extent to which age-related physiological differences are driven by static biological changes (e.g., senile miosis, reduced HRV) versus dynamic task engagement. We use machine learning classification to isolate baseline traits from cognitive load reactivity.
- **Evaluation of Cross-Age-Group Model Generalisability:** We develop and assess a Mixed Population predictive model. We evaluate whether training algorithms on diverse datasets improves their ability to generalise to new users compared to age-segregated models, as a step toward more inclusive adaptive systems.
- **Identification of Age-Specific Load Markers:** Using SHAP (SHapley Additive exPlanations) analysis, we explore how the hierarchy of predictive features shifts with age, determining whether the physiological signals of cognitive load (e.g., attentional stability vs. oculomotor mechanics) remain consistent across the lifespan.

4.2 Methodology

4.2.1 Experimental Design Overview

This chapter tests RQ2/H2 and RQ3/H3 (Chapter 1). It asks whether eye-tracking remains a reliable proxy for cognitive load in older adults. It also asks whether the design considerations, specifically which features carry that signal, differ between younger and older populations. The design combines a between-subjects comparison (age group) with the same within-subjects, repeated-measures task used in Chapter 3.

Independent variables. (1) Age group: Younger (18-58, $N = 54$, from Chapter 3) versus Older (60-87, $N = 19$, this chapter), a between-subjects factor. (2) Objective task difficulty, the same four within-subjects levels used in Chapter 3 (1.6s, 2.1s, 2.6s, 3.1s spawn interval).

Dependent variables. The primary classification target is, as in Chapter 3, the three-level subjective difficulty rating (*Too Easy*, *Just Right*, *Too Difficult*). Alongside this, the chapter compares the two age groups directly on a battery of secondary dependent measures: weighted game performance, NASA-TLX subscales (Total Workload, Mental Demand, Physical Demand, Frustration), and physiological/oculometric baselines (mean pupil diameter, HRV/RMSSD, saccade peak velocity, mean GSR). Model classification accuracy (Accuracy, F1, AUC) is itself treated as a dependent measure when compared across three population configurations: Younger-only, Older-only, and Mixed. Keeping the same perceived-difficulty label as Chapter 3 makes the age groups comparable. Self-report may behave differently in older adults (Chapter 2), so the chapter also compares ratings with objective performance and physiological baselines. It treats any divergence between them as a finding.

Measures. The same eye-tracking and physiological feature set as Chapter 3 (Section 3.3.6) was extracted for the older cohort using an identical pipeline, allowing like-for-like comparison. NASA-TLX subscales operationalise subjective workload; the MoCA served only as an inclusion screen for Mild Cognitive Impairment and was not used as an analysis variable.

The Results section below follows the same structure as the Data Analysis Strategy. Age-group differences on the secondary dependent measures are reported first, testing whether older adults' physiology and performance differ from younger adults' at all. Model performance across the three population configurations is reported second, testing RQ2, whether the signal is usable in older adults specifically. SHAP feature importance is reported last, testing RQ3, whether the specific measures driving prediction differ by age group.

4.2.2 Participants

An additional sample of 20 participants was recruited through a local University of the Third Age (U3A), a peer-led lifelong-learning organisation for retired and semi-retired adults, and completed the same task as described in Chapter 3, using the VR application and data collection pipeline described there. Recruitment was advertised as open to adults aged 60 to 89; the oldest participant who actually took part was 87, which is why the achieved sample is described as aged 60-87 in the analyses that follow. The sample consisted of 6 male and 14 female participants, mean age = 71.7, SD = 7.28. All participants were right-handed. One participant was excluded due to large amounts of missing data. This 'older adults' sample was compared to the younger sample from Chapter 3 ($N = 54$); both are part of the same publicly available dataset (Chapter 3). Ethical approval was granted by the University of Glasgow, College of Science and Engineering, School of Computing Science (application number 300230006).

At the end of the experiment, similarly to the previous sample, participants completed a Simulator Sickness Questionnaire [84] once each ($N = 20$ administrations, the full recruited sample). Only 3 participants reported moderate symptoms. One reported moderate difficulty focusing, one moderate sweating, and the last moderate general discomfort, headache, difficulty focusing and blurred vision. No participant reported symptoms in the high-discomfort range that would have warranted exclusion, and, as in Chapter 3, no participants were excluded on the basis of SSQ responses. Participants also completed the MoCA test once each, as a screen for Mild Cognitive Impairment; all 20 recruited participants passed the assessment. No participants withdrew from the study. As in Chapter 3, participants completed a NASA-TLX questionnaire after each of the three sessions (6 subscales per administration; $20 \times 3 = 60$ administrations as delivered). This uses the full recruited sample of 20, since the single participant excluded from the classification analyses below was excluded for missing physiological data rather than incomplete NASA-TLX data.

This sample is not representative of the general population of older adults, and the ways in which it is not are relevant to interpreting the results reported below. Recruiting through a University of the Third Age is itself a documented source of selection bias. U3A membership is consistently found to skew toward the more educated and socially engaged, and in particular toward women, to the extent that commentators have characterised it as dominated by middle-class, urban women [54]. Gender differences in participation have been linked to retirement patterns and social group membership [181]. This is consistent with the composition of the present sample, which is predominantly female (14 of 20 recruited, 70%), a more pronounced skew than the roughly even split in the younger cohort from Chapter 3 (27 of 54, 50% female). The sample is also exclusively right-handed, whereas the younger cohort included six left-handed participants. Every participant passed the MoCA screen for Mild Cognitive Impairment. The sample therefore reflects cognitively healthy, socially engaged older adults specifically, consistent with the scope of healthy ageing set out in Chapter 1. It does not represent the wider older population, which includes individuals who are less socially engaged, more educationally varied and, with increasing age, more likely to show cognitive impairment. More generally, volunteers for technology-focused studies are unlikely to be representative even of a U3A population. A comparable study recruiting older adults for a robotics workshop found that volunteers scored higher than the national average on personality traits such as openness. This suggests that willingness to take part in novel technology research selects for further traits on top of whatever bias the recruitment channel itself introduces [77]. The present sample is therefore best read as informative about eye-tracking and physiological cognitive load estimation in relatively healthy, socially engaged, technology-willing older adults, rather than as representative of ageing more broadly, a point returned to in the Discussion.

4.2.3 Task design

Task design remained unchanged from the one in Chapter 3. Each participant completed 3 sessions, each consisting of 8 levels with 4 levels of difficulty cycled through twice in each session. The experiment was conducted using the same materials and same pre-processing strategy as described in Chapter 3.

4.2.4 Data Analysis Strategy

To investigate the impact of aging on physiological computing in VR, we performed a comparative analysis between the *General Population* cohort (Age 18-58, $N = 54$) and the *Older Adult* cohort (Age 60-87, $N = 19$). The analysis had four components:

- **Statistical Comparison:** We aggregated physiological and performance metrics at participant level to account for session-to-session variance. We utilised the non-parametric Mann-Whitney U test to compare groups on behavioural metrics, NASA-TLX subscales, and physiological baselines. P-values were corrected using the Benjamini-Hochberg (FDR) procedure.
- **Population Differences vs Within Participant Differences:** To determine whether age-related physiological differences were driven by baseline shifts (e.g., senile miosis) or task-related reactivity, we trained a Random Forest classifier to predict participant age group. We compared the model's performance (AUC) using *Global Scaling* (preserving absolute values) versus *Within-Participant Z-Scoring* (preserving only relative reactivity).

- **Subjective Difficulty Distribution:** To assess differences in the subjective perception of task challenge, we performed a Chi-Squared test of independence on the distribution of the reported difficulty ratings (*Too Easy, Just Right, Too Difficult*). A post-hoc analysis using standardized residuals (Z-scores) was conducted to identify specific rating categories that significantly deviated from the expected frequency for each age group.

- **Model Performance Comparison across Cohorts:** We evaluated the predictive performance of the XGBoost model (identified as the best-performing architecture in the experiment including the general population only) across three distinct population configurations: *Younger Adults*, *Older Adults*, and a *Combined* population. To assess generalisability to new users, we employed a Leave-One-Participant-Out (LOPO) cross-validation scheme. Each model underwent Bayesian hyperparameter optimization (Optuna) and used the 30-feature set selected in Chapter 3, to identify whether age-specific training yields higher accuracy than a global one-size-fits-all approach.

Statistical Assumptions and Test Selection

Prior to the comparative analysis, the distribution of all dependent variables was assessed for normality using the Shapiro-Wilk test and visual inspection of histograms. The analysis revealed significant deviations from normality across the older adult cohort for all key metrics, necessitating a non-parametric analytical approach.

Specifically, game performance scores exhibited a non-normal distribution in both groups ($p < 0.001$), characterised by a distinct ceiling effect in the younger cohort (negative skew) and high variability in the older cohort. Physiological metrics such as Heart Rate Variability (RMSSD) followed a typical positively skewed biological distribution, violating normality assumptions for both groups ($p < 0.001$).

Notably, while oculomotor metrics (Mean Pupil Diameter and Saccade Peak Velocity) followed a normal distribution in the larger younger cohort ($p > 0.05$), they deviated significantly from normality in the older cohort ($p < 0.05$). This divergence is likely attributable to the smaller sample size ($N = 19$) and the higher inter-subject variability inherent to aging physiology.

Given these universal violations of normality in the older cohort and the unequal sample sizes ($N_{Young} = 54$ vs. $N_{Old} = 19$), the **Mann-Whitney U test** was utilised for all between-subjects comparisons. This test compares mean ranks rather than raw means, making it robust to the observed outliers and skewed distributions.

4.3 Results

Following the Experimental Design Overview above, results are reported in three parts: age-group differences on the secondary dependent measures (performance, workload, physiology), model classification performance against the subjective-difficulty target across the three population configurations, and the SHAP feature-importance comparison that speaks to RQ3 directly.

4.3.1 Participant Game Performance Comparison

Older adults demonstrated a significantly worse game performance compared to younger adults. This was calculated using Weighted Performance, which is the percentage of all cubes that were destroyed multiplied by the inverse of the difficulty level, ranging from 0 to 3.1. As shown in the left panel of Figure 4.1, Weighted Performance scores were significantly lower for the older cohort ($M = 1.52, SD = 0.45$) compared to the younger cohort ($M = 2.08, SD = 0.27$; U test, $p < 0.001$). The effect size for this difference was very large (Cohen's $d = 1.51$).

This difference is further broken down in the right panel of Figure 4.1. Due to the unequal variances observed at higher difficulty levels (Levene's test $p < 0.01$), a Linear Mixed Model (LMM) was conducted to assess the interaction between Age and Difficulty. The analysis revealed a significant *Age* × *Difficulty* interaction ($p < 0.001$). While younger adults maintained relatively stable performance as difficulty increased (Mean drop from the easiest to the hardest ≈ 0.16), older adults exhibited a significantly steeper rate of decline (Mean drop ≈ 0.31), effectively doubling the performance penalty associated with high task demands. This is consistent with older adults reaching a performance ceiling at lower difficulty levels than younger adults.

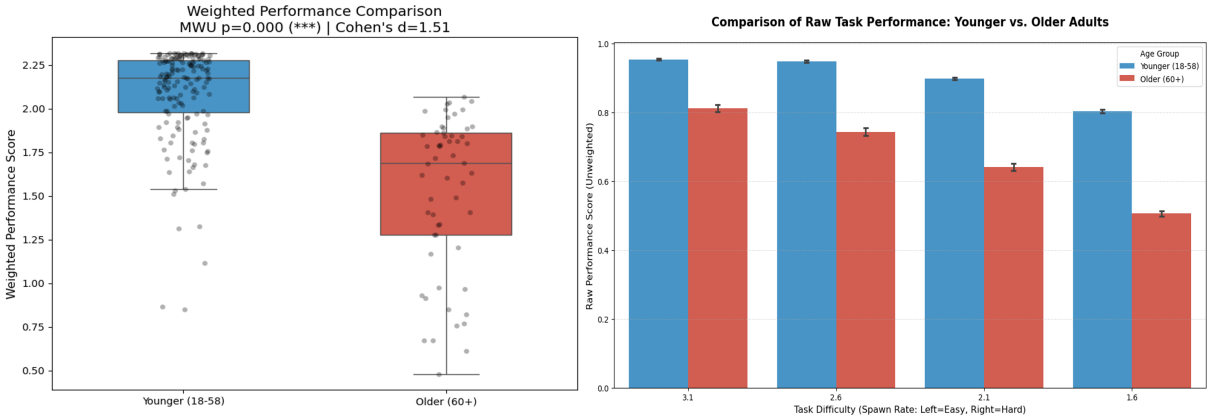


Figure 4.1: Performance differences between age groups. Left: Older adults show significantly lower Weighted Performance scores ($p < 0.001, d = 1.51$). Right: Comparison of Raw Task Performance across difficulty levels (Spawn Rates 3.1 to 1.6), showing a steeper decline in older adults as task difficulty increases.

4.3.2 Subjective Workload (NASA-TLX)

Despite the clear gap in objective performance, the subjective experience of workload differed mainly in emotional response rather than in cognitive demand. As detailed in Table 4.1, there was no statistically significant difference in reported *Mental Demand* ($p = 0.766$) between the two groups. However, older adults reported significantly higher *Frustration* ($p < 0.001, d = -0.61$). This discrepancy suggests that the perceived difficulty

for older adults may be driven by performance limitations and the resulting emotional response, rather than an incapacity to process the cognitive load of the task. Full NASA-TLX descriptive statistics for all subscales, including those not analysed here, are reported in Appendix 10, Section 10.1.

Table 4.1: Summary of the Age Comparison Results (Session Level). Significant differences were observed across performance, frustration, and physiological metrics, despite similarly reported mental and physical demand.

Metric	**Younger (M)**	**Older (M)**	***p*-value**	**Cohen's d**
Subjective Workload (NASA-TLX)				
TLX Total Score	51.71	59.37	0.033	-0.42
Mental Demand	10.39	10.72	0.766	-0.07
Physical Demand	7.46	6.93	0.386	0.11
Frustration	5.95	8.79	< 0.001	-0.61
Physiological Measures				
Mean Pupil Diameter (mm)	4.20	3.90	0.007	0.31
HRV (RMSSD)	61.02	47.55	< 0.001	0.31
Saccade Peak Velocity	1.95	1.65	< 0.001	0.69
Mean GSR (Ohms)	1.08×10^6	1.19×10^6	0.040	-0.09

4.3.3 Physiological Differences

Physiological metrics also differed between the age groups (see Table 4.1), in line with the age-related differences observed in behavioral performance.

Oculometrics: A significant reduction in *Saccade Peak Velocity* was observed in older adults ($MOlder = 1.65$ vs $MYounger = 1.95$; $p < 0.001, d = 0.69$), consistent with known age-related slowing of motor control. Similarly, older adults exhibited significantly smaller *Mean Pupil Diameter* ($p = 0.007$), likely attributable to physiological senile miosis rather than a lack of cognitive engagement.

Autonomic Measures: Heart Rate Variability (RMSSD) was significantly lower in the older group ($M = 47.55$ ms) compared to the younger group ($M = 61.02$ ms; $p < 0.001$). While HRV naturally declines with age, the lower values in the older cohort, combined with the performance data, may suggest a reduced physiological reserve to adapt to the high-demand VR task. Additionally, Galvanic Skin Response (GSR) indicated slightly higher skin resistance in older adults ($p = 0.040$, $d = -0.09$), a small difference consistent with age-related reductions in skin conductance.

To determine if the physiological metrics alone could distinguish between the age groups, a Random Forest classifier was trained using only oculometric and physiological features. **Training set:** pooled sub-session-level rows from both cohorts (Younger $N = 54$ plus Older $N = 19$ participants; 1,736 rows in total), using the same 112-column engineered feature pool described in Chapter 3 (Section 3.3.6), restricted here to its oculometric and physiological columns. The model was evaluated using 5-fold Stratified Group Cross-Validation to ensure no participant's data appeared in both training and testing sets.

As shown in Table 4.2, the classification strategy depended heavily on the scaling method. The Global Standardization approach yielded moderate predictive power, with an AUC of 0.77 and an Accuracy of 75.1%, which is close to the 74% obtained by always predicting the younger group (54 of 73 participants). Feature importance analysis revealed that *Pupil Diameter Standard Deviation* (27.0%) and *HRV (SDNN)* (16.7%) were the top predictors.

In contrast, the Within-Participant Z-Score strategy, which normalizes data relative to each individual's mean, caused performance to drop to near-random levels (AUC 0.55). This disparity indicates that the distinguishing physiological markers of age are *baseline traits* (e.g., consistently lower HRV or reduced pupillary range due to senile miosis) rather than transient, task-induced reactions.

Table 4.2: Age Classification Results (Physiological Features Only). The drop in performance for Within-Participant Z-Scoring suggests that age-related differences were driven by absolute baseline values rather than relative task reactivity.

Scaling Strategy	Accuracy	ROC AUC
Global Standardization	0.751	0.770
Within-Participant Z-Score	0.689	0.547

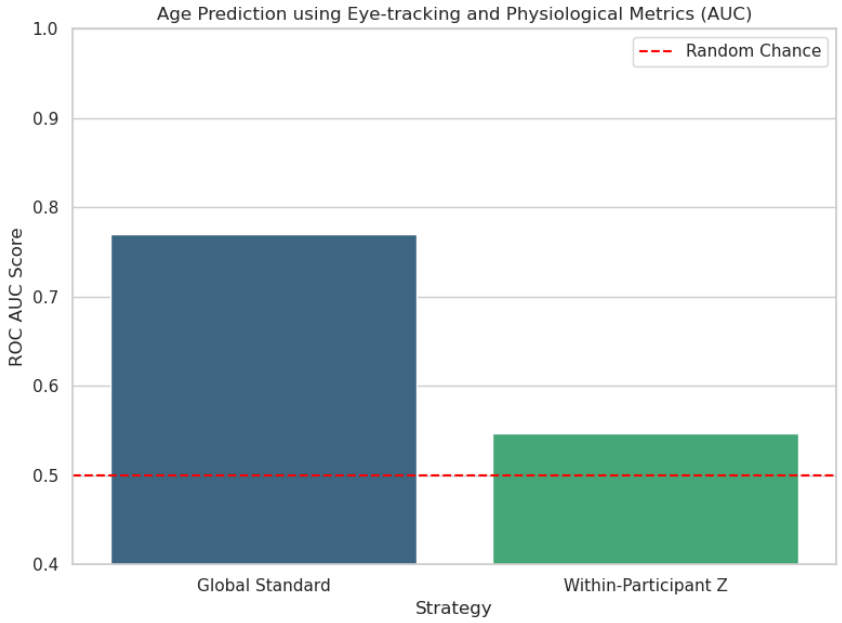


Figure 4.2: Comparison of the Age Prediction models. The Global model (left) effectively distinguishes age groups, whereas removing baseline information (Within-Participant Z, right) eliminates predictive capability.

4.3.4 Subjective Perception of Difficulty

To evaluate whether the two age cohorts perceived the task challenge differently, a Chi-Squared test of independence was performed on the distribution of the subjective difficulty ratings (*Too Easy, Just Right, Too Difficult*).

As illustrated in Figure 6.4, there was a significant difference in how younger and older adults categorized the task difficulty ($\chi^2 = 99.38, p < 0.001$). While younger adults frequently rated the task as *Too Easy* (42.2% of responses), older adults rarely did so (18.6%), instead clustering their responses around *Just Right* and *Too Difficult*.

A post-hoc analysis using standardized residuals (Z-scores) was conducted to identify the specific drivers of this divergence. A residual magnitude greater than $|1.96|$ indicates a statistically significant deviation from the expected frequency. The analysis revealed that:

- **Younger Adults** significantly over-selected *Too Easy* ($Z = 3.70$) and under-selected *Too Difficult* ($Z = -3.41$).
- **Older Adults** significantly avoided the *Too Easy* rating ($Z = -6.19$) and were significantly more likely to rate the task as *Too Difficult* ($Z = 5.72$) compared to the expected distribution.

This shift suggests that the window of optimal challenge differs between the groups, as the task parameters that constituted a trivial challenge for younger participants often represented a substantial demand for older adults.

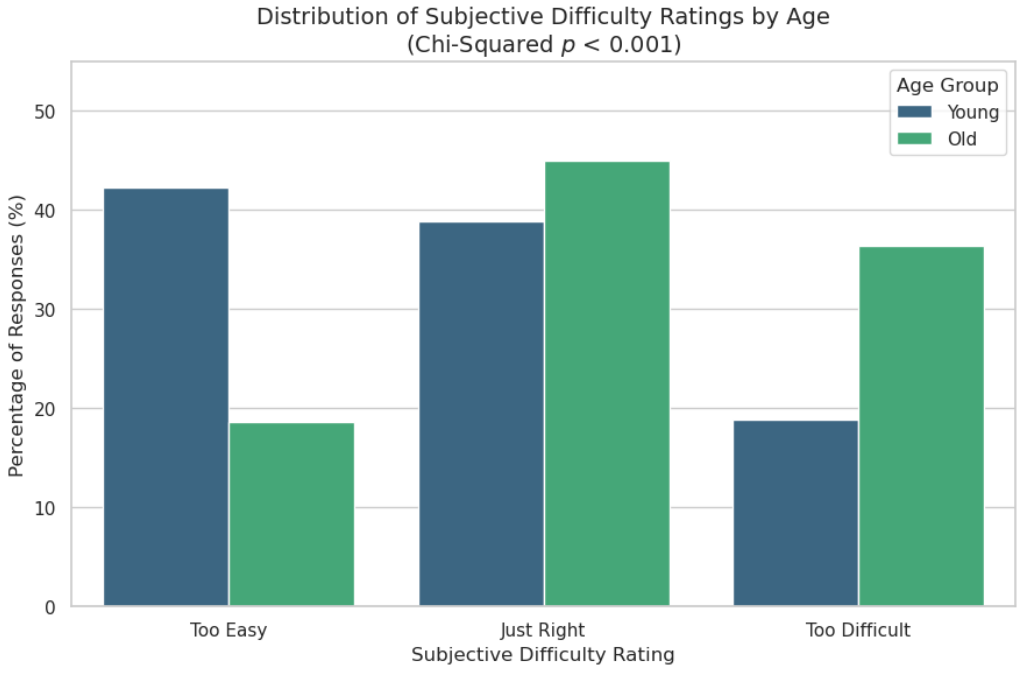


Figure 4.3: Distribution of the Subjective Difficulty Ratings by Age Group. A significant distributional shift ($p < 0.001$) is observed, with older adults significantly shifting towards Just Right and Too Difficult ratings while younger adults predominantly selected Too Easy.

4.3.5 Predictive Modelling Results

To evaluate the feasibility of the real-time cognitive load detection across age groups, we trained three separate XGBoost classifiers (Mixed, Younger, and Older) using the top 30 features identified by the SHAP-based feature selection introduced in Chapter 3. That selection was computed on the combined younger and older dataset, so that the same features can be compared across the three populations. **Training set sizes:** Mixed ($N = 73$ participants; 1,736 sub-session-level rows), Younger ($N = 54$ participants; 1,280 rows, identical to the dataset described in Chapter 3), and Older ($N = 19$ participants; 456 rows), each using the same 30-feature set as Chapter 3. Each model was optimized using Bayesian hyperparameter tuning (Optuna) and evaluated using Leave-One-Participant-Out (LOPO) cross-validation to ensure the results reflect generalisation to unseen users.

Hyperparameter Configuration: The optimization process yielded distinct architectural preferences for each demographic, reflecting the underlying complexity of their physiological data:

- **Mixed Population:** A moderately complex model was required to capture the variance across the full age spectrum ($n_estimators = 578$, $max_depth = 6$, $learning_rate = 0.042$, $subsample = 0.78$).
- **Younger Adults:** This group favoured a standard depth but higher learning rate, suggesting more consistent signal patterns ($n_estimators = 489$, $max_depth = 5$, $learning_rate = 0.084$).
- **Older Adults:** The optimizer converged on a significantly shallower and slower-learning architecture ($n_estimators = 383$, $max_depth = 3$, $learning_rate = 0.023$). This conservative configuration may limit overfitting to the higher noise levels and the individual variability observed in the older cohort's smaller dataset.

Model Performance Analysis: Table 4.3 summarizes the classification metrics across the three cohorts. The *Younger* and *Mixed* models achieved comparable performance (Accuracy $\approx 54-55\%$, AUC ≈ 0.72), indicating that the inclusion of the older adults in the mixed training set did not noticeably degrade the model's ability to classify the majority demographic. It also did not improve it, which is plausible given the differences between the datasets.

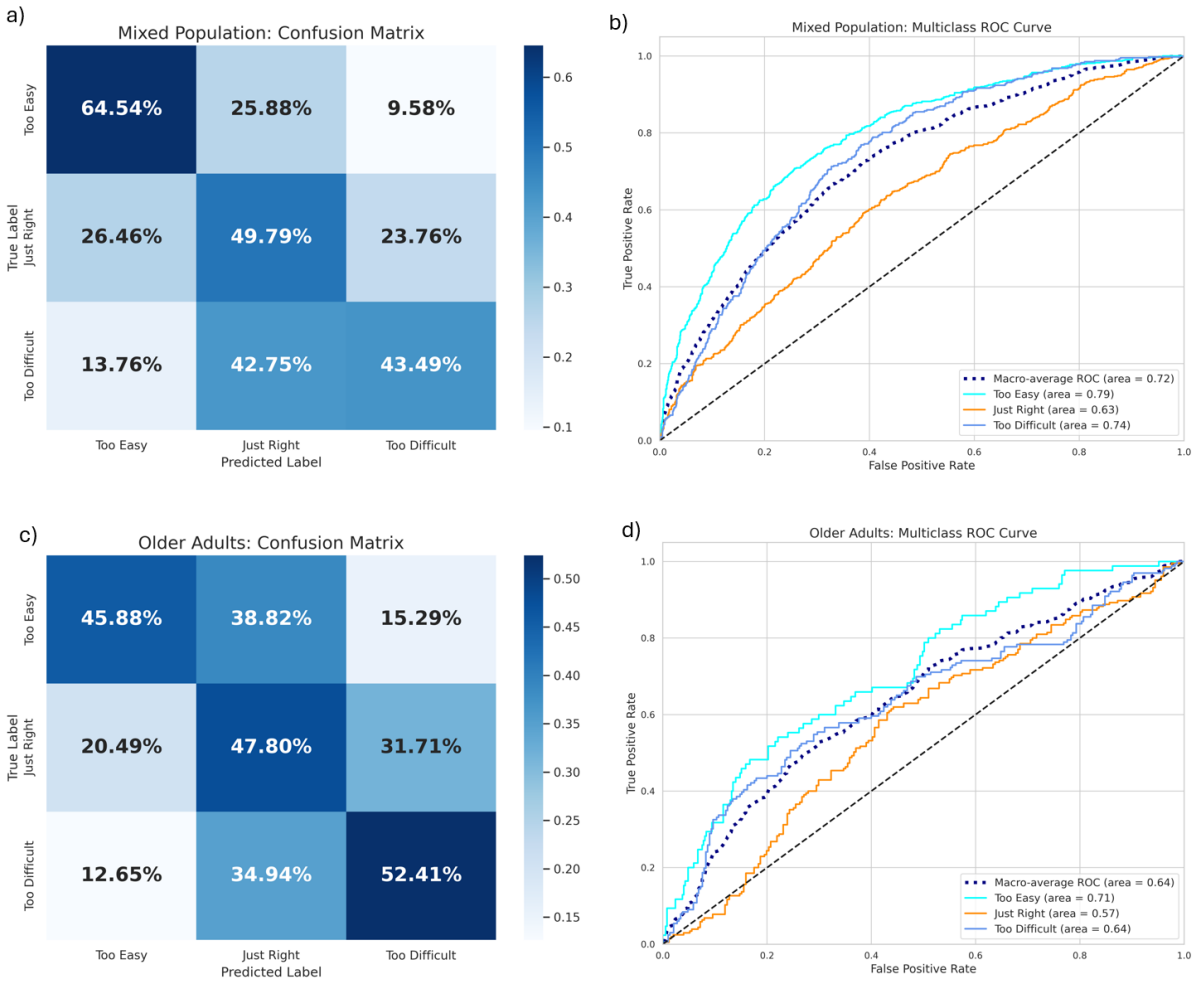


Figure 4.4: Performance comparison between Mixed Population (Top) and Older Adults (Bottom). Panels (a) and (c) show confusion matrices, highlighting the higher misclassification rate in the older cohort, particularly between Too Easy and Just Right. Panels (b) and (d) display Multiclass ROC curves, revealing a drop in Macro-Average AUC from 0.72 to 0.64 in the older population.

However, the *Older Adult* specific model demonstrated a notable reduction in predictive power (Accuracy = 49.1%, AUC = 0.66), which may be due in part to the smaller sample size. Learning curves in Chapter 7 plateau within the 19 older participants available, but this does not show that a much larger and more varied older sample (for example 100 or 200 participants) would not improve performance. This performance gap is further visualized in Figure 4.4, which compares the Confusion Matrices and ROC curves for the Mixed and Older populations.

- **Confusion Matrices (Fig 4.4a, c):** The Mixed model successfully identified *Too Easy* states with 64.5% accuracy. In contrast, the Older Adult model struggled to distinguish *Too Easy* (45.9%) from *Just Right* (38.8%), likely due to the floor effect observed in the subjective ratings where older adults rarely reported low difficulty.
- **ROC Curves (Fig 4.4b, d):** The Macro-Average AUC of the pooled ROC curves dropped from 0.72 in the Mixed population to 0.64 in the Older cohort (Table 4.3 reports the mean of per-participant AUCs: 0.724 and 0.665). While the Mixed model maintained balanced detection across classes, the Older Adult model showed specific weakness in the *Just Right* category (AUC = 0.57), suggesting that the physiological transition between optimal challenge and overload is less distinct or more variable in older users.

Table 4.3: Summary of XGBoost Classification Performance (LOPO Cross-Validation). Values represent Mean ± Standard Deviation across all participants. The Older Adult cohort shows reduced classification accuracy and AUC compared to the Younger and Mixed populations.

Population	Accuracy	Precision	Recall	F1-Score	ROC AUC	AUPRC
Mixed ($N = 73$)	0.537 ± 0.14	0.503 ± 0.16	0.506 ± 0.14	0.466 ± 0.14	0.724 ± 0.12	0.615 ± 0.12
Younger ($N = 54$)	0.556 ± 0.15	0.528 ± 0.15	0.527 ± 0.16	0.491 ± 0.16	0.724 ± 0.12	0.606 ± 0.14
Older ($N = 19$)	0.491 ± 0.11	0.479 ± 0.15	0.505 ± 0.12	0.457 ± 0.13	0.665 ± 0.09	0.562 ± 0.10

Feature Importance Analysis: To investigate the physiological drivers behind the predictive models, we analyzed the SHAP values for the Younger and Older adult cohorts. As shown in Figure 4.5, the hierarchy of influential features reveals differences in how cognitive load manifests across age groups.

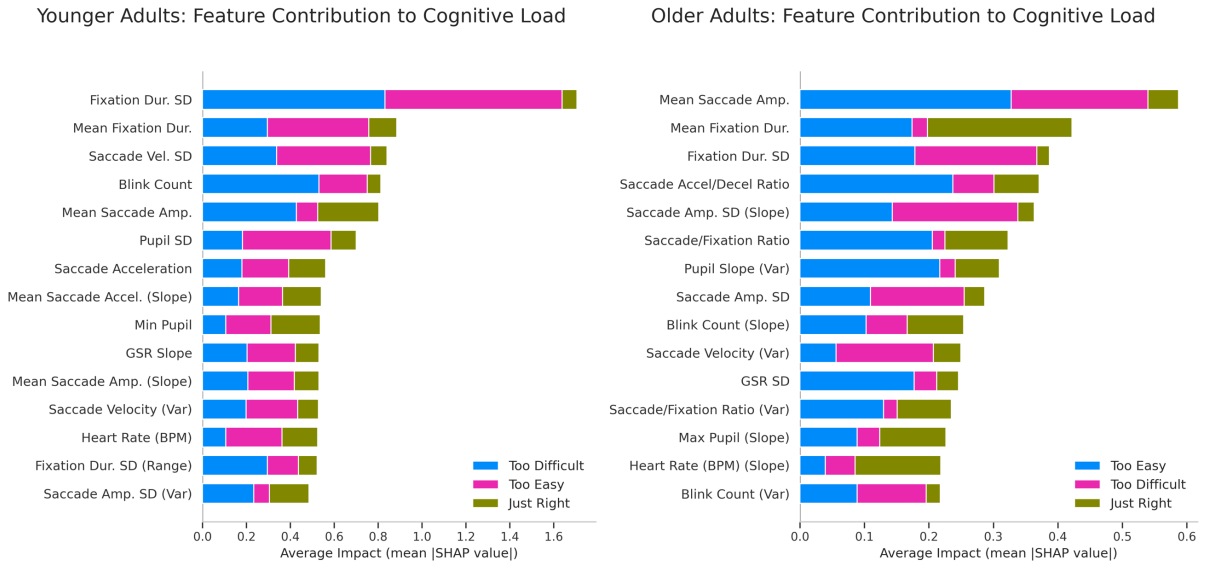


Figure 4.5: Feature Importance Comparison (SHAP Summary Plots). Left: Younger Adults. Right: Older Adults. The Younger model is dominated by Fixation Stability (Fixation Dur. SD), whereas the Older Adult model relies heavily on Oculomotor Mechanics (Saccade Amplitude and Acceleration/Deceleration Ratios).

Younger Adults: For the younger cohort (Left Panel), the model is overwhelmingly dominated by *Fixation Duration Standard Deviation*. This single feature exhibits an average impact value (≈ 1.7) nearly double that of the next closest feature. This indicates that for younger adults, cognitive load is primarily characterized by fluctuations in the ability to maintain steady visual focus. The dominance of fixation-based metrics (occupying the top two spots) suggests that their physiological bottleneck is cognitive/attentional rather than motoric.

Older Adults: In contrast, the older cohort (Right Panel) shows a distinct shift toward motor-control metrics. The top predictor shifts to *Mean Saccade Amplitude*. This suggests that the mechanics of scanning (specifically how far older adults move their eyes to engage with targets) becomes the most critical indicator of their workload state. Notably, the *Saccade Acceleration/Deceleration Ratio* appears as a top 4 predictor. This metric captures the velocity profile of eye movements, likely reflecting the age-related degradation in oculomotor control (e.g., slower initiation or termination of saccades) that becomes more pronounced under high load.

This comparison suggests that while cognitive load in younger adults is detectable via gaze stability (an index of attention), in older adults it is more strongly predicted by eye movement dynamics (an index of physiological scanning effort). In other words, when younger participants struggle with the task, they struggle to keep up with processing of what they see. However, when older participants struggle, they struggle to physically move their eyes at the speed required by the task.

4.4 Discussion

The findings of this research point to a divergence between younger and older adults in how cognitive load is manifested and perceived in VR. Physiological reactivity to load appears to remain usable with age, but it is accompanied by baseline differences that standard computing models may not account for.

4.4.1 Performance Limits and Neural Compensation

A key finding of this study is the disconnect between objective performance and subjective mental demand. Despite a significant performance decline in older adults ($d = 1.51$), their reported *Mental Demand* did not differ significantly from the younger cohort ($p = 0.766$). This can be interpreted through the *Compensation-Related Utilization of Neural Circuits Hypothesis (CRUNCH)*, introduced in Chapter 2 [138]. According to CRUNCH, older adults recruit additional neural resources to maintain performance parity on tasks of low-to-moderate difficulty. This neural over-recruitment may explain why reported mental demand remained comparable. As difficulty increased, older adults may have reached a ceiling beyond which further compensation was no longer possible, which would be consistent with the steeper performance drop and the rise in *Frustration* ($d = -0.61$).

4.4.2 Eye-tracking Differences: Attentional vs. Motoric Bottlenecks

SHAP analysis revealed that younger adults' workload is primarily predicted by attentional stability, measured through fixation duration variability. For younger cohorts, the bottleneck is primarily cognitive: when task demands increase, they struggle to keep up with the processing of visual information [21].

In contrast, older adults' workload is driven by oculomotor mechanics, specifically mean saccade amplitude and velocity profiles. This suggests an age-related contraction of the **Useful Field of View (UFOV)** [18]. As load increases, the functional area from which older adults can extract information shrinks, necessitating more frequent or larger-amplitude saccades to scan the environment. This effectively substitutes physical eye movement for a diminished attentional span. The struggle for older adults is thus sensorimotor: the mechanics of physically moving the eyes to keep up with the task spawn rate may reach a biological ceiling before cognitive processing limits are met [119].

4.4.3 The Value of Mixed-Population Models for Inclusivity

The slight degradation in performance observed in the Mixed Population model (53.7% accuracy) compared to the Younger model (55.6%) is a small trade-off for inclusivity. By training algorithms on a dataset that includes the absolute baseline shifts of older adults, the system learns to navigate the continuum of human physiology [164].

A system trained solely on younger adults could misinterpret an older adult's low HRV and constricted pupil as permanent indicators of high stress or low arousal, respectively, leading to maladaptive difficulty adjustments. The Mixed Population model may offer a workable route around the cold start problem, allowing a system to interpret diverse user states across a wide age range without requiring prior knowledge of the user's age or baseline. Each model in this chapter was, however, trained and evaluated within its own population, and no younger-to-older transfer test was run.

4.4.4 Implications of Sample Representativeness

The divergence documented above, attention-driven load markers in younger adults, motor-driven markers in older adults, was obtained from the cognitively healthy, majority-female U3A sample described in the Participants section, and the practical implications of this chapter's findings should be scoped accordingly. For the healthiest and most engaged segment of the older population, of the kind likely to belong to a U3A and then volunteer for a VR study, the results suggest that a single sensing pipeline can be adapted to older users with a change in feature weighting rather than a wholesale redesign. An older population may include emerging cognitive impairment, greater frailty, less social engagement, or less willingness to engage with headset-based VR at all. This sample was not positioned to detect any of these factors. The baseline shifts documented here, particularly the shift toward a motor-driven bottleneck identified via SHAP, are therefore more likely to represent a lower bound on the divergence from younger-cohort models than an upper one. A larger and more demographically varied older cohort, recruited through channels less likely to select for social engagement and openness to novel technology [77], would be needed before the Mixed Population model's applicability could be extended beyond the relatively narrow band of the older population sampled here.

4.4.5 Implications for VR Design: From Cognitive to Motoric Scaffolding

These findings suggest that VR design for older adults may benefit from motoric scaffolding as well as cognitive simplification. Following the **Selective Optimization with Compensation (SOC)** framework [19], intervention designs may benefit from tools that compensate for physiological losses.

Adaptive VR systems could consider motor accommodation such as slowing target movement, increasing target size, or centralizing UI elements when the model detects motoric strain. By providing motoric compensation, the system may allow older users to dedicate their limited resources to the *germane* cognitive task, which may reduce the frustration-driven withdrawal often observed in clinical VR interventions for the elderly.

4.5 Summary

This chapter tested RQ2 and RQ3: whether eye-tracking remains a reliable proxy for cognitive load in older adults, and what design considerations are needed when comparing models across age groups. The answer to both is qualified rather than a simple yes or no. Eye-tracking and physiological sensing retained meaningful predictive validity in the older cohort (AUC = 0.66 under LOPO validation), supporting the first part of H2 (meaningful predictive validity). The features driving that validity differed by age: where younger adults' load is best captured by attentional stability (fixation duration variability), older adults' load is better captured by oculomotor mechanics (saccade amplitude and acceleration/deceleration ratios). This is consistent with H3's prediction that the markers most predictive of load would differ between cohorts, although the older-adult

model was also weaker (AUC 0.66 versus 0.72), so the older cohort's signal is not simply a differently weighted copy of the younger cohort's. It is also consistent with the baseline-shift argument set out in Chapter 2: age-related physiological change appears to alter which markers carry the load signal, in addition to adding noise to it.

This chapter also found lower performance at higher difficulty levels in older adults, and a significant rise in frustration despite comparable mental demand ratings. It suggested that a Mixed Population model may provide a workable, if imperfect, route to a single generalisable system rather than requiring fully age-segregated models. As discussed above, these conclusions are scoped by the sample's representativeness. The cohort recruited here is a healthy, technology-willing subset of the older population. The true extent of divergence between younger and older models in a broader, more representative older sample remains an open question for future work.

Chapter 5 takes the combined younger-and-older dataset assembled across this chapter and Chapter 3. It uses this dataset to build a model architecturally suited to real-time use. Chapter 6 then deploys that model in a real-time, closed-loop setting, using the live physiological stream to drive dynamic difficulty adjustment rather than relying on the offline, retrospective validation used throughout this chapter. Instead of relying on pre-set difficulty curves or potentially unreliable self-reports, the system evaluated in Chapter 6 adapts task difficulty in real time, with the aim of keeping users within their optimal window of challenge (the real-time cohort was aged 18 to 59), addressing RQ4 through RQ6.

Chapter 5

Development and Pre-Deployment Evaluation of a Real-Time Cognitive Load Estimation Model

5.1 Introduction

Chapters 3 and 4 developed and validated an offline cognitive load estimation model on session-level, hand-aggregated features, evaluated after the fact against data collected under fixed, discrete difficulty conditions. Testing RQ4 through RQ6, introduced in Chapter 1, requires a model that can consume a continuous stream of eye-tracking and physiological data and produce a cognitive load estimate while a session is still under way, so that task difficulty can be adjusted in real time rather than assessed retrospectively. This chapter describes and evaluates the model, providing the necessary context for chapter 6, which then deploys the model in a live, closed-loop VR system and reports the results of that deployment.

A note on terminology. This thesis uses "real-time" to describe the system built in this chapter and deployed in Chapter 6. Difficulty in this system is updated once per 60-second task level rather than continuously from moment to moment, so it is worth being explicit about what the term claims. There is no agreed technical threshold in this literature that separates "real-time" from "too slow to count": a system that updates every second is not obviously more or less real-time than one that updates every ten,

and the field has not converged on a number. What the term is doing here is marking a causal and temporal distinction, not a speed claim. In the offline paradigm of Chapters 3 and 4, data collection, modelling and evaluation are fully decoupled in time. The participant completes the session and leaves, and only afterwards is the data analysed and a model fitted. By construction, the model then has zero influence over the experience that generated the data it is trained on. In the system built in this chapter and deployed in Chapter 6, by contrast, inference happens while the participant is still in the session (Chapter 6 reports sub-500ms end-to-end latency from data acquisition to estimate). Its output is acted upon, adjusting the difficulty the participant then experiences, before the session ends. Once that adjustment is made, it cannot be revisited or fine-tuned with hindsight the way an offline analysis can; the prediction has already had its effect. The 60-second cadence itself is a property of the task design, each level is a fixed-length unit within which difficulty is held constant, consistent with the task structure used throughout this thesis, rather than a limitation of how quickly the model can produce an estimate. It is this in-session, causally consequential relationship between prediction and experience, and not a specific refresh rate, that this thesis treats as the operative difference between "real-time" and "offline". The term is used consistently in that sense throughout.

This use of "real-time" is also consistent with how the term functions elsewhere in this literature, where it is more often an aspiration than an implemented property. Bhatti et al.'s CLARE dataset is named for exactly this goal, "Cognitive Load Assessment in Real-time", yet the system it describes is an offline benchmark, evaluated with 10-fold and leave-one-subject-out cross-validation on previously collected data, with no live, closed-loop deployment reported [25]. Pinilla et al. similarly present a technique for "real-time affect detection in virtual reality", but validate it entirely offline, on EEG recorded while participants passively watched a series of pre-selected videos in a VR headset, again with no live adaptation of the VR experience itself [129]. Eckert et al.'s method for correcting pupil measurements for the pupillary light reflex, discussed in Chapter 2, is framed around calibration mapping functions "that can be employed in real-time during a VR experience", but the paper itself only evaluates those functions offline, against recorded pupillometry data [50]. The HP Labs Omnicept dataset, discussed at length in Chapter

2, is framed in the same terms. Its authors report having "developed a novel machine learning solution to predict cognitive load in real time" and aim "to introduce a new generation of adaptive VR training tools that personalise the experience in real-time". Yet the study itself is a large-scale ($N = 738$) offline data collection and model-training exercise, releasing a static test dataset rather than reporting any live, closed-loop deployment [178]. Mortazavi et al.'s systematic review of dynamic difficulty adjustment similarly identifies real-time, physiologically-driven adaptation in VR as a promising direction while noting that its application in real-time VR settings remains comparatively underexplored in the literature it reviews [115]. Against that backdrop, and whatever the merits of its specific 60-second cadence, the system in this chapter and Chapter 6 does something that, to my knowledge, a substantial part of this literature proposes but does not yet implement. It produces an estimate from live physiological data and uses it to change the participant's experience before that experience is over. "Real-time" is not a term coined for this thesis; it is already used in this literature, almost always as a stated goal rather than a delivered one.

Building a model capable of real-time operation is not simply a matter of retraining the XGBoost classifier from Chapter 3 more often. That model relied on session-level aggregate features (means, standard deviations, slopes and ranges computed over an entire task level) derived offline from recorded data. This chapter instead adopts a windowed, sequential representation, four 15-second windows per task level, that can be computed from the live data stream and passed to a recurrent architecture (a bidirectional LSTM with self-attention) capable of learning temporal structure across those windows rather than relying on precomputed summary statistics. The estimate is produced once, after the 60-second level ends, and is used to set the difficulty of the next level. This also gives a natural opportunity to check an assumption carried over from earlier chapters: that hand-engineered temporal features are actually necessary, rather than something a sufficiently expressive sequence model could learn to reconstruct from minimally processed signal on its own. Section 5.2.6 tests this directly.

5.1.1 Chapter Aim and Contributions

The aim of this chapter is to build and offline-validate a cognitive load estimation model architecturally suited to real-time deployment, using the combined dataset from Chapters 3 and 4 ($N = 73$: 54 younger and 19 older adults). It also aims to establish, before any live deployment is attempted, how much confidence that offline validation actually supports. This matters because Chapter 2 identified pooled-subject cross-validation and task-stacking as two of the field's central methodological weaknesses. Avoiding them here does not by itself guarantee that offline performance will be either strong or representative of real-time performance. This chapter therefore reports both the model's offline accuracy and a direct test of whether its feature engineering pipeline earns its complexity, rather than treating a working architecture as sufficient justification on its own.

Contributions:

- **A windowed, sequence-based architecture suited to real-time inference:** an attention-enhanced bidirectional LSTM that consumes four sequential 15-second windows per task level, rather than the session-level aggregate features used in Chapters 3 and 4, so that a cognitive load estimate can be produced from the live data stream as soon as a level ends.
- **A direct test of whether hand-engineered features are necessary:** a raw-signal baseline, trained on minimally processed sensor output with no window-level summary statistics, to establish whether the engineered feature pipeline used throughout this thesis contributes meaningfully beyond what a sequence model can learn unaided (design in Section 5.2.6, results in Section 5.3.2).
- **A like-for-like comparison with XGBoost and a retrospective assessment of the model choice:** the deployed BiLSTM is compared with the Mixed XGBoost model of Chapter 4 on the same participants, and the choice of the more complex model is reviewed in Section 5.3.3.

- **An offline validation baseline against which Chapter 6's deployment results can be judged**, obtained under the same leave-one-participant-out protocol used throughout this thesis.

5.2 Methodology

5.2.1 Experimental Design Overview

This chapter does not test a research question directly; it builds the infrastructure RQ4 through RQ6 (Chapter 1) are tested against in Chapter 6, and its own evaluation is best described as a controlled ablation rather than a participant experiment. No new participant data collection took place: the training data is the combined offline dataset from Chapters 3 and 4.

Manipulated factors (analogous to independent variables). (1) Feature representation: hand-engineered, windowed summary features (the approach used throughout this thesis) versus minimally processed raw signal (Section 5.2.6), to test whether feature engineering is actually necessary. (2) Model architecture, held roughly constant in capacity but varied in structure: a compact unidirectional LSTM (raw-signal baseline) versus the attention-enhanced bidirectional LSTM used for deployment, tested on identical raw input as a secondary check on whether architecture alone can substitute for feature engineering. (3) Comparison model: the Mixed XGBoost model of Chapter 4, trained on the same 73 participants and 1,736 task levels, as a reference for the deployed model.

Dependent variable. As in Chapters 3 and 4, the classification target is the three-level subjective difficulty rating (*Too Easy*, *Just Right*, *Too Difficult*), now predicted from a sequence of four 15-second windows per task level rather than a single session-level aggregate. Model quality against this target is itself the outcome measure, reported as accuracy, macro-precision/recall/F1, and AUC-ROC/AUPRC under Leave-One-Participant-Out cross-validation. Training against the participant's own label keeps the

model comparable with Chapters 3 and 4. It also lets the participant define what *Too Easy* and *Just Right* mean. It inherits the weaknesses of that label, which Chapter 7 examines. Accuracy against this label is therefore read alongside the error analysis below and, in Chapter 6, alongside outcomes that do not depend on any rating.

Measures. The same eye-tracking and physiological feature families as Chapters 3 and 4 (Section 5.2.3), recomputed per 15-second window rather than per session, plus the minimally processed raw channels described in Section 5.2.6 for the ablation.

The Results section below reports the windowed, handcrafted-feature model's offline performance against this design first, establishing the baseline Chapter 6 deploys. The raw-signal baseline and the comparison with XGBoost follow it (Section 5.3.2). A note on the choice of the deployed model closes the Results (Section 5.3.3).

5.2.2 Training Data

The model described in this chapter was trained on the combined offline dataset from the two preceding empirical chapters: the younger-adult sample from Chapter 3 and the older-adult sample from Chapter 4, together comprising 73 participants (54 younger, 19 older) after the exclusions applied separately in each chapter. This is a fixed-difficulty, single-task dataset in both cases; no data from the real-time deployment described in Chapter 6 was available at the time this model was trained, since a real-time-deployable model can, by construction, only be trained on data collected before deployment.

Sample and feature count. Each task level contributes one input sequence of $T = 4$ windows, each window described by $F = 28$ features (Section 5.2.3 below), giving an input tensor of shape 4×28 per level. Across the 73 participants (54 younger, 19 older), this yields 1,736 training sequences, the same underlying task-level population used by the raw-signal baseline in Section 5.2.6 below, differing only in feature representation (engineered per-window features here versus minimally processed raw channels there).

5.2.3 Feature Engineering

Eye-tracking and Physiological Data Preprocessing

The pipeline for eye-tracking and physiological signal preprocessing was analogous to the one detailed in Chapter 3.

Temporal Windowing and Feature Aggregation

To capture temporal dynamics while maintaining computational efficiency for real-time processing, data from each 60-second experimental level was divided into four sequential 15-second windows, as in Chapter 3 (1,325 raw samples per window at 90 Hz for eye-tracking and 765 samples per window for physiological data at 51Hz) [190]. This windowing approach enables the model to learn temporal patterns in cognitive load changes throughout each task level while providing sufficient temporal resolution for real-time applications [146].

The justification for these windowing parameters is given in Chapter 3.

For each window, 21 eye-tracking features and 7 physiological features were extracted, resulting in 28 features per window. The complete feature set included:

Eye-tracking features per window: blink count, blink rate, fixation count, mean/std fixation duration, saccade count, mean/std saccade amplitude, saccade-fixation ratio, advanced saccade metrics (peak velocity, acceleration/deceleration characteristics), and comprehensive pupil measures (mean, std, range, slope, constriction/dilation velocities).

Physiological features per window: GSR metrics (mean, std, slope) and heart rate variability metrics (BPM, SDNN, RMSSD, PNN50).

As opposed to earlier Chapters, the temporal metrics such as mean, std, slope for the entire level were not calculated and instead the model was expected to learn those representations itself based on the 4 windows presented to it per level in a sequence.

5.2.4 Real-time Multi-modal Cognitive Load Estimation Model

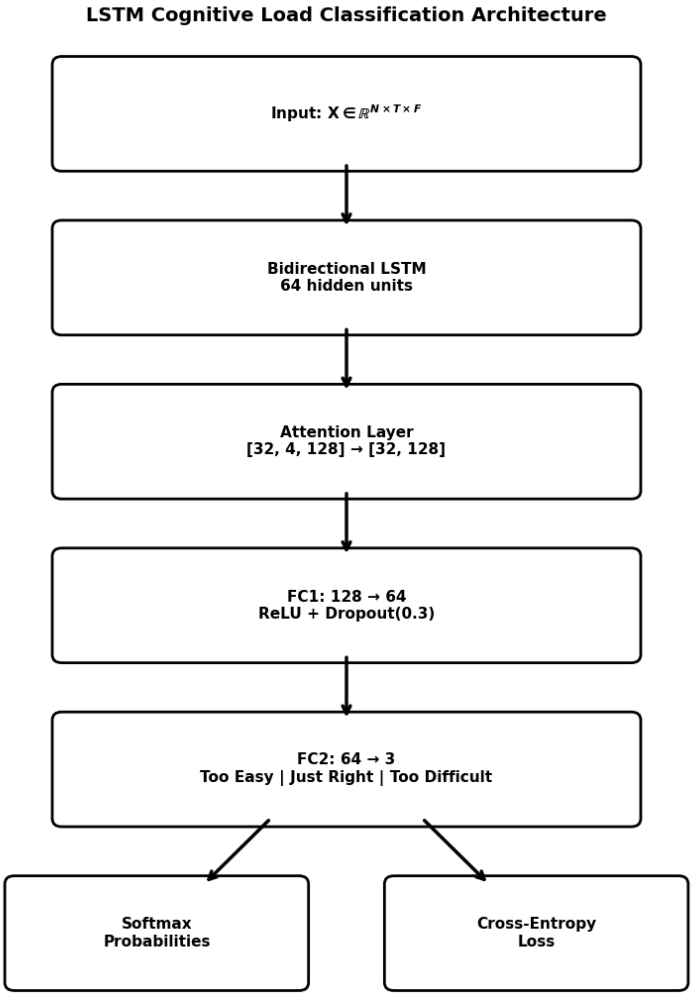


Figure 5.1: Diagram of the LSTM classifier architecture.

Attention-Enhanced Bidirectional LSTM Architecture

The cognitive load estimation model employed a bidirectional Long Short-Term Memory (LSTM) architecture with attention mechanisms, specifically designed to capture temporal dependencies in physiological and eye-tracking signals [127], see Figure 5.1. The model architecture consisted of several key components optimized for cognitive load classification from multimodal physiological data.

Input Layer Configuration: The model processed input tensors $\mathbf{X} \in \mathbb{R}^{N \times T \times F}$, where N represents the batch size, $T = 4$ denotes the temporal sequence length, and $F = 28$ represents the dimensionality of the concatenated feature vector per time step. The feature space comprised 21 oculometric parameters and 7 physiological markers extracted through standardized preprocessing pipelines to ensure temporal alignment and feature normalization across modalities.

Bidirectional LSTM Processing: A bidirectional LSTM with 64 hidden units processed the input sequences, following prior work on physiological signal classification [10]. Bidirectionality enabled the model to capture both forward and backward temporal dependencies, useful for capturing cognitive load patterns that may manifest with temporal delays or anticipatory effects. The network used hyperbolic tangent for state activation and sigmoid for gate activation.

Attention Mechanism: The four windows of a task level are not equally informative, and averaging them, or relying on the final hidden state alone, would treat them alike. A self-attention layer was therefore added to learn a weight for each window from the training data [193]. The attention mechanism computed contextual importance through:

$$\text{attention_weights} = \text{softmax}(W_a \cdot h_t + b_a) \tag{5.1}$$

$$\text{context_vector} = \sum_t (\text{attention_weights}_t \odot h_t) \tag{5.2}$$

where h_t represents the LSTM hidden state at time t, W_a, b_a are learnable attention parameters, $\cdot$ denotes matrix-wise multiplication and $\odot$ denotes element-wise multiplication. In words, each window receives a score, the scores are normalised across the windows, and the context vector is the score-weighted combination of the window representations. This self-attention mechanism enabled the model to focus on the most informative segments while maintaining comprehensive feature integration across the entire sequence of eye-tracking and physiological data.

Classification Head: The final classification layer consisted of a two-layer fully connected network with ReLU activation and dropout regularization (0.3). The network mapped the attention-weighted context vector to class probabilities for three difficulty categories: "Too Easy," "Just Right," and "Too Difficult" based on a softmax activation.

Class Imbalance Mitigation: To address class imbalance in cognitive load datasets, a multi-stage augmentation pipeline was implemented. Initial class balancing was achieved through participant-wise median target count normalisation, followed by synthetic sample generation. The augmentation pipeline comprised three complementary strategies: (1) **temporal jittering** with Gaussian noise injection ($\sigma = 0.05$) to improve model robustness to temporal variations. (2) **Intra-class window mixing** where temporal segments from different sequences within the same 'difficulty' class were combined (e.g., concatenating windows 1-2 from participant A with windows 3-4 from participant B, both labeled "Low" difficulty), (3) **cross-class interpolation** using weighted linear combinations of "Low" and "High" difficulty samples to generate synthetic "Medium" difficulty sequences with interpolated weight of $\alpha \in [0.45 - 0.55]$. To improve the classification of 'medium' level difficult, we employed **structured temporal pattern synthesis**

using predefined arrangements: ($Medium_1 \rightarrow Low_3 \rightarrow High_2 \rightarrow Medium_4$ and $Medium_4 \rightarrow High_3 \rightarrow Low_2 \rightarrow Medium_1$), where subscripts indicate temporal position within the original sequence. This approach captured transitional cognitive load states characteristic of intermediate difficulty levels.

Dynamic Class Weighting: Class weights were computed based on the inverse frequency scaling with manual adjustments to enhance edge case detection. Applied weights were 1.5× for *Too Easy*, 1.5× for *Just Right*, and 1.7× for *Too Difficult* categories, multiplied by their respective inverse frequency weights to address challenging classification boundaries [2].

5.2.5 Model Training Configuration

Key training parameters included:

- **Loss Function**: Weighted categorical cross-entropy for standard classification
- **Optimizer**: Adam optimizer (learning rate = 0.001, weight decay = 0.01)
- **Regularization**: Dropout (0.3) and early stopping (patience = 15 epochs) to prevent overfitting
- **Batch Size**: 32, optimized for memory constraints and convergence stability
- **Maximum Epochs**: 100 with early stopping criteria based on validation accuracy and loss
- **Network Architecture**: Single bidirectional LSTM layer with 64 hidden units
- **Trainable Parameters**: 56,708 in total, made up of the bidirectional LSTM layer (48,128), the attention scoring layer (129), the first fully connected layer (8,256) and the output layer (195)

- **Data Augmentation**: Multi-strategy approach including Gaussian jitter ($\sigma = 0.05$), temporal window mixing, and cross-class interpolation generating up to 380 additional synthetic samples per class
- **Feature Preprocessing**: Standard normalization (z-score) applied to all input features fitted on training data only
- **Validation Split**: 10% stratified split from training data for model selection and early stopping
- **Training Strategy**: Edge case prioritization with weighted loss to improve detection of *Too Easy* and *Too Difficult* classifications over *Just Right* categories

5.2.6 Cross-Validation and Generalization Assessment

Leave-One-Participant-Out Protocol: The model was trained using data from 73 participants (54 younger, 19 older) collected in the initial model development phase. Training employed leave-one-participant-out (LOPO) cross-validation to assess generalization across participants [91]. This validation approach is widely recommended for physiological signal classification as it prevents data leakage between training and testing sets while providing realistic estimates of cross-participant generalization performance [176].

Nested Validation Protocol: Within each LOPO fold, the training set (72 participants) was further partitioned using stratified 90/10 training-validation splits to enable unbiased hyperparameter selection and early stopping. This nested validation approach ensured unbiased model selection within each fold while providing robust generalization estimates across the participant population.

Feature Scaling and Normalization

All features were standardized with a z-score scaler fitted on the training data only, without per-user calibration, following established protocols for cross-participant physiological models [25]. The same scaler was applied during real-time deployment to ensure feature distributions matched those seen during training. Under this zero-calibration approach, baseline differences between participants in heart rate, pupil size and skin conductance can affect model performance (see Chapter 6).

Raw-Signal Baseline (No Handcrafted Features)

As discussed later in this chapter, the LSTM model trained on the features processed as described above is likely too complex for this simple dataset and therefore leaves a lot of performance capacity unused. To test whether the hand-engineered temporal features (Section 5.2.3) are necessary for cognitive load classification, or whether a recurrent model could instead learn directly from minimally-processed sensor output, a number of potential configurations was trained on raw, per-sample physiological and eye-tracking signal. Due to the number of configurations attempted, the evaluation used 5-fold group cross-validation, grouped by participant, so that no participant's sessions appeared in both the training and held-out sets of any fold. This is likely to inflate the results slightly compared to LOPO-CV as described for the model trained on preprocessed features. However, trying all the configurations on LOPO-CV would be extremely time and computationally consuming, especially when using raw data. Moreover, the purpose of this analysis is to find the potential performance ceiling of the raw data model, and 5-fold group cross-validation is likely to achieve that.

Data was drawn exclusively from the original (offline) task sessions across the full participant population ($N = 73$: 54 younger and 19 older adults; 1,736 task levels), same population as the model described above. For each task level, the 13 available raw channels (pupil diameter, combined and per-eye; eyelid openness, per-eye; 3D gaze direction

vectors, per-eye; galvanic skin response; and heart rate) were extracted at native sampling rate (90 Hz for eye-tracking, 5,400 samples over 60 s per level, with physiological metrics upsampled to match the frequency). They were then resampled via linear interpolation to a fixed length of a number timesteps (tested a range of 50-800, as well as not resampling at all). This resampling was the only transformation applied to the signal; no window-level summary statistics of the kind used elsewhere in this thesis (e.g. blink count, mean pupil diameter, temporal slopes or ranges) were computed at any stage. Multiple architecture variants were tested, including a single layer unidirectional LSTM, bidirectional LSTM, adding another layer to both, adding self-attention, a sweep of hidden states, network depth, dropout, as well as various grid combinations of dataset balancing (upsampling minority classes, downsampling majority classes, class weighting).

The results are reported in Section 5.3.2.

5.3 Results

Consistent with the Experimental Design Overview above, the windowed, handcrafted-feature model is evaluated first against the subjective-difficulty target under LOPO cross-validation. The raw-signal baseline (Section 5.3.2) then shows how much of this performance depends on the feature representation rather than the architecture. The comparison with XGBoost places it against a simpler alternative.

5.3.1 Model Evaluation (pre-deployment)

Model evaluation used held-out participants and participant-wise analysis with aggregated cross-validation results. The evaluation protocol employed leave-one-participant-out cross-validation to give generalization estimates across the participant population while preventing data leakage between training and testing phases.

Table 5.1: Overall Metrics Averaged Across LOPO-CV Folds (pre-deployment)

Metric	**Mean**	**Std**	**Min**	**Max**
Accuracy	52.3%	14.5%	16.7%	83.3%
F1-Score (Macro-Averaged)	49.7%	16.2%	13.7%	80.8%
Precision (Macro-Averaged)	58.9%	17.4%	15.2%	87.5%
Recall (Macro-Averaged)	54.0%	15%	20%	90%
AUC-ROC (Macro-Averaged)	73.7%	10.7%	47%	98%
AUPRC (Macro-Averaged)	62.5%	12.3%	35.8%	97.6%

Classification Performance: Detailed performance metrics are summarized in Table 5.1 and Figure 5.2. Table 5.1 includes accuracy, precision, recall, AUC-ROC, AUPRC and F1-scores. Figure 5.2 presents performance metrics including a confusion matrix (Figure 5.2a) showing classification accuracy across difficulty levels and ROC curves (Figure 5.2b) demonstrating discriminative power for each class with corresponding area-under-curve values. Mean accuracy was 52.3% and macro-F1 49.7% for three classes, with fold accuracy ranging from 16.7% to 83.3%, so offline performance was not uniform across participants. How much these figures matter depends on which errors the model makes, since difficulty in the adaptive system changes in small steps after every level (Chapter 6: 0.3 s after a *Too Easy* or *Too Difficult* rating, 0.1 s after *Just Right*). Difficulty is then re-estimated from new data before the next level. Predicting *Too Easy* or *Too Difficult* for a level the participant rated *Just Right* produces a single small step away from the *Just Right* band, which later estimates can offset, so this error carries little cost in this system. Predicting *Just Right* for a *Too Easy* level still moves difficulty in the right direction by a smaller step, whereas predicting *Just Right* for a *Too Difficult* level leaves that step uncorrected. The costly errors are those between the two edge cases, which move difficulty in the wrong direction. These made up 8.81% of all predictions (*Too Difficult* predicted as *Too Easy* in 71 cases and *Too Easy* predicted as *Too Difficult* in 84 cases; Figure 5.2a). This offline result is the baseline for the deployment results in Chapter 6, where the success of the model-driven condition is judged by whether outcomes are equivalent to those of the participant-driven condition.

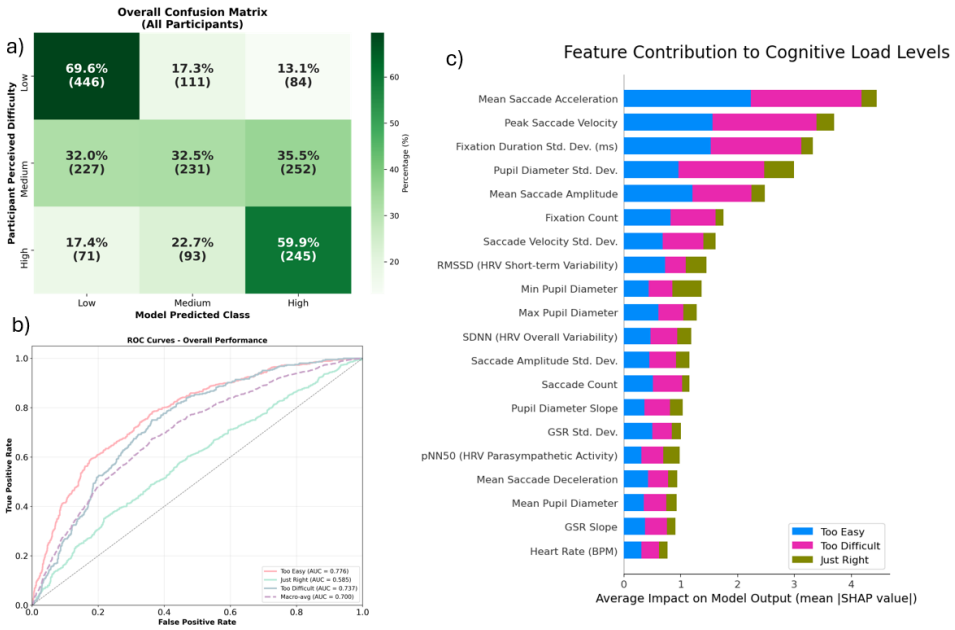


Figure 5.2: Confusion matrix, ROC curves and feature importance rating for the offline (pre-deployment) model.

Feature Importance Analysis: The SHAP analysis of feature importance revealed that oculometric measures dominated model predictions, with the top 7 contributing features derived exclusively from eye-tracking data (Figure 5.2c). The most informative features were primarily saccadic metrics (specifically mean acceleration, peak velocity, and mean amplitude) and fixation-related parameters (standard deviation of duration and count), alongside pupil diameter variability. Physiological measures played a secondary role, with short-term Heart Rate Variability (RMSSD) appearing as the eighth most important feature, while blink frequency patterns were notably absent from the top contributions. This is consistent with the dominance of oculometric features already established in Chapter 3.

5.3.2 Raw-Signal Baseline and Comparison with XGBoost

The raw-signal baseline (Section 5.2.6) performed poorly. The best performing model was a single-layer unidirectional LSTM with 64 hidden units, a dropout layer ($p = 0.3$), and a single linear output layer over the final hidden state (20,419 trainable parameters). The network was trained with inverse-class-frequency-weighted categorical cross-entropy,

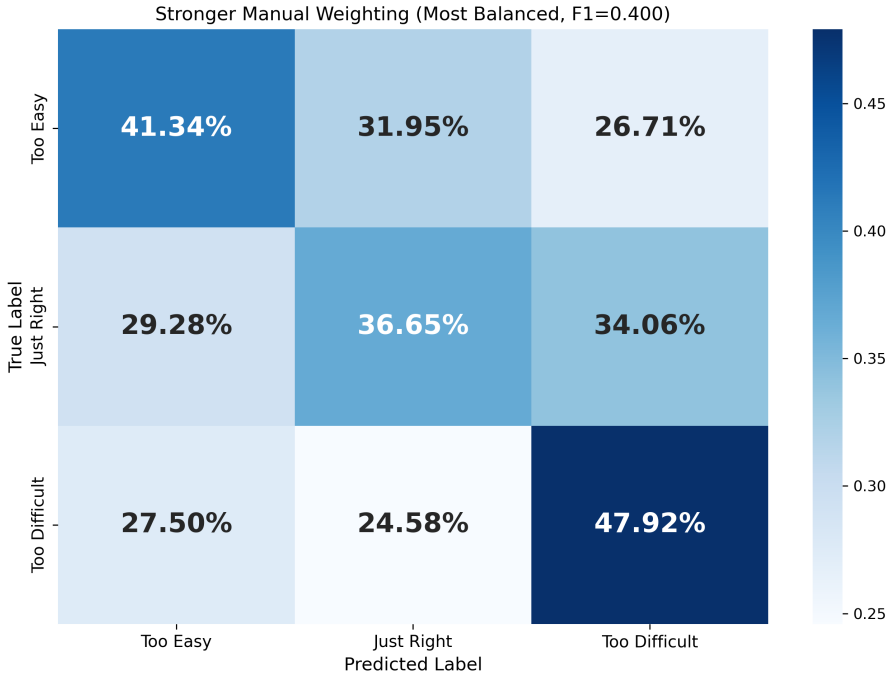


Figure 5.3: Confusion matrix of the balanced raw data model (F1 = 0.4)

the Adam optimizer (learning rate = 0.001, weight decay = 1×10^{-5}), batch size 32, and early stopping (patience = 10, maximum 200 epochs). Trained on minimally processed signal (13 channels resampled to 200 timesteps, 1,736 task levels from 73 participants), the model reached an out-of-fold accuracy of 0.43 and macro-F1 of 0.42. Macro-precision was 0.42, macro-recall 0.43 and AUC 0.58, against a chance level of about 0.33 for three classes. The other configurations explored did not change this picture. Across every variant tested, including the bidirectional LSTM with self-attention used for deployment, macro-F1 stayed between about 0.34 and 0.45. No combination of architecture, capacity or temporal resolution closed the gap to the models built on engineered features. However, the accuracy was derived from a largely *Too Easy* prediction. While none of the balancing techniques improved accuracy beyond baseline, the most balanced model achieved F1 score of .40 and a confusion matrix seen in Figure 5.3 (41.34% *Too Easy*, 47.92% *Too Difficult*), which is the most comparable result to the model deployed in Chapter 6, because if a model trained on raw data were to be deployed, it would have been balanced the same way the edge case prioritisation is handled there.

Table 5.2 places the raw-signal models next to the models built on engineered features, all using the same 73 participants and 1,736 task levels. The windowed BiLSTM used for deployment and the Mixed XGBoost model of Chapter 4 (30 session-level features, leave-one-participant-out validation) were close. The BiLSTM reached the higher macro-F1 (0.497 against 0.466), while accuracy and AUC were slightly lower (0.523 against 0.537, and 0.700 against 0.724). Both are clearly above the raw-signal models. The differences between the two feature-based models appear small relative to the spread across participants (standard deviations of 0.12 to 0.16 in Chapter 4). Two caveats apply: the raw-signal models were evaluated with 5-fold participant-grouped cross-validation while the other two used leave-one-participant-out validation, and the AUC of the BiLSTM comes from pooled predictions while that of XGBoost is the mean of per-participant values.

Table 5.2: Models trained on the 73-participant offline dataset (1,736 task levels). Accuracy and macro-F1 for the two feature-based models are means across held-out participants; the raw-signal results are out-of-fold.

Model	**Input per level**	**Validation**	**Accuracy**	**Macro-F1**	**AUC**
Raw Data LSTM	13 channels × 200 timesteps (2,600 values)	5-fold, grouped	0.43	0.42	0.58
Deployed LSTM	4 windows × 28 features (112 values)	LOPO	0.523	0.497	0.70
XGBoost (Chapter 4)	30 session-level features	LOPO	0.537	0.466	0.72

5.3.3 Choice of the Deployed Model

The deployed model was a bidirectional LSTM with self-attention. On the same 73 participants it reached a higher macro-F1 than the Mixed XGBoost model (0.497 against 0.466; Table 5.2), with slightly lower accuracy and AUC. It also discriminated the *Too Easy* and *Too Difficult* classes from the rest well (one-versus-rest AUC of 0.78 and 0.74, against 0.59 for *Just Right*; Figure 5.2b), which is the property the adaptation rule of Chapter 6 depends on most. When the real-time experiment had to be deployed, the advantage of the LSTM over the XGBoost models available at that stage was slightly

larger than it is in the final version of this thesis. The better-performing XGBoost models reported in Chapters 3 and 4 were trained later, with improved feature selection among other refinements. The LSTM was therefore the stronger candidate at the time, and the real-time experiment cannot be repeated with a different model.

The network is large relative to the data, and this is acknowledged as a shortcoming. It has 56,708 trainable parameters, roughly 37 for each of the 1,540 or so real training sequences in a leave-one-participant-out fold (up to about 2,700 with augmentation). Several measures constrain this capacity: dropout (0.3), weight decay (0.01), early stopping on a nested 10% validation split, and evaluation on held-out participants throughout. Most of the parameters (48,128) sit in a recurrent layer that processes only four windows per level. The margin over XGBoost was modest, and the raw-signal comparison above suggests that added capacity did not substitute for feature engineering. A simpler model, such as XGBoost with the improved feature set, might therefore offer a comparable trade-off at lower cost in training, maintenance and reproducibility. The implications are taken up in Chapters 8 and 9.

5.4 Summary

This chapter built a windowed, sequence-based cognitive load estimation model intended for real-time use, and evaluated it offline before any live deployment was attempted. Three findings shape how the remainder of this thesis uses this model. First, offline accuracy (52.3% mean, macro-F1 49.7%) varied across participants (fold accuracy ranged from 16.7% to 83.3%). Most errors were between *Just Right* and one of the edge classes, which the small-step adaptation rule tolerates, while errors between the two edge cases accounted for 8.81% of predictions. The LOPO performance is therefore best read as an average over a wide spread. Second, the raw-signal baseline (Section 5.3.2) performed poorly (macro-F1 about 0.42, against about 0.50 for the windowed model), and no variant tried managed to close that gap. The engineered features carried over from Chapters 3 and 4 are therefore still doing useful work and were not made redundant by a more expressive sequence model.

Third, on the same 73 participants the windowed BiLSTM reached a higher macro-F1 than the Mixed XGBoost model of Chapter 4 (0.497 against 0.466), with slightly lower accuracy and AUC. The margin was slightly larger at the time of deployment, which is why the LSTM was chosen for the real-time experiment. The LSTM also discriminated the *Too Easy* and *Too Difficult* classes well, which the adaptation rule depends on most. The network is large relative to the data (56,708 parameters), and a simpler model might offer a comparable trade-off at lower cost. This is acknowledged as a shortcoming (Section 5.3.3).

Chapter 6 takes this model, built and validated here entirely on offline, fixed-difficulty data. It deploys the model into a live, closed-loop VR system where difficulty is adjusted in real time after every 60-second level, addressing RQ4 through RQ6. That chapter evaluates the model by whether model-driven and participant-driven control lead to equivalent outcomes, rather than by offline accuracy alone.

Chapter 6

Real-Time Deployment of an Adaptive Cognitive Training System in Virtual Reality

6.1 Introduction

Chapter 5 built and offline-validated an attention-enhanced bidirectional LSTM capable of estimating cognitive load from a continuous stream of eye-tracking and physiological data. It reported an offline accuracy of 52.3% on average, varying across participants. This chapter deploys that model into a live, closed-loop VR system and tests what it actually does once the offline-to-real-time gap is no longer hypothetical. It addresses RQ4, RQ5 and RQ6, introduced in Chapter 1. RQ4 asks whether a model trained on single-task data generalises to a dual-task context introduced specifically to test this. RQ5 asks whether a model-driven adaptive setting is a suitable replacement for a participant-driven one. RQ6 asks whether real-time adjustment after each level changes participant difficulty judgments relative to a fixed design.

Chapter 2 (Section 2.7) discusses at length the distinction between reactive, performance-triggered adaptation and proactive, physiologically-informed adaptation. It also discusses the broader gap between systems that react to performance or affect and systems that estimate intrinsic cognitive load. That review found that the cognitive load datasets and models it covered were almost all evaluated offline. It also found that adaptive VR sys-

tems are typically driven by simple heuristics, affect or extraneous load rather than an estimate of intrinsic cognitive load. This chapter is a direct test of a system that attempts the latter, using the model built in Chapter 5, and it reports that test's results without assuming in advance that offline validation guarantees live performance.

6.1.1 Rationale and Contributions

This chapter makes the following contributions. First, it reports what is, to my knowledge, among the first live deployments of a model-driven adaptive cognitive load estimation system in an interactive VR environment using eye-tracking and physiological signals only, rather than performance heuristics or affect. Second, it evaluates the real-time model-driven design directly against a participant-driven design running in the same live system, rather than against an offline benchmark. Third, it compares a dual-task paradigm (sustained attention plus mental arithmetic) against the single-task paradigm the model was trained on, within a single deployment study, providing a direct test of RQ4 rather than an offline proxy for it.

6.2 Methodology

6.2.1 Experimental Design Overview

This chapter tests RQ4, RQ5, and RQ6 (Chapter 1), through hypotheses H4, H5 and H6 respectively. It uses a 2×2 within-subjects, repeated-measures design, with every participant experiencing all four combinations of the two manipulated factors below.

Independent variables. (1) Control Type: Model-driven versus Self (participant)-driven difficulty adjustment, directly testing RQ5. (2) Task Complexity: Single-task (sustained attention only) versus Dual-task (sustained attention plus mental arithmetic), directly testing RQ4's question of generalisation from the single-task training data. Difficulty itself was adjusted incrementally after every level within each condition (real-time adjustment vs. the fixed, discrete levels of Chapters 3 and 4), which is what RQ6 compares.

Dependent variables. Objective: raw game performance, difficulty-weighted game performance, arithmetic error rate (dual-task only), and final difficulty level reached. Subjective: NASA-TLX subscales (Total Workload, Mental Demand, Physical Demand, Temporal Demand, Effort, Frustration, Performance) and the categorical difficulty rating (*Too Easy*, *Just Right*, *Too Difficult*) used identically to Chapters 3 and 4. Model-specific: real-time classification accuracy, precision, recall, F1, and AUC against the same subjective-difficulty target, computed after every level throughout each session. Two kinds of reference are therefore used to judge the model-driven condition. One is the participant's own rating after each level. The other is outcomes that do not depend on any rating (game performance, arithmetic errors and final difficulty). The NASA-TLX is collected at the end of each session to compare conditions. Using both means the conclusions do not rest on a single ground truth (Chapter 2), and the reliability of the per-level rating is examined in Chapter 7.

Measures. The same eye-tracking and physiological feature pipeline as Chapter 5 (windowed, real-time-computed), NASA-TLX subscales for subjective workload, and the game engine's own logs for raw and difficulty-weighted performance and arithmetic accuracy.

The Results section below is organised around these variables in turn. Model classification performance against the subjective-difficulty target is reported first - this is the mapping of the model predictions over those same judgments made by the participants throghout each session. The manipulation check on subjective difficulty ratings follows, then the Control Type × Task Complexity effects on game performance, difficulty convergence, NASA-TLX workload and learning/fatigue. A dedicated analysis of individual differences in model performance comes last.

6.2.2 Participants

We recruited 54 participants aged 18-59 ($M = 28.5, SD = 8.4$) of whom 30 were female, 22 male and 2 non-binary. Three of the participants were excluded due to issues with the real-time eye-tracking streaming, leading to invalid model predictions. The final sample included 51 participants aged 18-59 ($M = 28.6, SD = 8.6$), with 27 being female, 22 male, 2 non-binary. Of those, 49 were right-handed, 2 left-handed. Participants were recruited through online advertising and offered a £10 Amazon voucher for their participation. This real-time cohort forms part of the same publicly available dataset introduced in Chapter 3, and the real-time deployment software described in this chapter, alongside the VR application code, is publicly available in the same code repository. Ethical approval was granted by the University of Glasgow, College of Science and Engineering, School of Computing Science (application number 300230006).

6.2.3 Experimental Design - Framework for Real-time Model Adaptation

Following the first two rounds of data collection aimed at developing a model that can be deployed in real-time (Chapters 3 and 4), we decided to make some modifications to the experimental design. Firstly, the difficulty became adaptive instead of fixed. This means that participants always started at 2.6s difficulty (lower medium level from model devel-

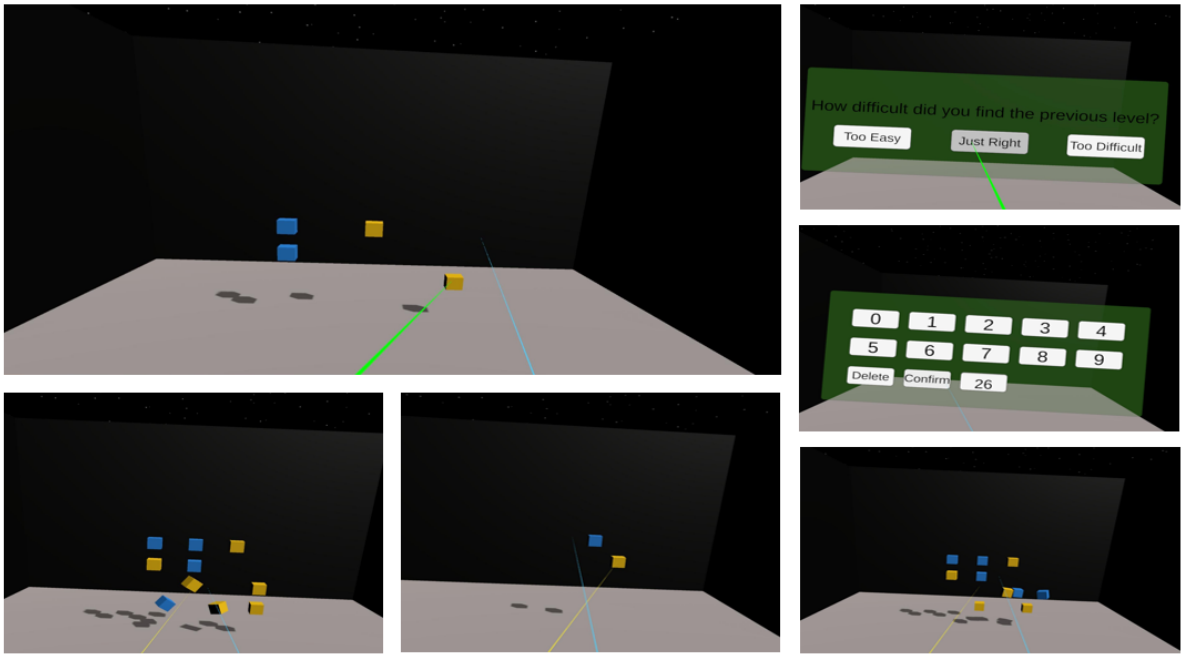


Figure 6.1: In-game view at different difficulty levels including mental arithmetic menu (middle right) and perceived difficulty selection menu (top right).

opment stage) and the difficulty was then guided by the model or participant, depending on condition. If the difficulty selection was *Too Easy* the task would become more difficult by a factor of 0.3s. Conversely, if the difficulty selection was *Too Difficult* the task would become easier by a factor of 0.3s. If the difficulty selection was *Just Right*, it became harder by a factor of 0.1s to avoid stale difficulty throughout a session. As a reminder the difficulty value refers to the time interval between two cubes of the same colour appearing (e.g. 2.6s indicates that between two blue cubes spawning there is exactly 2.6s).

Secondly, we added a dual task element to the design for two reasons. a) It allows difficulty to increase independently of physical dexterity, which was a risk at higher difficulty levels in the single task design: if the cubes appeared too quickly, participants could struggle to physically move their wrists fast enough to destroy them. b) It tests how well the model generalises to different types of tasks than the one it was trained on. Therefore, we added a simple mental arithmetic task, which involved 5 numbers between 1 and 9 being played (audio) throughout each 60s level at 10s intervals, starting 10s after level start. The participants were asked to provide the sum of the numbers at the end of each level using a keypad integrated into the VR experiment. This was followed by a menu presenting the difficulty choice (*Too Easy, Just Right, Too Difficult*) (see Figure

6.1 for both menus). Following the difficulty choice, participants were given another 5s break before starting the next level. The participants played 8 levels in each condition, for a total of 4 conditions - two single task conditions, one controlled by the model, one by the participant, and two dual task conditions, one controlled by the model, one by the participant. The order in which the conditions were presented was randomised. However, single task was always in spots 1 and 3, and dual task in spots 2 and 4. This was to allow the participants to familiarise themselves with the experiment and to allow equal spacing between single/dual task conditions. The four random orders were: 1. model-guided single task, model-guided dual task, self-guided single task, self-guided dual task, 2. self-guided single task, self-guided dual task, model-guided single task, model-guided dual task, 3. self-guided single task, model-guided dual task, model-guided single task, self-guided dual task, 4. model-guided single-task, self-guided dual task, self-guided single task, model-guided dual task.

For a visual breakdown of the 4 conditions see figure 6.2. The conditions were counter-balanced with 13 participants in all conditions except for condition 3 which had 15 participants. After exclusions order 1 had 13 participants, order 2 - 11 participants, order 3 - 15 participants and order 4 - 12 participants. The rest of the task remained the same as in the first phase of the experiment (model development). Participants were asked to complete a NASA-TLX questionnaire at the end of each condition during the 5 minute breaks (6 subscales per administration; 4 conditions per participant, so $54 \times 4 = 216$ administrations as delivered). This uses the full recruited sample of 54, since the 3 participants later excluded from the model-performance analyses were excluded for real-time eye-tracking streaming issues rather than incomplete NASA-TLX data.

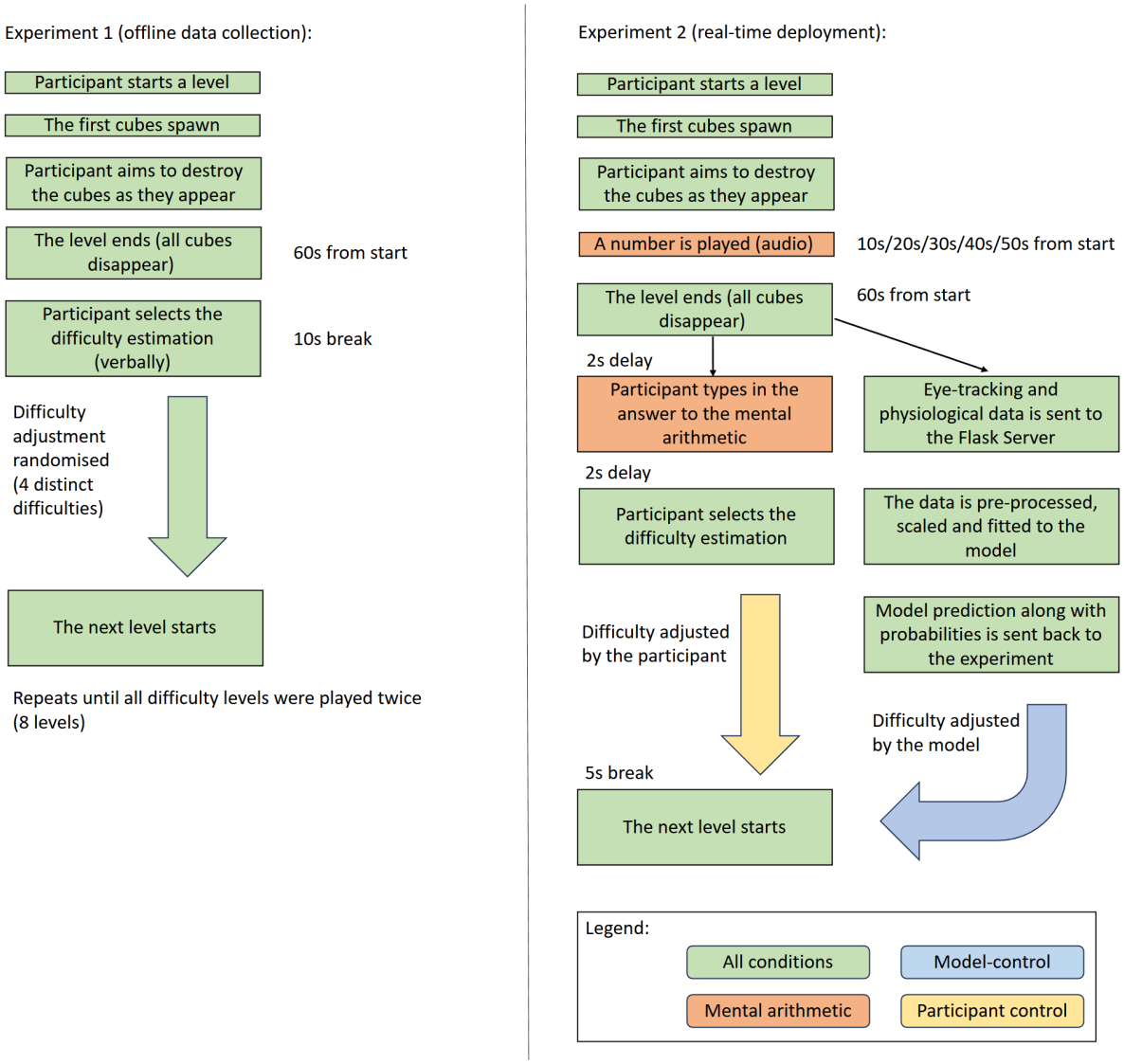


Figure 6.2: The flowchart depicting the experimental flow of the adaptive experimental design compared to the original flow used in Chapters 3 and 4.

6.2.4 Materials

The study was carried out using an HTC VIVE PRO EYE headset featuring an integrated eye-tracker. The headset was linked to a desktop computer equipped with an NVIDIA RTX 2070 SUPER graphics card, AMD Ryzen 5 3600 processor and 16GB of RAM. Raw eye-tracking metrics were gathered via the HTC Vive Pro Eye headset at a 90 Hz frequency through a free Unity plugin. These encompassed eye openness, pupil diameter, sensor position, gaze origin and direction, and convergence distance. Physiological data, comprising photoplethysmography (PPG) and galvanic skin response (GSR), were acquired with a wrist-based Shimmer sensor at 51Hz and handled using the pyshimmer [109] library in Python. A data collection pipeline was implemented to synchronize timestamps for eye-tracking data, physiological data, and in-task events via LabStreamingLayer [157]. Participants filled out a standard NASA-TLX questionnaire [66] after each condition.

6.2.5 Real-time Deployment for VR integration

The trained cognitive load estimation model was deployed using a distributed microservice architecture built on Flask, enabling streaming and processing of eye-tracking and physiological data in real-time [2]. No new model training took place in this chapter: the model deployed here is the identical, already-trained artefact described in Chapter 5 (training set and feature count reported there), used only for inference on newly streamed, real-time data. The deployment pipeline achieved sub 500ms end-to-end latency, from data acquisition to cognitive load estimation, suitable for responsive adaptive VR applications.

Data Stream Management and Synchronization: Multimodal data acquisition was orchestrated through LabStreamingLayer (LSL) protocol [157], which handled temporal synchronization between heterogeneous sensor streams operating at different sampling rates.

Real-time Feature Extraction Pipeline: The preprocessing and feature extraction methods used during training were applied to streaming data, maintaining consistency between training and deployment conditions. This includes identical filtering parameters, windowing strategies, and normalization procedures.

Model Inference Engine: PyTorch models were loaded with pre-trained weights and executed on incoming feature sequences to generate real-time predictions with class probabilities. The system maintained model state across inference calls to ensure temporal continuity.

Adaptive VR Integration: Model predictions were integrated with the VR task difficulty adjustment system, enabling automatic difficulty modulation based on estimated cognitive load levels.

This real-time architecture enabled cognitive load estimation with minimal latency while maintaining the robustness necessary for continuous operation in interactive VR environments.

6.3 Results

Following the Experimental Design Overview above, model classification performance against the subjective-difficulty target is reported first, since it bears on how much weight to place on every result that follows. The remaining subsections then work through the dependent variables in the order set out there: subjective ratings, game performance and difficulty convergence by Control Type and Task Complexity, NASA-TLX workload, learning and fatigue, and individual differences.

6.3.1 Evaluation of Physiologically-Informed Adaptive Difficulty Prediction in Virtual Reality (Analysis of the Classifier Performance During Deployment)

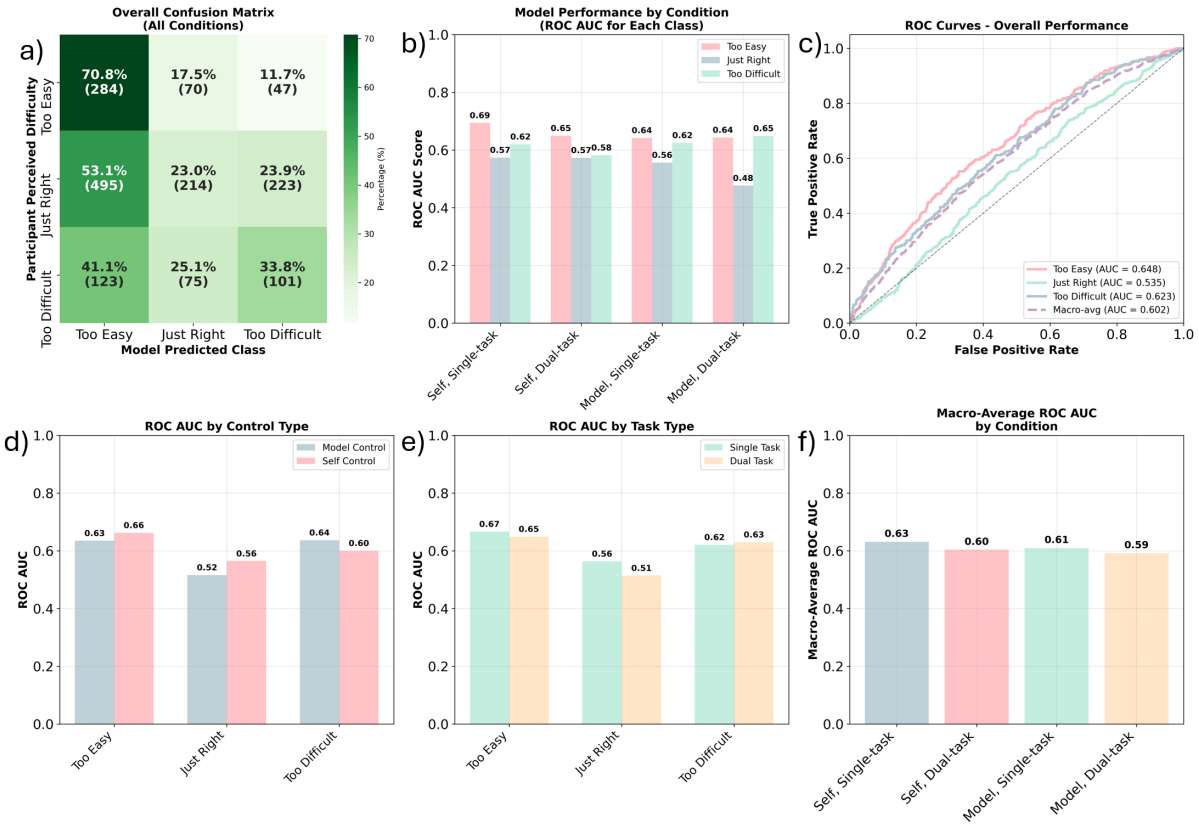


Figure 6.3: Complete model evaluation across all experimental conditions. (a) Overall confusion matrix showing prediction accuracy with percentages and counts, (b) Model performance by condition showing ROC AUC scores for each difficulty class, (c) ROC curves for overall performance, (d) ROC AUC comparison by control type, (e) ROC AUC comparison by task type, (f) Macro-average ROC AUC by individual condition.

For reference, Table 5.1 in Chapter 5 reports the corresponding offline, pre-deployment metrics for the same model.

Real-time classification performance was assessed through comprehensive analysis of model predictions against participant-reported difficulty ratings across all experimental conditions. The evaluation leveraged continuous model predictions throughout the experimental protocol, enabling comparison of 32 trials per participant against ground-truth subjective difficulty assessments, yielding 1,632 (32x51) total prediction-rating pairs for analysis.

Table 6.1: Overall Metrics Averaged Across All Real-Time Predictions (During Deployment)

Metric	**Mean**	**Std**	**Min**	**Max**
Accuracy	36.7%	11.9%	6.2%	59.4%
F1-Score (Macro-Averaged)	32.6%	10.0%	8.0%	56.0%
Precision (Macro-Averaged)	43.2%	11.3%	13.7%	72.0%
Recall (Macro-Averaged)	42.1%	12.5%	4.6%	65.7%
AUC-ROC (Macro-Averaged)	63.0%	10.7%	41.3%	82.0%
AUPRC (Macro-Averaged)	49.5%	11.6%	31.4%	80.4%

Overall Classification Performance

The deployed model reached 36.7% accuracy against participants' own ratings (macro AUC 0.63; Table 6.1), compared with 52.3% pre-deployment (Chapter 5). *Just Right* was the most frequent rating (932 of 1,632 trials, 57.1%) and the model predicted *Too Easy* for 55.3% of trials, most often for levels rated *Just Right* (30.3% of all trials; Figure 6.3a). As in Chapter 5, the errors that matter most for the system are those between the two edge cases, since they push difficulty in the wrong direction. These made up 10.4% of predictions (*Too Difficult* predicted as *Too Easy* in 123 trials and *Too Easy* predicted as *Too Difficult* in 47 trials). Predicting an edge case for a *Just Right* level changes difficulty by one small step that later predictions can offset, so per-trial agreement with self-report understates how the model works as a controller. Part of the gap between pre-deployment and real-time accuracy may also reflect the instability of participants' ratings under continuous, low-contrast conditions (difficulty steps of 0.1 to 0.3 s in real time, against 0.5 to 1.5 s offline), which Chapter 7 examines. Class-specific precision was 0.60 for *Just Right*, 0.31 for *Too Easy* and 0.27 for *Too Difficult*. The measure of success for the model-driven deployment is outcome equivalence between model-driven and participant-driven conditions, examined in the sections that follow.

Model performance varied modestly across experimental conditions (Figure 6.3b). Comparison between control types (Figure 6.3d) and task types (Figure 6.3e) were in line with these findings, with self-control conditions showing marginally better AUC values (0.610 vs. 0.597) and single task conditions outperforming dual task conditions (0.618 vs. 0.599). Individual condition analysis (Figure 6.3f) showed that Self Single Task achieved the highest macro-average AUC (0.631), while Model Dual Task showed the lowest performance (0.592).

6.3.2 Subjective Difficulty Analysis

To validate the efficacy of the experimental manipulation, we analysed the distribution of subjective difficulty ratings (*Too Easy, Just Right, Too Difficult*) across conditions. Figure 6.4 compares the rating distributions of the original baseline dataset against the Single and Dual Task conditions of the current study.

Perceived Difficulty Comparison Between Single and Dual Real-Time Designs

The introduction of the secondary auditory task successfully induced a shift in perceived difficulty. In the Single Task condition, 30.6% of segments were rated as *Too Easy*, whereas this dropped significantly to 17.6% in the Dual Task condition ($\chi^2 = 158.70, p < 0.001$). Conversely, reports of *Too Difficult* ratings rose from 17.3% to 20.9% ($\chi^2 = 14.65, p < 0.001$). An omnibus Chi-square test confirmed that the overall distributional shift was statistically significant ($\chi^2 = 159.58, p < 0.001$), indicating that the dual-task protocol effectively increased cognitive load as perceived by participants.

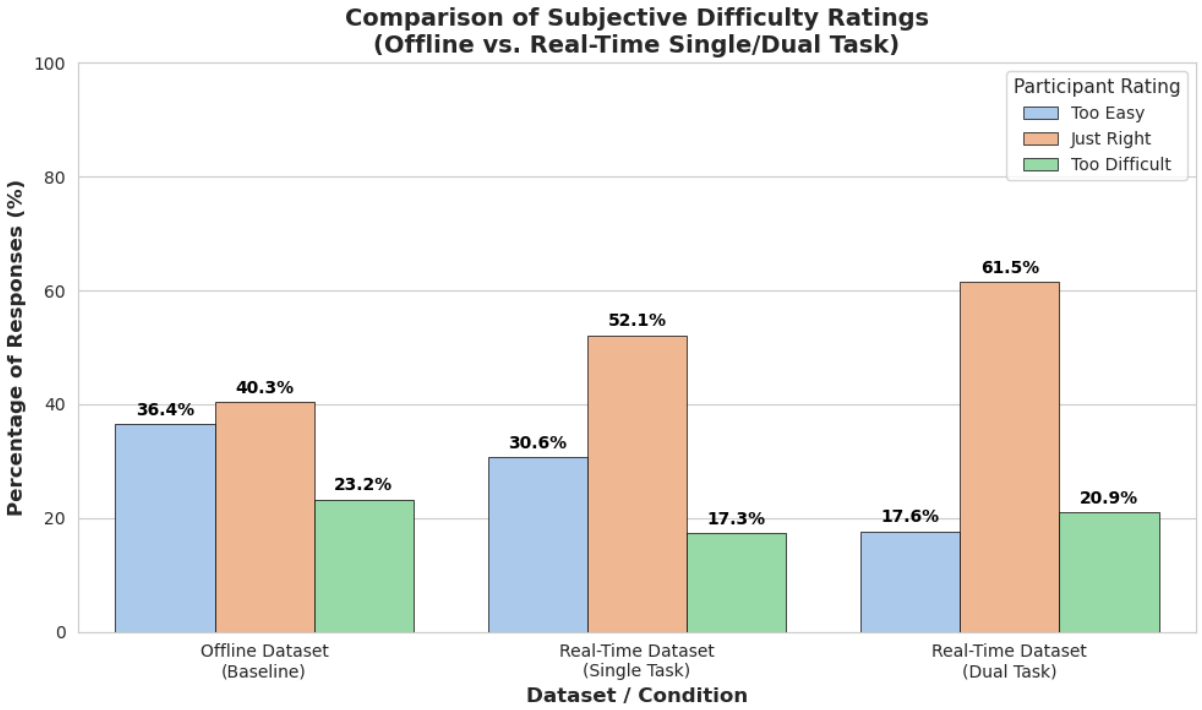


Figure 6.4: Distribution of Subjective Difficulty Ratings. The chart compares the original dataset (baseline) with the current study's Single and Dual Task conditions.

Comparison of the Real-Time and Offline Datasets

We also compared the real-time Single Task condition to the offline dataset. The offline dataset showed a more polarized distribution compared to the real-time Single Task, where responses converged more heavily toward the *Just Right* category. Specifically, the real-time Single Task condition had significantly fewer *Too Easy* ratings ($\chi^2 = 34.18, p < 0.001$) and fewer *Too Difficult* ratings ($\chi^2 = 48.29, p < 0.001$) compared to the offline baseline.

6.3.3 Game Performance Analysis: Influences of Control Mechanism and Task Complexity

To evaluate how control type (model-controlled vs. participant-controlled) and task complexity (single-task vs. dual-task) affect game performance in adaptive virtual reality environments, we conducted a 2×2 repeated measures ANOVA. In this context, game performance refers to the participant's ability to complete in-game objectives under vary-

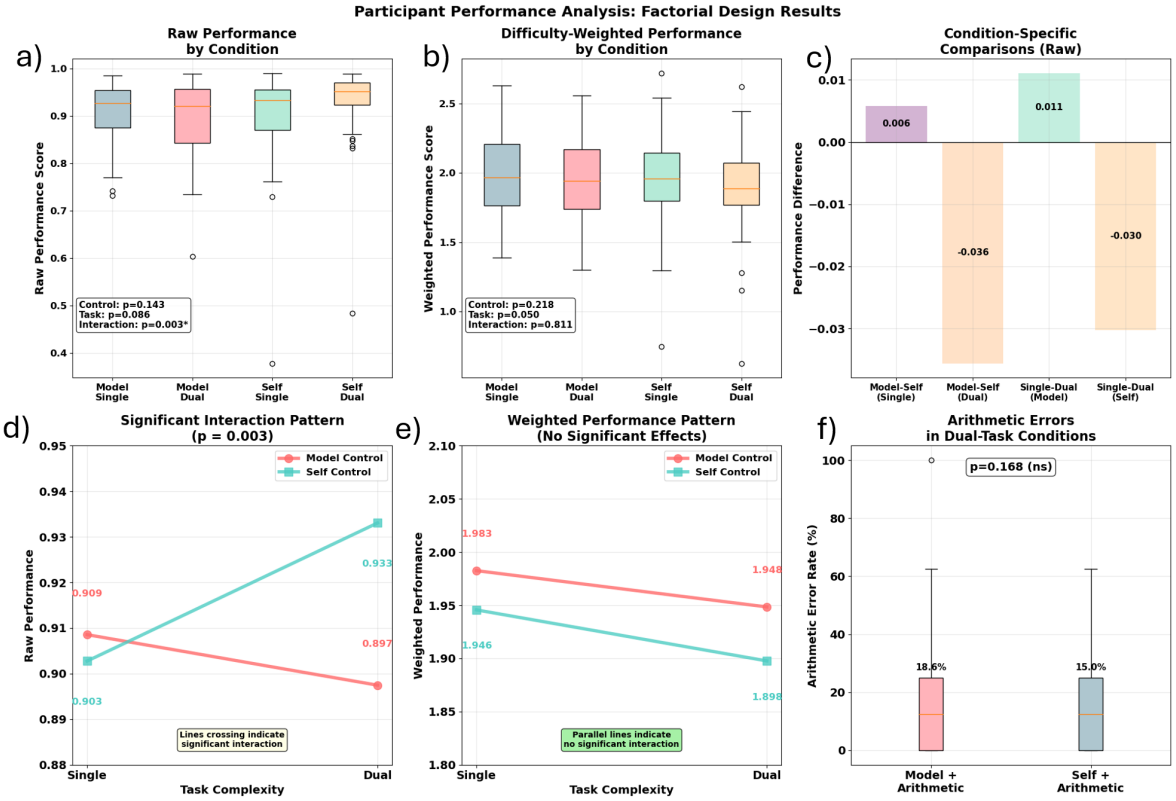


Figure 6.5: Integrated performance and error analysis across experimental conditions based on 2×2 repeated measures ANOVA. (a) Raw game performance scores by condition showing significant Control × Task Complexity interaction ($p = 0.008$) with equivalent model-self game performance in single tasks but self-control advantage in dual tasks; (b) Difficulty-weighted game performance revealing no significant effect of control type ($p = 0.219$) and borderline effects of task complexity and of the Control × Task Complexity interaction (both $p = 0.050$); (c) Condition-specific performance comparisons highlighting the interaction pattern, showing how model-self differences vary by task complexity; (d) Raw game performance interaction plot demonstrating significant crossing pattern ($p = 0.008$) where control type effectiveness depends on task complexity; (e) Weighted game performance interaction plot (Control × Task Complexity interaction borderline, $p = 0.050$); (f) Arithmetic error rates in dual-task conditions showing numerically higher errors for model-control (18.6%) versus self-control (15.0%, $p = 0.168$, ns). Statistical analyses used repeated measures ANOVA for factorial comparisons and paired t-tests for specific contrasts. Box plots show median, quartiles, and outliers; interaction plots display condition means with statistical annotations.

ing difficulty conditions. This statistical test was designed to identify: (1) whether the control mechanism influences performance outcomes, (2) whether task complexity alters performance, and (3) whether there is an interaction effect between these two factors. For example, whether the impact of control type depends on task complexity.

Measuring game performance in adaptive virtual reality tasks is a nuanced problem, particularly when task difficulty varies dynamically based on participant ability. Here we use three complementary metrics of game performance to obtain a robust view of performance:

- **Raw game performance:** proportion of cubes destroyed out of total cubes spawned (range 0-1; higher is better). This can be used to illustrate the bias of participants in judging their performance but is not an accurate representation of actual game performance as it does not take into account the difficulty at which a given score was achieved.
- **Difficulty-weighted game performance:** raw score multiplied by a difficulty modifier (range 0-4.3; higher is better). This is the most accurate representation of participant game performance as it provides a nuanced score taking into account both the difficulty and proportion of cubes destroyed. In other words, destroying a high proportion of cubes at a high difficulty level is objectively a better performance than destroying the same amount of cubes at a lower difficulty level.
- **Arithmetic error rate:** number of incorrect responses to the mental arithmetic in dual-task conditions (the less the better)

A series of 2×2 repeated measures ANOVAs was conducted with Control Type (Model vs Self) and Task Complexity (Single vs Dual) as within-subjects factors and raw game performance as well as difficulty-weighted game performance as dependent variables. A paired t-test was used to determine whether there was a difference in arithmetic errors in the dual conditions depending on Control Type. Prior to analysis, assumption checks were performed for the primary dependent variables. Outlier detection utiliz-

ing the Interquartile Range (IQR) method identified a limited number of outliers across conditions. Shapiro-Wilk tests indicated that raw game performance data violated the assumption of normality ($p < .001$) due to ceiling effect expected in levels that were too easy and did not sufficiently challenge participants. We conducted the primary analysis using standard repeated measures ANOVA and validated significant findings using a robustness check with arcsine square-root transformed data. The key results for this section are visualised in Figure 6.5. Analysis of raw game performance revealed a significant Control Type × Task Complexity interaction ($F(1,50) = 7.16, p = .008$). While performance was statistically equivalent in single-task conditions ($M_{Model} = 0.91$ vs. $M_{Self} = 0.90$), participants in the self-control condition achieved significantly higher raw accuracy in dual-task conditions ($M_{Self} = 0.93$) compared to the model-control condition ($M_{Model} = 0.90$). This interaction effect was confirmed by the robustness check on transformed data ($F(1,50) = 11.44, p = .001$), indicating that the interaction is not an artifact of the data distribution.

Difficulty-weighted game performance analysis revealed a borderline significant main effect for Task Complexity ($F(1,50) = 4.045, p = 0.050$), indicating that game performance differed between single and dual tasks when accounting for task difficulty, with no significant effect for Control Type ($F(1,50) = 1.55, p = .219$) and a borderline significant result for the Control × Task Complexity interaction ($F(1,50) = 4.04, p = .050$). This suggests that while the adaptive system maintained equivalent challenge calibration across control types, task complexity influenced game performance outcomes when weighted by difficulty.

Mental arithmetic errors showed no significant differences ($t(50) = 1.400, p = 0.168, d = 0.20$) for Control Type. The absence of significant differences in weighted game performance as well as arithmetic errors is consistent with the adaptive system maintaining appropriate challenge levels.

6.3.4 Adaptive Control Effectiveness: Difficulty Trends and Convergence

The individual predictions might not be the most reliable way of evaluating the accuracy of the model, given that they rely on the participant and model agreeing about individual 60s long level. That method of evaluation, although providing an absolute measure of accuracy, misses out on the general trend in difficulty adjustment. We argue that if the model converges to a similar difficulty level as the participant does by the end of each condition, this indicates that the overall cognitive load trend has been captured. If the model diverges from the participant-driven difficulty, this may indicate a failure of the adaptive system. It may also reflect a limitation of the participant-driven reference itself, for example if participants set difficulty conservatively, which can be examined by checking whether performance is maintained at the harder levels the model selects. Therefore, we investigated both the temporal dynamics of difficulty adjustment and the final convergence points between conditions.

To capture the continuous nature of this adaptation process, we first analyzed the slope of difficulty changes over time using a Linear Mixed-Effects Model (LMM). The model included fixed effects for Time (Level Sequence), Control Type (Model vs Self), and Task Complexity (Single vs Dual), along with random intercepts and slopes for Time grouped by Participant. Residual diagnostics indicated a minor deviation from normality ($W = 0.953$). This deviation was attributed to the discrete, step-wise nature of the difficulty levels. Given the large number of observations ($N = 1,632$) and the balanced study design, the model was deemed robust to these minor violations without the need for transformation. We followed that by a 2×2 repeated measures ANOVA was conducted on the final level played, with Control Type and Task Complexity as factors. Prior to analysis, the distribution of final difficulty levels was inspected. Shapiro-Wilk tests indicated deviations from normality in all conditions (W range: $0.83 - 0.95$, $p < .05$).

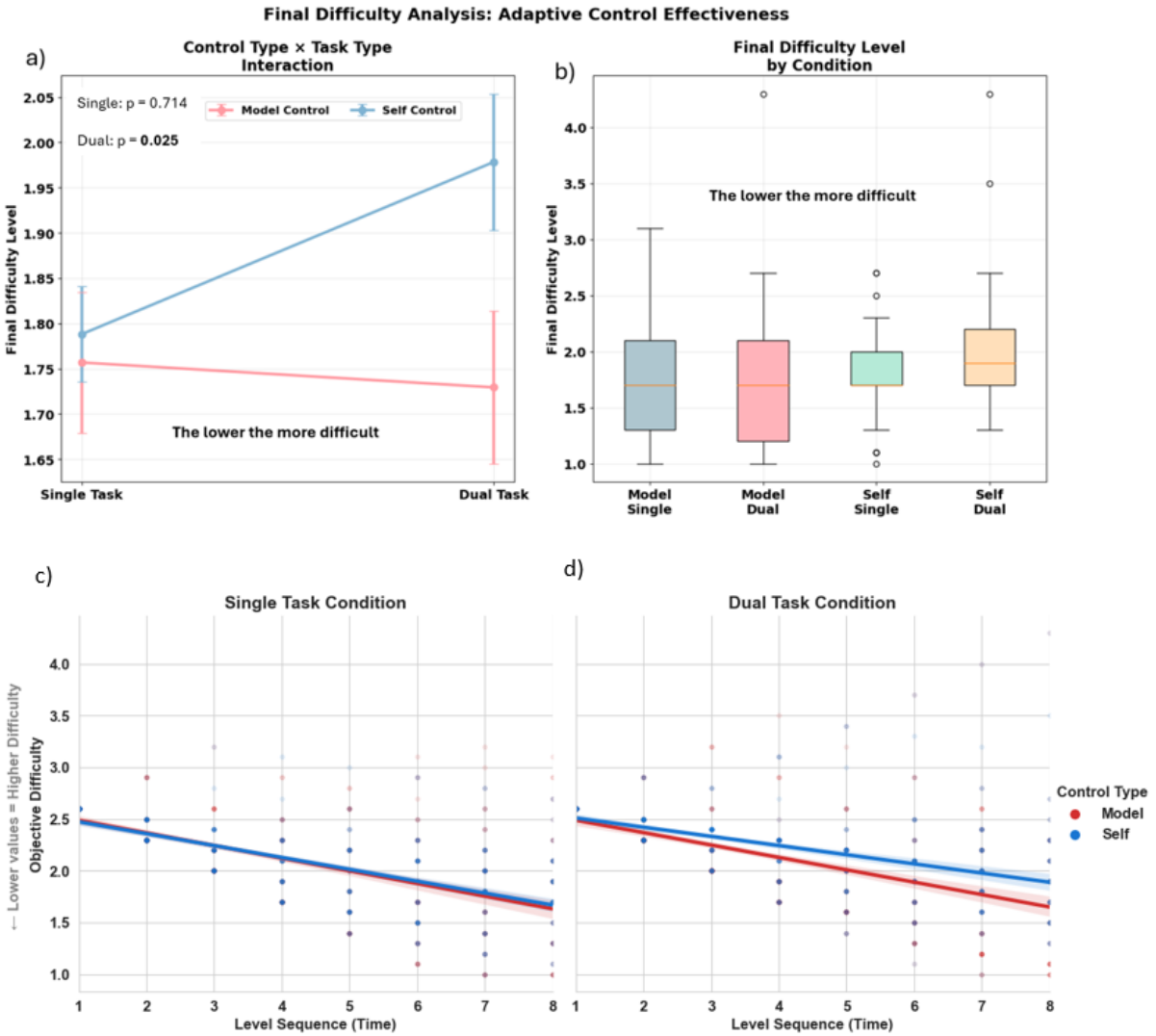


Figure 6.6: Final difficulty analysis revealing control type effectiveness and task complexity interactions. (a) Control type × task type interaction plot demonstrating equivalent performance in single-task conditions but divergence in dual-task conditions, with model control achieving harder final difficulties. Statistical comparisons show non-significant differences in single-task conditions (p=0.714) but marginally significant differences in dual-task conditions (p=0.025, corrected p=0.05), with model-based control reaching harder final difficulties than self-assessment control when cognitive resources were divided between VR task performance and arithmetic processing; (b) Final difficulty distributions across all four experimental conditions, showing the magnitude of differences between control types within each task complexity level; (c) Difficulty trajectories over level sequence in single-task conditions and (d) in dual-task conditions, with linear fits for model-driven (red) and participant-driven (blue) control. Lower difficulty values indicate harder task settings. Box plots show median, quartiles, and outliers; error bars in the interaction plot represent standard error of the mean.

Visual inspection of Q-Q plots revealed that these deviations were driven by the discrete nature of the difficulty levels (step-wise increments) rather than extreme outliers. Given the balanced design and sample size ($N = 51$), the repeated measures ANOVA was deemed robust to these violations.

The analysis revealed a significant main effect of Time ($\beta = -0.122, SE = 0.009, p < .001$), confirming that objective difficulty generally increased (numerical value decreased) as the session progressed across all conditions. A marginally significant three-way interaction was observed between Time, Control Type, and Task Complexity ($\beta = 0.024, SE = 0.013, p = .073$). As visualized in Figure 6.6c and 6.6d, the difficulty trajectories diverge under load. While the participants' adaptation slope was aligned with the model's in the single task condition (Figure 6.6c), it diverged in the dual-task condition (Figure 6.6d, red line), increasing difficulty at a slower rate than in the model-controlled condition.

This divergence in adaptation rates accumulated over the session, culminating in a difference at the final difficulty level in dual-task conditions. The results of the 2×2 repeated measures ANOVA revealed no significant main effect of task complexity ($t(50) = 1.62$, $p = 0.112$), and the main effect of control type approached but did not reach significance ($t(50) = -1.85$, $p = 0.070$). Although the omnibus interaction was not statistically significant ($t(50) = -1.79$, $p = 0.080$), planned simple main effects analysis was in line with the trend observed in the slope analysis.

In single-task conditions, model-based and self-assessment control achieved virtually equivalent final difficulty levels (Model: 1.76 ± 0.55, Self: 1.79 ± 0.38, $t = -0.37$, $p = 0.714$), indicating comparable performance when cognitive load was low (see Figure 6.6a). However, in dual-task conditions, model-based control assigned harder final difficulties than participants selected for themselves (Model: 1.73 ± 0.60, Self: 1.98 ± 0.53, $t = -2.31$, $p = 0.025$, $d = 0.33$). After applying a Bonferroni correction for two comparisons, this difference was borderline (corrected $p = 0.05$).

Table 6.2: Cognitive Load Judgment Agreement Analysis for Raw Game Performance and Difficulty-Weighted Game Performance by Task Type and Perceived Difficulty Category - . Bold values indicate statistical significance. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Positive t-statistics indicate higher self-control game performance relative to model-control. Cohen's d effect sizes: small (0.2), medium (0.5), large (0.8).

Task Type	**Category**	**Raw Game Performance**			**Weighted Game Performance**		
		t-stat	*p-value*	*Cohen's d*	*t-stat*	*p-value*	*Cohen's d*
Single	Too Easy	5.189	**0.000*** **	0.749	-1.551	0.128	-0.224
	Just Right	0.491	0.626	0.070	-1.054	0.297	-0.151
	Too Difficult	-5.906	**0.000*** **	-0.958	-0.185	0.854	-0.030
Dual	Too Easy	4.611	**0.000*** **	0.758	-1.128	0.267	-0.185
	Just Right	3.241	**0.002** **	0.463	-3.322	**0.002** **	-0.475
	Too Difficult	-2.437	**0.019***	-0.376	0.347	0.731	0.054
Pooled	Too Easy	5.146	**0.000*** **	0.743	-0.998	0.324	-0.144
	Just Right	2.136	**0.038***	0.299	-3.972	**0.000*** **	-0.556
	Too Difficult	-5.080	**0.000*** **	-0.749	1.185	0.242	0.175

The divergence in final difficulty played, along with no difference in weighted game performance across dual tasks, may indicate that participants are more conservative with their assessments. They adopted a flatter adaptation slope (Figure 6.6d) and ended at easier levels than the model determined they were capable of handling but ended up achieving similar weighted game performance and equivalent arithmetic error rates. This suggests that despite the model pushing the participants to play at harder levels in dual task designs, it does not come at a cost in difficulty-weighted performance.

6.3.5 Divergence in Cognitive Load Judgments: Human vs. Model Thresholds

Analysis of the cognitive load judgment thresholds revealed systematic differences in how the computational model and human participants categorised task difficulty, providing insight into the reasons behind prediction discrepancies observed throughout the experiment (Figure 6.7). Specifically, we examined the mean game performance levels associated with each difficulty label (*Too Easy, Just Right and Too Difficult*) to compare how model predictions relate to participant perceptions of optimum difficult levels, see Table 6.2.

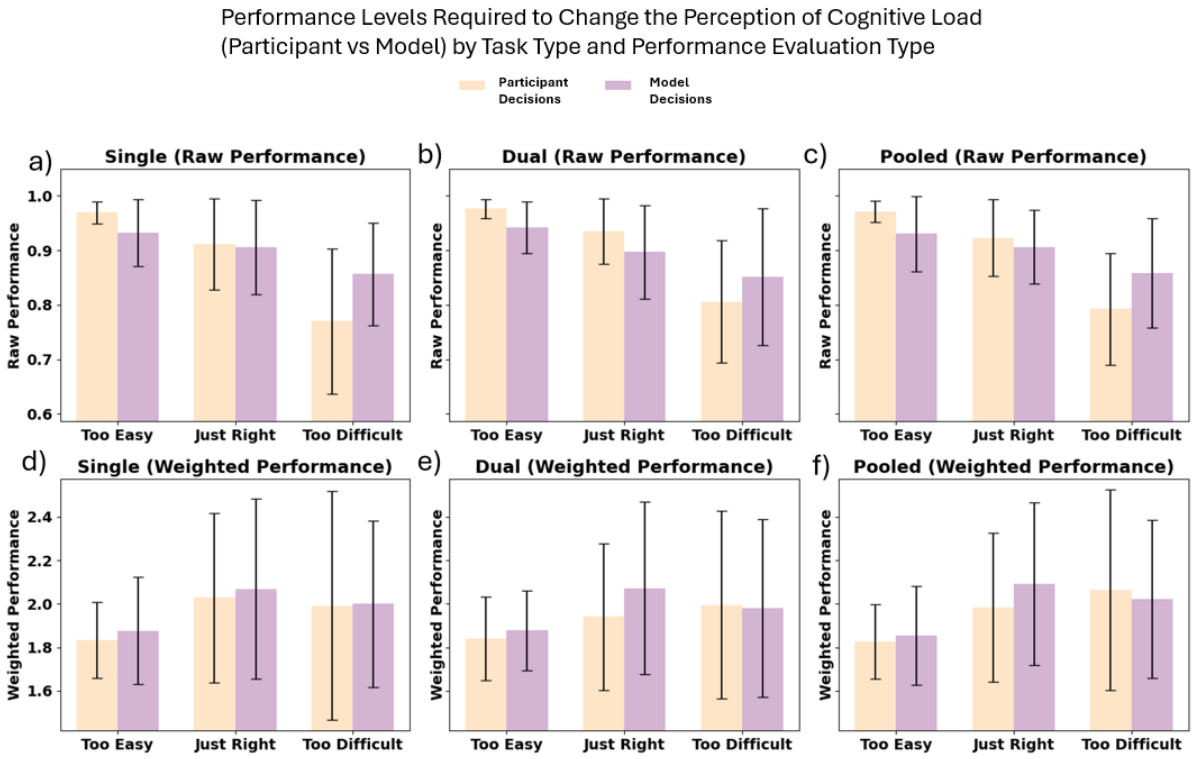


Figure 6.7: Game performance analysis showing the relationship between difficulty judgments/predictions and actual game performance. The 2×3 grid compares raw performance scores (top row) and weighted game performance scores (bottom row) for both model and self-driven conditions, including an additional breakdown for single (left column) and dual tasks (middle column) as well as both conditions pooled (right column). Boxplots display the distribution of mean performance scores aggregated by participant and difficulty (Too Easy, Just Right and Too Difficult). This visualization presents the mapping of raw and weighted game performance across difficulty estimations.

In the raw game performance analysis, participants tended to overestimate their ability when they perceived a task as *Too Easy* and underestimate it when they judged it as *Too Difficult*. These systematic biases disappeared in the weighted game performance analysis, where game performance was adjusted for objective task difficulty, suggesting that participants did not fully factor in difficulty when forming their judgments. By contrast, the computational model maintained consistent judgment thresholds across conditions, with differences persisting only in the *Just Right* category, which is consistent with the training strategy prioritising *Too Easy* and *Too Difficult* classification over *Just Right*. This suggests that the model accounted for task demands more consistently, while participant judgments appeared to be more influenced by a simpler sense of success based on the number of cubes left at the end of the level, regardless of how many of them appeared.

6.3.6 Subjective Workload Assessment with NASA TLX

While real-time cognitive workload evaluation based on categorical difficulty labels (*Too Easy, Just Right, Too Difficult*) provides insight into moment-to-moment task adaptation, it may not fully capture the broader experiential and affective dimensions of user engagement. To complement this model-driven assessment, we employed the NASA Task Load Index (TLX) at the end of each experimental condition, as a measure for comparing conditions rather than for tracking participant state during play. Full NASA-TLX descriptive statistics, overall and broken down by condition, are reported in Appendix 10, Section 10.1.

To examine the effects of the adaptive system and the secondary task on subjective experience, a 2×2 repeated measures ANOVA was conducted on the weighted TLX scores, with Control Type (Model vs. Self) and Task Complexity (Single vs. Dual) as within-subject factors. Shapiro-Wilk tests confirmed that NASA-TLX workload scores satisfied the assumption of normality across all four experimental conditions ($W > 0.96$, $p > .19$), validating the use of parametric analysis without correction.

Task Complexity

The analysis revealed a robust main effect of Task Complexity across nearly all workload dimensions, confirming the efficacy of the arithmetic task in inducing cognitive load. For the overall Total Workload score, participants reported significantly higher burden in the dual-task condition compared to the single-task condition ($F(1,50) = 38.48, p < .001$).

Breakdown by subscale indicated that this increase was driven primarily by **Mental Demand** ($F(1,50) = 101.03, p < .001$) and **Effort** ($F(1,50) = 31.30, p < .001$), reflecting the high cognitive cost of concurrent arithmetic processing. **Frustration** levels were also significantly higher in the dual-task condition ($F(1,50) = 24.28, p < .001$), suggesting that the divided attention requirement negatively impacted affective state. **Physical Demand** ($F(1,50) = 10.87, p = .002$) and **Temporal Demand** ($F(1,50) = 6.45, p = .014$) showed smaller but statistically significant increases. Notably, the **Performance** subscale did not differ significantly between task conditions ($F(1,50) = 2.40, p = .127$), indicating that participants felt their level of success was comparable regardless of the added complexity.

Control Type and Interaction Effects

In contrast to the strong effects of task complexity, the main effect of Control Type was not statistically significant for Total Workload ($F(1,50) = 1.87, p = .177$), nor for any individual subscale (all $p > .08$). This indicates that participants did not perceive the Model-controlled adaptive difficulty as significantly more demanding or intrusive than their own self-regulated adjustments.

The interaction between Control Type and Task Complexity was not significant for Total Workload ($F(1,50) = 0.27, p = .608$) or any subscale. This suggests that the increase in workload observed during the dual-task condition was consistent across both control methods. The adaptive model did not exacerbate the workload of the secondary task, nor did it mitigate it significantly more than users could achieve on their own.

Game Performance Relationships with NASA TLX-based Workload Ratings

Significant negative correlations emerged between TLX ratings and task performance, particularly for both raw and weighted game performance measures. For raw game performance, moderate negative correlations were observed with performance perception ($r = -0.442,\ p < 0.001$), frustration ($r = -0.399,\ p < 0.001$), and total workload ($r = -0.323,\ p < 0.001$). For weighted game performance, which accounts for task difficulty, the correlations were weaker but still significant for game performance perception ($r = -0.287,\ p = 0.002$), frustration ($r = -0.245,\ p = 0.008$), and total workload ($r = -0.206,\ p = 0.023$). Other TLX subscales showed non-significant correlations with weighted game performance, including mental demand ($r = -0.152,\ p = 0.103$), physical demand ($r = -0.087,\ p = 0.346$), and temporal demand ($r = -0.114,\ p = 0.208$). These relationships suggest that higher subjective workload was associated with poorer game performance, with raw game performance showing stronger associations than weighted game performance. Additionally, mental demand showed a notable positive correlation with arithmetic error rates ($r = 0.353,\ p < 0.001$), consistent with dual-task interference being reflected in both objective game performance decrements and subjective cognitive load ratings.

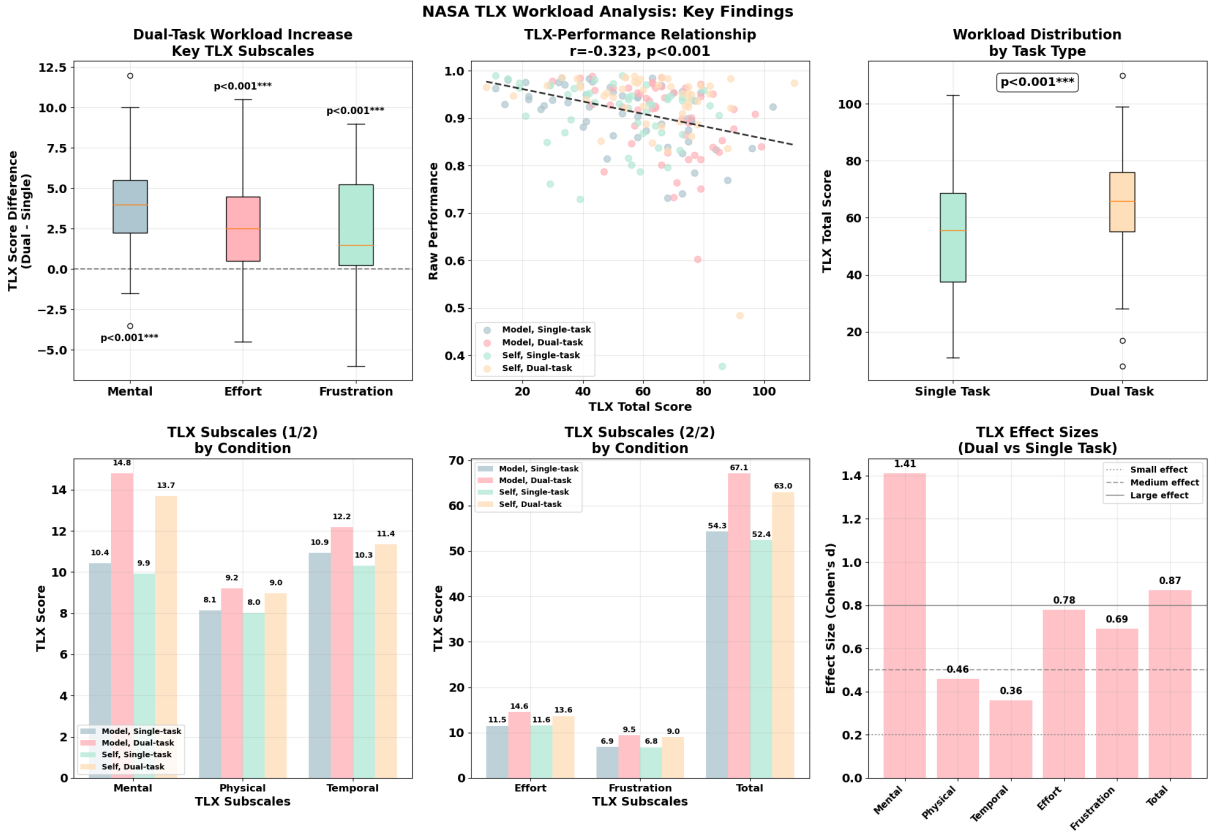


Figure 6.8: NASA TLX workload analysis revealing dual-task cognitive burden and performance relationships. (a) Dual-task workload increases showing significant elevations in mental demand ($p < 0.001$, $d = 1.41$), effort ($p < 0.001$, $d = 0.78$), and frustration ($p < 0.001$, $d = 0.69$) when comparing dual-task to single-task conditions within participants; (b) TLX-performance correlation demonstrating a significant negative relationship ($r = -0.323$, $p < 0.001$) between subjective workload and spawn-destroy task performance across all conditions; (c) Total workload distribution confirming significantly higher TLX scores in dual-task compared to single-task conditions ($p < 0.001$, $d = 0.87$); (d) TLX subscale profiles for mental demand, physical demand, and temporal demand across the four experimental conditions, with Model Dual-task showing the highest mental demand ratings; (e) TLX subscale profiles for effort, frustration, and total workload, demonstrating consistent patterns where dual-task conditions (Model Dual-task and Self Dual-task) exceed single-task conditions across all measures; (f) Effect size analysis quantifying the magnitude of dual-task workload increases, with mental demand showing the largest effect ($d = 1.41$), followed by total workload ($d = 0.87$), effort ($d = 0.78$), and frustration ($d = 0.69$), while physical and temporal demands showed smaller but significant effects ($d = 0.46$ and $d = 0.36$, respectively). Box plots display median, quartiles, and outliers; bar charts show condition means with individual values labeled; effect size reference lines indicate Cohen's conventions for small (0.2), medium (0.5), and large (0.8) effects.

Table 6.3: Repeated-Measures ANOVA Results for Single-Task Conditions

Metric	F Value	Num DF	Den DF	p-value
Raw Game Performance	3.3118	1	50	0.0748
Weighted Game Performance	38.2463	1	50	<0.0001
Final Difficulty	14.3117	1	50	0.0004
TLX Total	3.8970	1	50	0.0539
TLX Mental	5.3976	1	50	0.0243
TLX Physical	0.1066	1	50	0.7454
TLX Temporal	3.8713	1	50	0.0547
TLX Perceived Performance	4.0689	1	50	0.0491
TLX Effort	8.0692	1	50	0.0065
TLX Frustration	0.3982	1	50	0.5309

Table 6.4: Repeated-Measures ANOVA Results for Dual-Task Conditions

Metric	F Value	Num DF	Den DF	p-value
Raw Game Performance	6.1670	1	50	0.0164
Weighted Game Performance	14.4091	1	50	0.0004
Final Difficulty	2.6869	1	50	0.1075
TLX Total	12.7224	1	50	0.0008
TLX Mental	20.0635	1	50	<0.0001
TLX Physical	5.3026	1	50	0.0255
TLX Temporal	3.6519	1	50	0.0617
TLX Perceived Performance	7.9402	1	50	0.0069
TLX Effort	9.5341	1	50	0.0033
TLX Frustration	3.2905	1	50	0.0757

6.3.7 Learning and Fatigue Effects

To assess participant adjustment to the task over time, we analysed learning and fatigue patterns. We compared participants' game performance and NASA-TLX results between their first and second exposure to each task type (single-task vs dual-task conditions). The experiment design ensured consistent sequencing across participants. Single task blocks were always in positions 1 and 3, while dual task blocks appeared in positions 2 and 4. This fixed ordering allowed for direct within-subject comparisons, controlling for time-on-task and interval effects across all participants. Repeated-measures ANOVA was conducted separately for each task type with single-task results presented in Table 6.3 and dual-task results presented in Table 6.4.

Performance Dynamics and Skill Acquisition

To assess the sustainability of performance and the presence of learning effects over the course of the session, we analysed the trajectory of difficulty-weighted performance. A Linear Mixed-Effects Model (LMM) was fitted with **Level Sequence** (Time), **Control Type**, and **Task Complexity** as fixed effects, including random intercepts and slopes for participants. Prior to analysis, residual diagnostics were performed to validate the model. The Shapiro-Wilk test on model residuals indicated an excellent fit to the normal distribution ($W = 0.997$), and visual inspection of Q-Q plots confirmed that residuals closely followed the theoretical diagonal. A scatterplot of residuals versus fitted values displayed a random dispersion of points, satisfying the homoscedasticity assumption required for linear mixed-effects modeling.

The analysis revealed a robust and statistically significant main effect of **Time** (Level Sequence) on performance ($\beta = 0.090, SE = 0.009, p < .001$). As illustrated in Figure 6.9, participants in all conditions demonstrated a steady improvement in weighted performance as the session progressed, indicating a continuous learning effect rather than fatigue. The positive slope suggests that users were able to handle increasingly difficult challenges (or maintain accuracy at higher difficulty levels) without a degradation in capacity.

This learning rate was consistent across all experimental conditions. The interaction between Time and Control Type was not significant ($p = .942$), nor was the three-way interaction between Time, Control Type, and Task Complexity ($p = .722$). This indicates that the rate of improvement was identical regardless of whether the difficulty was controlled by the model or the user, and regardless of the secondary task load.

While the model-controlled condition did not induce a *steeper* learning curve, it successfully supported the same rate of skill acquisition as in the self-controlled conditions. The absence of a negative interaction effect in the dual-task condition further confirms that the additional cognitive load did not inhibit the users' ability to learn and improve at the primary task over time.

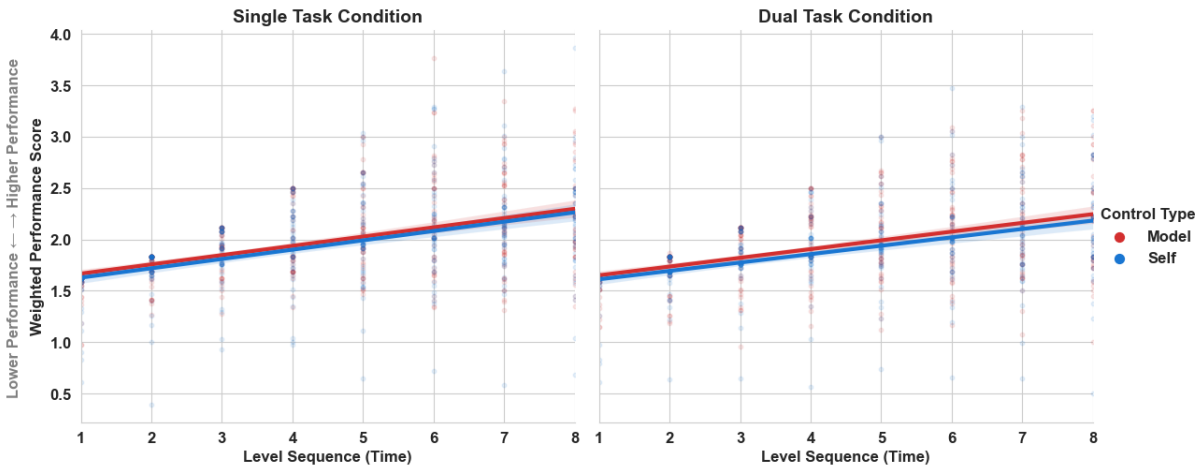


Figure 6.9: Linear regression of Weighted Performance trajectories over time. (a) Single Task condition (left) and (b) Dual Task condition (right). Red lines represent Model Control; Blue lines represent Self Control. The significant positive slope across all conditions ($p < .001$) indicates a consistent learning effect. Shaded regions represent 95% confidence intervals.

This continuous learning trajectory culminated in significant performance retention between the first and second exposures of each task. Repeated-measures ANOVA confirmed that Weighted Game Performance improved significantly in the second exposure for both Single-Task ($\Delta = +0.224, p < .001, d = 0.87$) and Dual-Task ($\Delta = +0.146, p < .001, d = 0.53$) conditions.

Raw Game Performance (unweighted) showed a more modest improvement, reaching significance only in the Dual-Task condition ($p = .016$). This discrepancy between weighted and raw metrics highlights the function of the adaptive system: as users learned (positive slope), the system increased the difficulty to match their new capacity. Consequently, while their raw score remained relatively stable, their difficulty-weighted performance increased because they were operating at a significantly higher difficulty level. This is confirmed by the analysis of Final Difficulty, which significantly increased (numerical value decreased) in the second Single-Task exposure ($p < .001, d = 0.53$), indicating that the learning observed in the LMM was successfully converted into higher challenge acceptance.

Workload Reduction Over Time

A pronounced learning effect was also observed in subjective workload measures. Total TLX scores decreased significantly for dual-task conditions by 8.5 points ($p < 0.001, d = 0.499$), as illustrated in the individual trajectory plot (Figure 6.10d). This workload reduction was driven by improvements across multiple TLX subscales (Figure 6.10e), with mental demand showing the largest decrease (-2.4 points, $p < 0.001$), followed by effort (-1.9 points, $p = 0.003$), perceived performance (-1.2 points, $p = 0.007$), and physical demand (-1.0 points, $p = 0.026$). Single-task conditions showed more modest workload reductions, with significant decreases in mental demand (-1.5 points, $p = 0.024$), effort (-2.0 points, $p = 0.007$), and perceived performance (-1.2 points, $p = 0.049$), while total TLX scores approached significance ($p = 0.054$).

6.3.8 Individual Differences in Real-time Model Performance

Analysis of per-participant model accuracy revealed substantial individual variability in prediction success, ranging from 6.2% to 59.4% accuracy across the 51 participants (M = 36.7%, SD = 11.9%). This wide distribution suggests that accuracy was considerably higher for some participants than for others.

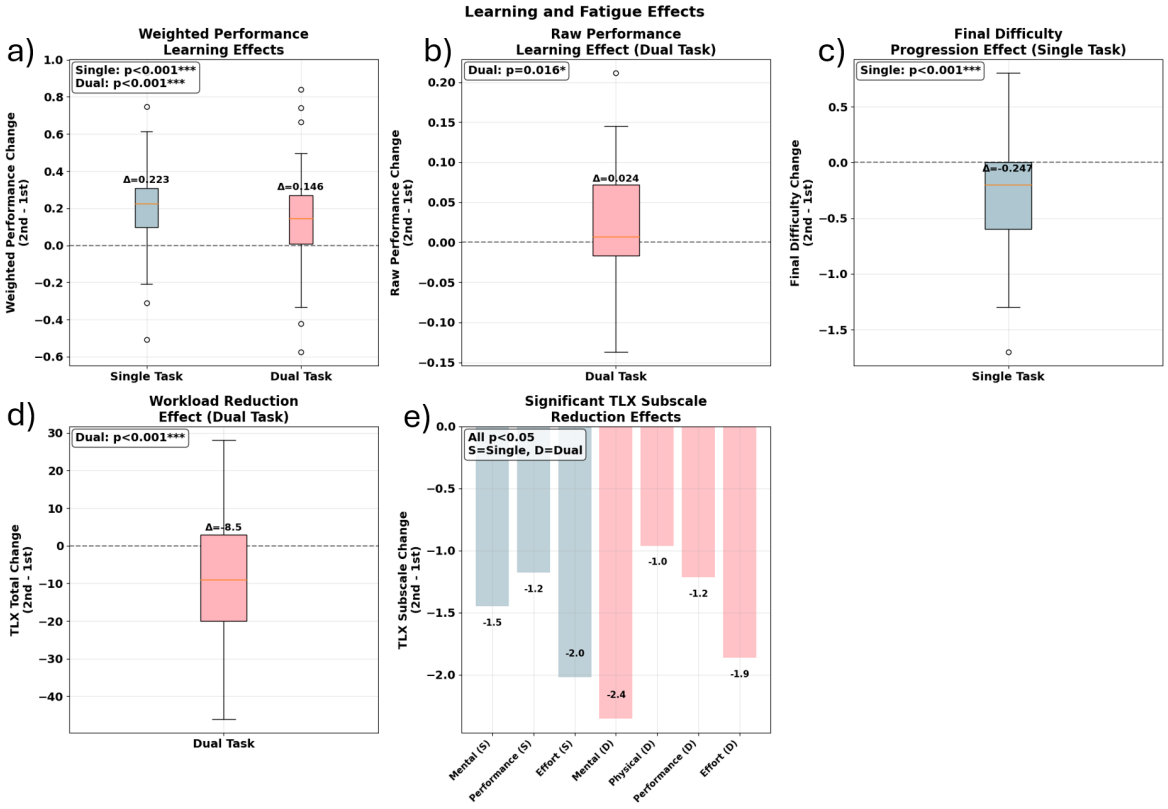


Figure 6.10: Significant learning and fatigue effects across experimental conditions. The figure displays five key findings: (a) Weighted game performance improvements in both single-task (Δ=0.224, p<0.001) and dual-task conditions (Δ=0.146, p<0.001); (b) Raw game performance improvement in dual-task conditions only (Δ=0.024, p=0.016); (c) Final difficulty level reduction in single-task conditions only (Δ=-0.162, p<0.001), indicating progression to harder tasks; (d) Overall workload reduction in dual-task conditions (Δ=-8.5, p<0.001); and (e) Specific TLX subscale reductions, with single-task conditions showing improvements in mental demand (Δ=-1.5, p=0.024), performance perception (Δ=-1.2, p=0.049), and effort (Δ=-2.0, p=0.007), while dual-task conditions demonstrated broader improvements across mental demand (Δ=-2.4, p<0.001), physical demand (Δ=-1.0, p=0.026), performance perception (Δ=-1.2, p=0.007), and effort (Δ=-1.9, p=0.003) subscales (all p<0.05). Box plots show median, quartiles, and outliers; bar charts display mean differences with statistical significance levels indicated.

Fairness Analysis of Demographic Predictors in Model Performance

Male participants had numerically higher model accuracy than female participants (40.3% vs 34.8%), although this difference did not reach statistical significance ($t = 1.565$, $p = 0.124$). Age showed no relationship with model accuracy ($r = 0.022$, $p = 0.877$), with equivalent performance across age groups (18-25: 35.5%, 26-35: 37.5%, 36+: 37.8%, $F = 0.188$, $p = 0.829$). Handedness effects could not be reliably assessed due to the small number of left-handed participants ($n = 2$).

Predictive Factors for Model Success

Correlation analysis identified probability-accuracy calibration as the strongest individual predictor of overall model performance ($r=0.385$,$p=0.005$). This metric is computed from the same predictions as accuracy, so it partly reflects accuracy itself. In the context of the zero-calibration deployment strategy, it may also represent the distributional alignment between an individual participant and the global training norms. Participants for whom the model performed best showed positive correlations between model confidence and accuracy, which may indicate that their standardized physiological features aligned well with the population-level distributions used for scaling. Low-performing participants showed weak or negative correlations, which may reflect a distributional shift in which the global scaler did not normalize their physiological baselines effectively.

In contrast to this physiological alignment metric, demographic variables (age, gender), baseline VR performance, and subjective workload measures showed only weak non-significant associations with model accuracy. A predictive model incorporating these readily observable characteristics explained only 13.7% of the variance in individual model

accuracy. This indicates that traditional screening criteria provide limited predictive power for identifying suitable candidates for model-based adaptive control, which is consistent with performance depending on the compatibility between the user's physiology and the global normalization model.

6.4 Discussion

6.4.1 Methodological Contributions

A prevalent methodological shortcut in the field of cognitive load estimation is the use of 'task stacking' (adding discrete secondary tasks) as a proxy for cognitive load manipulation [25, 178]. While this creates distinct physiological classes, it prevents the granular, incremental adjustments required for effective training within a single task. Moreover, in particular when using eye-tracking as means of detecting cognitive load shifts, task stacking can result in the model learning to distinguish the eye-tracking patterns required to perform the additional tasks, rather than the actual cognitive load increase. In the domain of adaptive design, existing systems often rely on simple game-based heuristics [9, 48], modulate visual complexity rather than intrinsic load [101, 36], or focus on affective states rather than cognitive capacity [129, 132].

Furthermore, real-time deployment of cognitive load models requires more than high offline accuracy. Some research prioritises methodological choices that maximize offline performance without considering real-world applicability. For instance, the majority of studies estimating cognitive load from eye-tracking and physiological data rely on within-participant cross-validation, which inflates accuracy by training and testing on the same individuals [149, 151, 68, 178]. When studies adopt leave-one-participant-out validation, which is a more realistic test of generalization, performance typically drops [15, 25].

This study addresses these gaps methodologically rather than claiming to have solved them: it uses a leave-one-participant-out protocol for the deployment model, manipulates difficulty intrinsically within the task, and deploys the system on consumer-grade hardware suitable for at-home use. As noted in a recent systematic review by Mortazavi [115], the field is shifting from rule-based heuristics to data-driven methods capable of capturing internal states. This study is an attempt at that shift, moving from offline analysis to a live, closed-loop system. As shown in the results, per-trial accuracy against participants' own ratings was lower in real time deployment than could have been expected based on the training benchmarks from 5, yet outcomes under model-driven control were comparable to those under participant-driven control.

6.4.2 Key Findings

Conservatism Bias vs. Overconfidence: The Divergence of Objective and Subjective Difficulty

One notable finding of this study is a divergence in difficulty regulation strategies under high cognitive load. In dual-task conditions, the model drove participants to modestly harder final difficulty levels than they selected for themselves ($d = 0.33$; borderline after Bonferroni correction, $p = 0.05$, so the effect should be treated as small and tentative). This increase in difficulty did not come with a measurable performance penalty on difficulty-weighted game performance or arithmetic error rate. One interpretation consistent with this pattern is a conservatism bias. Participants in the dual-task condition may underestimate their remaining capacity and self-regulate toward easier, safer difficulty levels. The model, working from physiological signal rather than self-report, was not subject to the same bias and pushed difficulty to a level participants could in fact sustain without cost. This is consistent with the model's real-time predictions being a noisier match to subjective labels than to what participants could actually handle, a distinction Chapter 7 examines directly.

This finding contrasts with Constant et al. [39], who observed that dynamic-difficulty adjustment (DDA) systems typically result in player *overconfidence*. However, their study employed a betting mechanism that explicitly incentivised risk assessment in a single-task context. The players needed to adjust the bet amount prior to each round, knowing that if they win, they will get the bet back along with a reward multiplier and if they lose, the money they bet will be taken away. As noted by Guo and Ren [60], such betting designs rely on *prediction* (future expectation), which typically triggers optimism biases, whereas post-hoc ratings or physiological monitoring rely on *reflection* (past evaluation), which often yields more rational or conservative estimates. In our study, the hidden nature of the control loop removed this gamification incentive as players were unaware that their difficulty judgments would influence the task difficulty. Instead, the observed conservatism may be better explained by dissociation theory [186]. Yeh and Wickens posit that subjective ratings reflect the cognitive resources invested to maintain performance, rather than the performance outcome itself. Consequently, participants may have felt the increased effort of the dual-task and rated it as *Too Difficult*, even though their objective performance output remained stable. Furthermore, this judgment may have been reinforced by a rule of thumb of the kind discussed below, where participants assumed that the structural presence of a secondary task necessitated higher difficulty ratings regardless of their actual remaining capacity.

This heuristic reliance is consistent with the disconnect between Objective Game Difficulty and Subjective Game Difficulty described in the literature [60, 40]. Our data suggests that participants anchored their judgments on raw performance (absolute number of cubes they missed despite difficulty fluctuations), whereas the model's output aligned more closely with difficulty-weighted performance. This is consistent with the metacognitive rule proposed by Hendy et al. [69], where participants rely on observable cues, in this case raw performance, as a mental shortcut for judging task difficulty. By setting difficulty from its own output rather than the participant's stated preference ('two tasks

must be harder than one'), the model-driven condition in dual-task settings pushed difficulty upward without an accompanying performance cost. This is a point worth weighing against task stacking as a proxy for cognitive load, since task stacking introduces exactly this kind of structural bias into participants' own judgments.

Adaptive Smoothing: The Shift to *Just Right*

We observed a significant distributional shift in difficulty ratings between the offline data collection (Experiment 1) and the real-time deployment (Experiment 2, a different cohort of participants). In the real-time study, participant ratings converged heavily on *Just Right*, with a significant reduction in *Too Easy* and *Too Difficult* edge cases. The adaptive system stabilises the difficulty through informed and incremental difficulty adjustments as opposed to random and large amplitudes observed in fixed difficulty systems [115]. Even though the classification model itself was pre-trained on a fixed difficulty task, the closed-loop nature of the system naturally minimised the occurrence of non-optimal states.

This is broadly consistent with the goal-based DDA framework proposed by Guo et al. [61]. That framework suggests that the primary value of DDA is its ability to control the difficulty progression toward a 'design curve'. In our case, the design curve aims to align the *Just Right* states with peak difficulty-weighted performance. By continuously correcting deviations, the closed-loop system kept most participants closer to the *Just Right* band than the random or fixed-step progressions used in the offline experiment. Some of this stabilisation is attributable to the closed-loop structure itself, since an incremental, small-step control law will tend to keep any session from drifting far from the target band, independent of the underlying classifier's precision on any single trial. This does not diminish the result so much as explain part of the mechanism behind it: the combination of a directionally-informative model and a responsive control loop was sufficient to produce a more stable experience than the fixed-difficulty design, even without high per-trial classification accuracy.

Model Utility: Automating the Training Loop

The clearest positive result of this study is that physiological, model-driven adaptation produced outcomes equivalent to participant-driven adaptation on every key metric tested: weighted game performance, subjective workload (NASA-TLX), and learning rate all showed no significant difference by control type. This is a meaningful result in its own right, independent of the model's raw per-trial classification accuracy against subjective labels. It is measured against the outcomes that actually matter for a training system, rather than against agreement with a self-report label that Chapter 7 suggests becomes a less reliable target under the low-contrast conditions used here. On this evidence, the explicit step of asking users for difficulty assessments, which breaks immersion and is itself subject to the conservatism bias discussed above, can be replaced by model-driven monitoring of physiological state without a measurable cost to the user's training outcomes.

This is relevant to the broader case for serious games and cognitive training, where, as Guo et al. [61] argue, the goal of maximising training capacity can differ from pure entertainment goals. In the dual-task condition specifically, the model-driven system pushed users to harder difficulty levels than they chose for themselves, without the frustration or disengagement that Mortazavi et al. [115] warn against, and without a measurable performance cost. Taken together with the equivalence result above, this is consistent with a physiologically-driven adaptive system being able to match what participants would choose for themselves. In the specific case of dual-task difficulty progression, it may modestly exceed it, at least over the single-session lengths tested here. Extending this to longer, multi-session training programmes remains for future work.

6.4.3 Limitations and Future Directions

The strongest evidence in this chapter concerns outcomes: model-driven and participant-driven control produced equivalent performance, workload, and learning trajectories, and the model pushed dual-task difficulty modestly higher without a performance cost. This is weaker evidence about the classifier's moment-to-moment accuracy: real-time accuracy against participants' own difficulty ratings was 36.7%, against a chance baseline of 33.3% for three classes, and below the 52.3% pre-deployment figure from Chapter 5. Most errors were between *Just Right* and an edge case, which the small-step adaptation rule tolerates. As argued above, a substantial part of this decline is plausibly attributable to the instability of the subjective label itself under low-contrast, real-time conditions rather than a proportional decline in the model's underlying signal, and Chapter 7 tests this directly by comparing model alignment with subjective versus objective targets. Even so, this chapter cannot fully separate how much of the accuracy drop is label instability versus genuine degradation in the model's real-time signal, and a like-for-like comparison against an objective real-time target would strengthen this claim considerably.

We also observed substantial individual variability in model performance, consistent with the individual differences often cited in physiological computing: per-participant accuracy ranged from 6.2% to 59.4%. This suggests that the disconnect between physiology and self-report is not uniform across participants. Future iterations of this system would benefit from a calibration phase to align the model's decision boundaries with the individual's baseline physiology, rather than relying on global normalization. Alternatively, a larger and more diverse offline dataset could result in a more general model, though Chapter 5's raw-signal baseline result suggests that more data alone would not substitute for the feature engineering pipeline this model already depends on.

6.5 Summary

This chapter tested RQ4, RQ5, and RQ6 by deploying the model built in Chapter 5 into a live, closed-loop VR system. The clearest result is on RQ5. Model-driven and participant-driven control produced statistically equivalent outcomes on weighted game performance, subjective workload and learning rate. In dual-task conditions, the model-driven system also sustained participants at modestly harder difficulty levels than they chose for themselves, without a measurable performance cost. On the metrics that matter for a training system's purpose, in other words, letting the model set difficulty worked as well as letting participants set it themselves, and in the specific case of dual-task load, arguably better.

On RQ4, the model trained on single-task data did generalise to the dual-task context, retaining a consistent directional effect on difficulty control, though its real-time classification accuracy against participants' own difficulty ratings (36.7%, against a 33.3% chance baseline) was well below the offline figure from Chapter 5. A substantial part of this decline may reflect the instability of the subjective label the model is scored against once the task moves to continuous, low-contrast, real-time conditions. Chapter 7 examines this by comparing the same signals against subjective and objective targets. On RQ6, real-time adjustment did change participant difficulty judgments relative to the discrete conditions used in earlier chapters, producing a marked shift toward *Just Right* ratings; this is attributable both to the model's directional accuracy and to the stabilising effect of any small-step closed-loop controller.

Chapter 7 follows up directly on the question raised by the accuracy figures in this chapter, addressing RQ7. It asks whether subjective difficulty judgments remain a trustworthy target at all once the task environment moves from the discrete, high-contrast conditions used to train and validate a model to the continuous, low-contrast conditions of real-time deployment used here.

Chapter 7

Subjective and Objective References for Cognitive Load in Adaptive Cognitive Training

7.1 Introduction

This chapter addresses RQ7, introduced in Chapter 1: whether subjective workload assessments remain a reliable ground truth once an adaptive system moves from the discrete, high-contrast conditions used to train and validate a model to the continuous, low-contrast conditions of real-time deployment. Chapter 6 deployed such a system and reported a real-time classification accuracy below the offline figure from Chapter 5. It suggested that part of this gap may reflect the subjective label it was being scored against becoming a less stable target under real-time conditions, and not only a degradation of the model's underlying physiological signal. This chapter tests that argument directly, rather than treating it as an assumption. It uses data from 124 participants tested across the offline and real-time cohorts of the preceding chapters: 54 younger and 19 older adults from the offline data collection, plus 51 from the real-time deployment. We contrast subjective difficulty ratings against objectively derived, difficulty-weighted performance as competing targets for the same physiological and eye-tracking features. We then ask which one the data supports as the more stable and predictable quantity.

The transition from static, high-contrast task environments (Chapters 3 and 4) to the real-time, low-contrast adaptive system of Chapter 6 introduces a perceptual shift in the human-machine loop. This shift is characterised by a divergence between objective sensory input, latent cognitive load and the observer's subjective experience [75]. This chapter examines it as a possible reason for the lower accuracy of the model deployed in real time. Chapter 2 introduced the general reasons why subjective ratings are an imperfect reference. Judgments are relative to the range experienced, and metacognitive sensitivity is limited. Performance offers an alternative reference with its own drawbacks. A model validated on unseen participants may also fail when the conditions of use change. This chapter develops the parts of that argument that bear on the specific design used here. In this design, the steps between difficulty levels shrank from 0.5-1.5 s in the offline experiments to 0.1-0.3 s in the real-time system. The relevant topics are perceptual anchoring and signal detection, the physiology of effort, and the machine learning description of the resulting shift. The question became pressing when the accuracy drop reported in Chapter 6 appeared, and the remainder of this introduction sets out the argument before turning to the chapter's aim and methodology.

As Chapter 2 noted, the reliability of human introspection weakens as training environments become more unpredictable and sensorily ambiguous. Specifically, as the task noise increases, our ability to judge our own performance, often referred to as metacognitive sensitivity, diminishes [53]. Instead of accurately assessing difficulty, users often rely on biased mental shortcuts and struggle to adjust their expectations even when new information is presented [133].

This breakdown in belief updating is particularly critical in adaptive systems, where the internal state of the user (arousal and effort) should ideally align with the task demand to maintain optimal performance [168]. In this chapter, we explore the theoretical reasons for this phenomenon and examine whether participants' physiological profiles

align better with objective metrics (weighted performance) than with their own subjective judgments. Furthermore, we investigate whether this gap between feeling and performing differs between the high-contrast offline designs and the low-contrast, online adaptive environments.

7.1.1 The Physiology of Mental Effort

Four strands of literature are needed to build this argument. The first explains why physiological measures should track effort at all. The second explains why subjective judgment specifically breaks down under low-contrast conditions. The third is the machine-learning theory that explains why a model trained on one and scored against the other would show exactly the kind of accuracy drop reported in Chapter 6. The fourth concerns the limits of objective performance as an alternative target. This section addresses the first of these.

The assessment of cognitive workload relies on objective measurements of physiological responses that reflect the recruitment of mental resources [80]. In sustained attention tasks, these responses are largely driven by the autonomic nervous system's reaction to task-induced demand [63].

Task-evoked pupil dilation (PD) is among the most sensitive indicators of mental effort [179]. Pupillary responses during cognitive tasks are directly linked to the Locus Coeruleus-Norepinephrine (LC-NE) system, which modulates cortical gain [11]. When task difficulty increases, the LC increases its activity to facilitate resource mobilization, resulting in pupil diameter increases [179]. However, in low contrast environments where difficulty shifts are subtle (e.g., 0.1 increments), the physiological signal may remain within the range of baseline noise, challenging the discriminability of the model [20].

Heart Rate Variability (HRV), particularly the standard deviation of normal-to-normal intervals (SDNN), provides a window into the sympathovagal balance [92]. According to Motivational Intensity Theory, effort rises with subjective task difficulty only as long as success is perceived as possible and the effort is justified [29, 140]. In a low contrast cohort, a stimulus may be perceived as *Too Difficult* simply because it represents the highest value encountered, leading to potential disengagement and a physiological withdrawal where HRV returns toward baseline despite high objective demand [150].

7.1.2 Psychological Mechanisms of Perceptual Judgment and Anchoring

Establishing that physiological responses track effort is not, by itself, enough to explain a divergence between physiology and self-report; it also has to be shown that self-report is vulnerable to a specific, predictable distortion under the conditions used in Chapter 6. The core challenge of the present experiment lies in the psychology of categorical judgment. Participants do not label difficulty against an absolute scale; instead, judgments are relative and contextual [112].

Parducci's Range-Frequency Theory (introduced in Chapter 2) posits that judgment is a compromise between the range principle (dividing the perceived range into equal segments) and the frequency principle (the tendency to use categories with equal frequency) [122, 123]. For a cohort exposed only to a narrow range (0.3 and 0.1 shifts), the subjective midpoint is anchored by these subtle values. Consequently, a difficulty level that the model (trained on 0.5 to 1.5 shifts) considers *Just Right* may be labeled *Too Difficult* by the user because it sits at the top of their current experienced range [177].

A further question is whether two difficulty levels that are close together can be told apart at all, by the physiological signal or by the participant, when both are noisy. Signal Detection Theory (SDT) is the standard framework for this, because it separates how far apart two states are (sensitivity) from where the decision boundary is placed (the cri-

terion), which corresponds to the distinction between how discriminable the physiological states are and how a participant's rating threshold shifts [57]. It is used here to reason about the expected direction of the effect, and d' itself is not estimated in this thesis. The distance between physiological signal distributions for different load levels is defined as d', while the internal threshold is the criterion C [108]:

$$d' = \frac{\mu_{signal} - \mu_{noise}}{\sigma}$$

Training on high-contrast data ($0.5 - 1.5$ shifts) supports a high d'. However, in real-time deployment with $0.1 - 0.3$ shifts, the distributions are likely to overlap more, lowering d' and potentially bringing it below the Just Noticeable Difference (JND) [108].

7.1.3 Computational Challenges: Covariate and Domain Shift

The anchoring account above is a psychological explanation; the same phenomenon also has a precise machine-learning description, which is what actually predicts the size and direction of an accuracy drop like the one reported in Chapter 6. The transition from a high-contrast training cohort to a low-contrast real-time cohort introduces the complex problem of covariate shift [159, 148], outlined in Chapter 2. The simplest case is that only the inputs change while their meaning stays the same. In a standard machine learning framework, we assume the marginal distribution of the input features $P(x)$ changes, while the conditional distribution $P(y|x)$, the relationship between features and label, remains invariant:

$$P_{train}(x) \neq P_{test}(x) \quad \text{and} \quad P_{train}(y|x) = P_{test}(y|x)$$

However, psychological anchoring effects suggest that this assumption is violated in two distinct ways, transforming a simple covariate shift into a more volatile domain shift [102]:

1. **Perceptual Recalibration:** Sustained exposure to low-contrast stimuli shifts user sensitivity thresholds. Over time, an identical physiological signal (x) may correspond to a different internal arousal state, as the autonomic nervous system adapts to the new environmental norm. This is consistent with Signal Detection Theory (SDT), where the internal criterion (C) is dynamic rather than fixed [57].

2. **Labeling Inconsistency:** Induced by psychological anchoring, users may apply the same behavioral label (y) to disparate internal states depending on the experienced range of difficulty. According to Range-Frequency Theory, the subjective midpoint is anchored to the context of the current task [122]. Consequently, a difficulty increment that feels Low in a low-contrast context might be perceived as High in a high-contrast context due to the magnitude of the range of the difficulty change.

Together, these effects suggest that $P(y|x)$ is non-stationary, constituting a form of concept shift that is not reducible to simple covariate shift. If $P(y|x)$ is unstable, the source domain labels, typically subjective reports, can struggle to serve as a stable reference for training [53]. From a system-design perspective, this implies that source domain labels may not be trusted without continual online recalibration. However, online recalibration introduces severe costs in system responsiveness and stability. To circumvent this, we propose that **objective performance metrics** serve as a more robust ground truth. By mapping physiology (x) to performance (y_{perf}), we bypass the psychological anchoring of the user, treating the subjective mismatch as a bias to be corrected by the model rather than as noise to be integrated into it.

7.1.4 Practical Limitations of Performance-Driven Ground Truths

If subjective labels are unstable for the reasons set out above, the natural response is to use objective performance as the training target, as proposed at the end of the previous section. Chapter 2 set out the limits of performance as a reference. It can mask compensatory effort, particularly in older adults, who may maintain performance through the recruitment of additional neural resources, so that physiological strain is not reflected in

the performance output [138]. It can also hide a ceiling: a system that assumes the user is in an optimal state because performance is high may fail to recognise that the user is operating at maximum capacity, leaving little room for growth or learning [80]. Two consequences follow for this chapter. Performance-based targets resolve the problem of label non-stationarity but do not replace physiology, so performance and physiological trends are interpreted together [168, 75]. In addition, the weighted-performance target used here is defined relative to each participant's own performance (Section 7.2.3), which shapes how its results should be read.

7.1.5 Chapter Aim and Contributions

As discussed in earlier chapters, the majority of the papers relating to cognitive load estimation using eye-tracking and physiological metrics use highly discrete, high-contrast ways of deciding low, medium and high cognitive load, the most prevalent one being task-stacking (adding a task for each cognitive load level, one for low, two for medium, three for high [178]). Chapter 6 deployed a model trained on discrete targets in a high-contrast, single-task scenario within low-contrast, single- and dual-task adaptive environments. This chapter tests RQ7 and H7 directly by examining how participant ratings of their own cognitive load relate to objective metrics of their performance, particularly in a real-time adaptive task design, and by testing whether the same physiological data used throughout this thesis predicts objective performance more reliably than it predicts subjective judgment.

7.2 Methodology

7.2.1 Experimental Design Overview

This chapter tests RQ7/H7 (Chapter 1) through secondary analysis of data already collected in Chapters 3 through 6; no new participants were recruited. Rather than manipulating a variable experimentally, the chapter treats two properties of the existing datasets as factors to be compared directly against each other using the same physiological and eye-tracking features as predictors throughout.

Factors compared (analogous to independent variables). (1) Target/ground-truth type: Subjective Judgment (the participant's own difficulty rating, used as the classification target in every preceding chapter) versus Weighted Performance (an objective, difficulty-adjusted performance score, Section 7.2.3 below), compared as competing prediction targets for the same input features. (2) Contrast/deployment context: high-contrast offline data (Chapters 3, 4) versus low-contrast real-time data (Chapter 6), and within the latter, single-task versus dual-task, compared as training and testing cohorts in the machine learning experimental matrix (Table 7.1).

Dependent variables. Model classification performance (accuracy, F1, precision, recall, AUC, AUPRC) for each of the 15 target/cohort combinations in Table 7.1; agreement statistics (Cohen's κ) between the subjective and objective three-class labellings of the same sessions; the sample size at which learning curves plateau, as a test of whether more data or a different ground truth is the limiting factor; and the qualitative cluster structure of the same physiological feature space when coloured by each ground truth (t-SNE).

Measures. The same windowed eye-tracking and physiological feature set used throughout this thesis (Section 5.2.3), plus the weighted-performance target variable defined in Section 7.2.3, plus the categorical subjective ratings already collected in earlier chapters.

The analyses below are presented in the order of the questions they address. Target distribution and subjective/objective agreement come first. The head-to-head predictive comparison between the two ground-truth types follows, which speaks most directly to RQ7. The learning-curve analysis of sample size comes next, and the t-SNE visualisation, used to interpret why one target predicts better than the other, comes last. Each is presented as a question, the method used to answer it, the results, and their interpretation.

7.2.2 Participants, Task, Setup, Preprocessing

This chapter examines the data as collected in earlier chapters, meaning the participants, experimental design and preprocessing methodology were already described in-depth in earlier chapters.

7.2.3 Target Variable Formulation

The only methodological addition in this chapter is the creation of discrete classes based on weighted performance, which was in-depth described in Chapter 6. The weighted performance was split for each participant after z-score normalization of the weighted performance for each participant based on the ± 0.75 SD threshold: Low Load: $Z \leq -0.75$ Medium Load: $-0.75 < Z < 0.75$ High Load: $Z \geq 0.75$.

7.3 Analyses and Results

The four analyses are presented in turn. Each begins with the question it addresses and the method used, followed by the results and their interpretation, so that the reasoning for each analysis can be read in one place. The overall discussion follows at the end of the chapter.

7.3.1 Analysis 1: Target Distribution and Subjective-Objective Alignment

Question and Method

The question is whether participants' own ratings of difficulty correspond to how they actually performed, and whether that correspondence differs between the high-contrast offline data and the low-contrast real-time data.

The initial analysis examines the distribution of the weighted performance classes across the 3 datasets of interest - the original dataset that the real-time model was trained on, the real-time single-task dataset, and the real-time dual-task dataset. We compare the class balances and contrast them against the subjective distributions presented in Chapter 6 using agreement analysis.

To ensure that the machine learning pipeline captures genuine physiological markers rather than dataset-specific task biases, it was necessary to first validate the comparability of the objective target variables. Participant-level Z-scoring (± 0.75 SD) was applied to the weighted performance metrics to normalize individual differences in baseline gaming skill.

Agreement between the subjective and objective labellings is quantified with Cohen's quadratic weighted kappa. This is chosen because the three load classes are ordered. A disagreement between Low and High is more serious than one between adjacent classes, and the weighting penalises it accordingly.

Results

A Chi-Square test of independence was performed to examine the relationship between the experimental cohort (Offline Original, Real-time Single-Task, Real-time Dual-Task) and the resulting distribution of objective cognitive load classes (Low, Medium, High). The test revealed that the cohorts possessed statistically similar load distributions, $\chi^2(4, N = 3456) = 5.03, p = .285$. Across all three cohorts, the participant-level normalization yielded a stable distribution. Approximately half of the gameplay levels fell into the Medium Load class (49.9 to 52.2%), and the remaining data were distributed symmetrically between the Low Load (24.2 to 26.3%) and High Load (21.6 to 25.9%) extremes (see Figure 7.1). This statistical parity confirms that the normalization procedure was effective, justifying the amalgamation of these datasets for the subsequent predictive models.

Following the validation of the objective performance labels, the alignment between these mathematically derived performance states and the participants' self-reported (subjective) cognitive load was evaluated using Cohen's quadratic weighted kappa (κ). Across all three cohorts, participants demonstrated only fair agreement between their perceived cognitive strain and their actual gameplay performance.

The Offline (Original) cohort exhibited the highest alignment, though it remained modest ($\kappa = 0.290$, 45.6% strict agreement). Notably, both Real-time cohorts yielded identical kappa scores ($\kappa = 0.244$), despite the Dual-Task paradigm showing a slightly higher strict agreement percentage (51.7% vs 44.2% in the Single-Task). To unpack the exact nature of this discrepancy, the proportional breakdown of actual objective states was plotted against each self-reported state (see Figure 7.2).

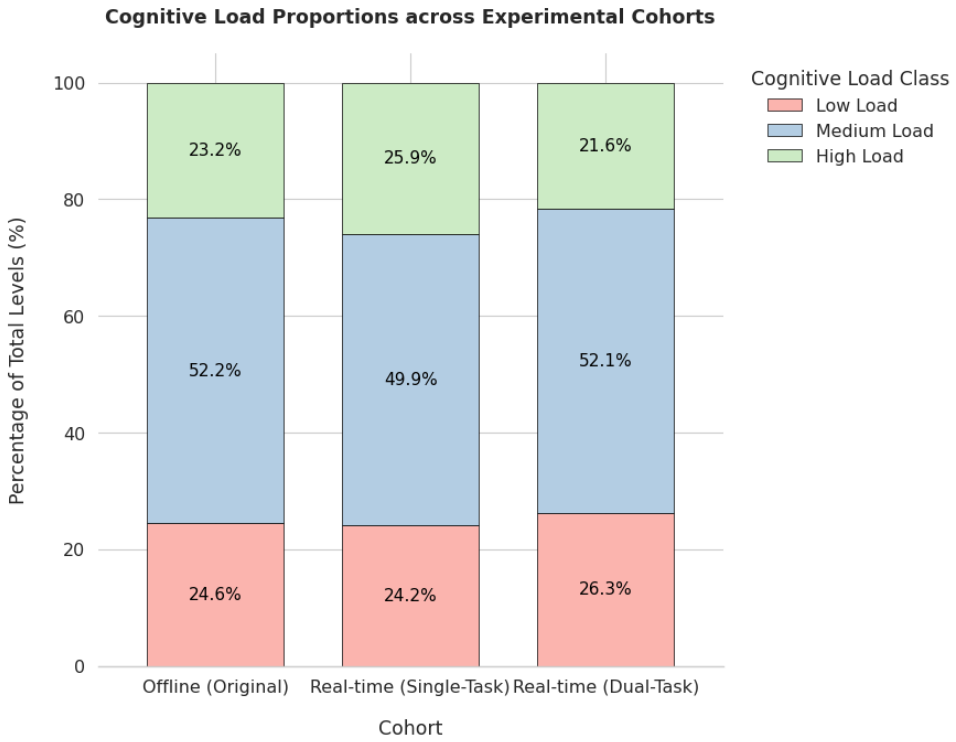


Figure 7.1: Distribution of discretized objective cognitive load classes across the three experimental cohorts. The normalization process successfully maintained comparable class balances regardless of the specific task paradigm.

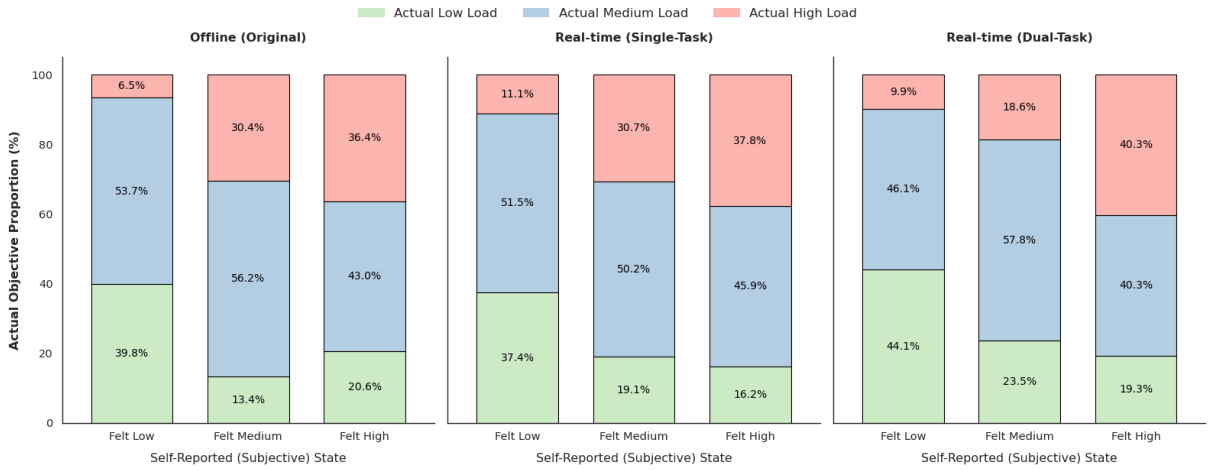


Figure 7.2: The Subjective-Objective Mismatch. Stacked bar charts illustrating the true objective performance state of participants conditional upon their self-reported cognitive load.

Visualizing these distributions reveals exactly why the agreement scores remained low. For instance, during the Real-time Dual-Task, when participants actively reported feeling High Load, only 40.3% of those instances corresponded to actual High Load objective performance (against a base rate of about 22 to 26% for High Load). The remaining 59.7% of the time, their performance was in the Medium or Low classes, indicating a substantial mismatch. This persistent subjective-objective mismatch is consistent with reports that users are often poor estimators of their own cognitive limits (see Chapter 2). Consequently, this provides a rationale for utilising physiological sensor data, as post-hoc self-reported metrics alone fail to consistently capture true task-induced cognitive degradation.

Interpretation

Analysis 1 revealed a substantial mismatch between self-reported cognitive load and actual gameplay performance ($\kappa \approx 0.24$ to 0.29). This divergence is consistent with a reduction in the metacognitive sensitivity of the observer as task environments transition toward sensory ambiguity [53]. This mismatch was asymmetric with participants frequently reporting experiencing High Load (*Too Difficult*) while their objective performance indicated optimal or even easy gameplay, potentially indicating that the lack of high-contrast anchors led to suboptimal belief updating [133, 168].

A possible explanation draws on Parducci's Range-Frequency Theory [122, 123] and Signal Detection Theory (SDT) [57]. In the low-contrast real-time environment, where difficulty increments are subtle, the physiological discriminability between states narrows [108]. Without absolute high-contrast anchors, participants may have dynamically recalibrated their internal criterion. Consequently, any local maximum in difficulty may have been perceived as *Too Difficult* simply because it was the hardest state encountered within that specific micro-context, rather than because it exceeded their actual cognitive capacity.

7.3.2 Analysis 2: Predictive Modelling of Objective Cognitive Load

Question and Method

The question is whether the same physiological and eye-tracking features predict objectively derived load better than they predict participants' subjective judgments (H7), and whether this holds across cohorts and task types.

The following analysis examines whether the objective targets (weighted performance) map better to the physiological data than the subjective judgments. 15 XGBoost models were fitted with the goal of optimising the performance on the real-time dataset. See Table 7.1 for the full breakdown, including the approximate number of participants contributing to each training cohort.

Table 7.1: The 15 XGBoost analyses performed to examine the difference in predictive performance between weighted performance and subjective judgments. Training-cohort participant counts are given for reference; row-level sample sizes (number of sub-session/level observations) vary by cohort in the same proportions as Chapters 3-6 and are not restated per row here.

Target Variable	**Training Cohort**	**Training N (participants)**	**Testing Cohort**
Subjective Judgment	Real-time	51	Real-time - Single Task
Subjective Judgment	Real-time	51	Real-time - Dual Task
Subjective Judgment	Real-time	51	Real-time - Combined
Subjective Judgment	Real-time + offline	$51 + 73 = 124$	Real-time - Single Task
Subjective Judgment	Real-time + offline	$51 + 73 = 124$	Real-time - Dual Task
Subjective Judgment	Real-time + offline	$51 + 73 = 124$	Real-time - Combined
Weighted Performance (Z-Score)	Real-time	51	Real-time - Single Task
Weighted Performance (Z-Score)	Real-time	51	Real-time - Dual Task
Weighted Performance (Z-Score)	Real-time	51	Real-time - Combined
Weighted Performance (Z-Score)	Real-time + offline	$51 + 73 = 124$	Real-time - Single Task
Weighted Performance (Z-Score)	Real-time + offline	$51 + 73 = 124$	Real-time - Dual Task
Weighted Performance (Z-Score)	Real-time + offline	$51 + 73 = 124$	Real-time - Combined
Weighted Performance (Z-Score)	Offline (all ages combined)	73	Offline (all ages combined)
Weighted Performance (Z-Score)	Offline (aged 18-58)	54	Offline (aged 18-58)
Weighted Performance (Z-Score)	Offline (aged 60+)	19	Offline (aged 60+)

The offline participant count used here (73: 54 younger, 19 older) is the post-exclusion figure used consistently in Chapter 5. For the offline cohorts, this corresponds to 1,736 sub-session-level rows (Offline, all ages), 1,280 rows (Offline, aged 18-58), and 456 rows (Offline, aged 60+), matching Chapters 3-5.

Accuracy, F1, Precision and Recall are calculated at the participant level to report mean and standard deviation. AUC and AUPRC are calculated globally, by aggregating all predictions into a single array, to account for participants who may not have experienced all three classes.

Table 7.1 lists the 15 models, which fall into four groups that each answer a different question. **Subjective versus objective target (rows 1 to 12).** The first twelve models are the head-to-head test of H7. The participants, features and validation scheme are the same, and only the target changes, from the participant's rating (rows 1 to 6) to weighted performance (rows 7 to 12). If ratings are a less stable reference under low contrast, the objective target should be predicted better. **Real-time data alone versus real-time plus offline data.** Each target is modelled twice, trained on the 51 real-time participants alone and on the real-time plus the 73 offline participants (124 in total). This tests whether adding high-contrast data helps or hurts prediction of the low-contrast sessions, which bears on the dataset-shift argument set out in the Introduction. **Test cohort.** Each real-time model is evaluated on single-task, dual-task and combined sessions, to show whether performance depends on task complexity. **Offline baselines (rows 13 to 15).** The last three models use the objective target on the offline data only, for all ages, for participants aged 18 to 58 and for participants aged 60 and over. They give a reference for the same features in the high-contrast setting and show whether older adults are predicted differently. They have no subjective counterpart in this table because the offline subjective results are reported in Chapters 3 and 4.

Results

Given the mismatch between participants' self-reported cognitive states and their actual performance, the machine learning phase focused on predicting the objectively derived cognitive load classes (Low, Medium, and High). A matrix of 15 models was trained to evaluate the predictive viability of physiological features across both the original high-

contrast environment (Offline cohort) and the new low-contrast environments (Real-time Single and Dual-Task cohorts). The complete performance metrics for all 15 experimental configurations, including strict participant-level Leave-One-Group-Out Cross-Validation (LOOCV), are detailed in Table 7.2.

To establish the foundational validity of the physiological feature space, the analysis first examined two pure baselines. One was a model trained and tested exclusively on the Real-time dataset. The other was a model trained and tested exclusively on the Offline dataset. Evaluating these models side-by-side indicates whether the autonomic nervous system's response to cognitive load remains consistent across vastly different visual and auditory VR paradigms.

As shown in Figure 7.3, both baseline models achieved moderate, comparable predictive performance. The Receiver Operating Characteristic (ROC) curves reveal that the models were particularly adept at identifying the extreme states of cognitive load. For the Real-time model, the Low and High Load classes achieved AUCs of 0.79 and 0.76, respectively. The Offline model mirrored this nearly identically, achieving AUCs of 0.77 for Low Load and 0.78 for High Load.

The normalized confusion matrices further contextualize this performance. Both models successfully identified their target classes at rates above random chance ($\sim 33\%$), with recall rates ranging from 50.4% to 60.6% for the extreme classes. Notably, when the models did err, the misclassifications were almost entirely restricted to adjacent classes. For example, in the Real-time model, when a participant was experiencing true High Load, the model incorrectly predicted Low Load only 6.1% of the time; the vast majority of its errors (43.5%) were adjacent misclassifications into the Medium Load category. This fuzzy boundary effect around the Medium class (reflected in lower AUCs of 0.62 and 0.64) is common in psychophysiological modeling, as the transition between optimal flow and cognitive overload is a continuous physiological gradient rather than a hard categorical threshold.

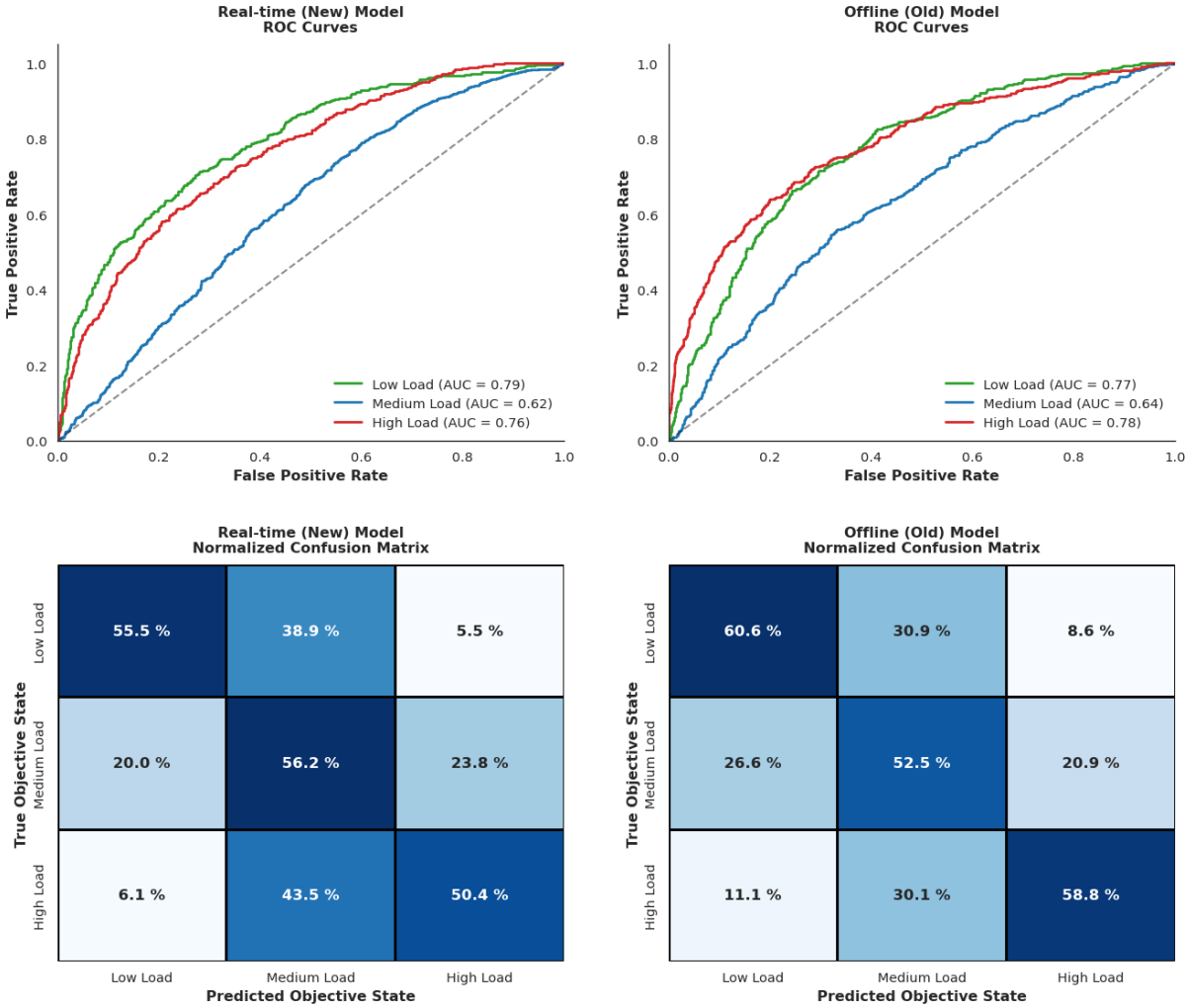


Figure 7.3: Performance baselines for the Real-time and Offline predictive models. The top row displays the One-vs-Rest Receiver Operating Characteristic (ROC) curves, highlighting moderate discrimination for extreme cognitive states. The bottom row displays the normalized confusion matrices, indicating that misclassifications primarily occur between adjacent classes.

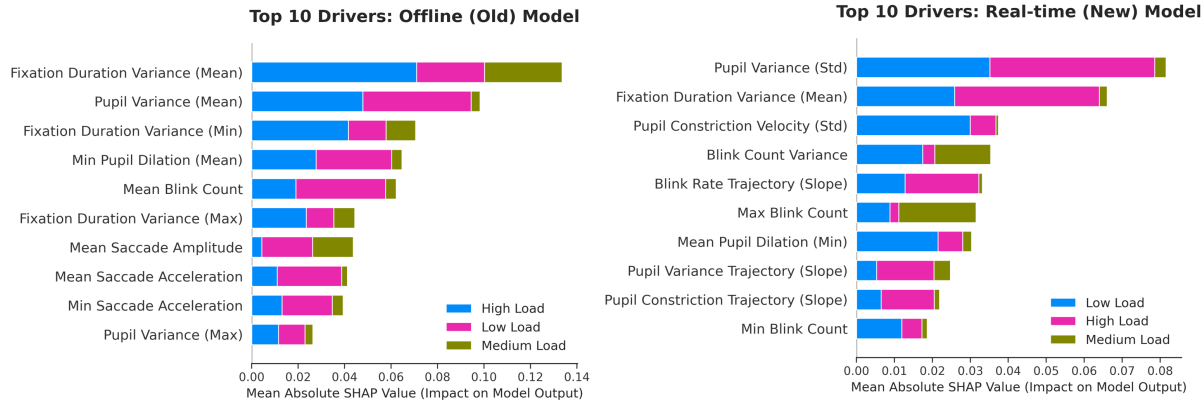


Figure 7.4: SHAP (SHapley Additive exPlanations) summary plots for the top 10 predictive features of both baseline models. Despite the differences in VR environments, both models overwhelmingly relied on identical ocular metrics, specifically pupil variance and fixation duration variance, to make their predictions.

To understand the driving forces behind these predictions, SHapley Additive exPlanations (SHAP) were extracted to determine global feature importance (see Figure 7.4). Despite the Real-time dataset featuring a low-contrast environment with a compounding auditory dual-task, and the Offline dataset featuring a high-contrast visual-only environment, both models selected highly similar physiological features.

Oculometrics dominated the predictive algorithms. Variance in Pupil Dilation (Std/Mean) and Variance in Fixation Duration emerged as the most critical drivers across both datasets. In the Real-time model, Pupil Variance exhibited the highest mean absolute SHAP value, strongly driving predictions for High and Low load states. The Offline model similarly relied on Fixation Duration Variance and Pupil Variance as its top two features. The presence of blink-related features (e.g., Blink Count Variance, Mean Blink Count) and saccadic movements in the top ten features of both models is consistent with ocular behaviour being a useful indicator of cognitive load across the two VR tasks studied here.

Reading across Table 7.2, the clearest difference appears to be between targets. In all six matched pairs, accuracy and AUC are higher for the objective target than for the subjective one. For example, accuracy is 0.582 against 0.488 for the real-time single-task models, and 0.537 against 0.490 for the combined models trained on real-time and offline data. Within the objective target, accuracy varies little across cohorts and training sets (0.529 to 0.587). Adding the offline data to the training set did not improve the objective models (accuracy fell by 0.01 to 0.035, for example from 0.582 to 0.547 for real-time single-task). It had mixed effects on the subjective models. This near-constant level is consistent with the plateau found in Analysis 3.

Table 7.2: Machine Learning Performance Metrics across 15 Experimental Configurations

Target	**Train Set**	**Test Set**	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC**	**AUPRC**
Subjective	Real-time	Real-time (Single)	0.488 ± 0.134	0.437 ± 0.137	0.441 ± 0.153	0.403 ± 0.136	0.676	0.494
Subjective	Real-time	Real-time (Dual)	0.490 ± 0.151	0.414 ± 0.133	0.403 ± 0.156	0.363 ± 0.123	0.590	0.402
Subjective	Real-time	Real-time (Combined)	0.506 ± 0.104	0.444 ± 0.098	0.466 ± 0.126	0.417 ± 0.098	0.641	0.454
Subjective	Real-time + Offline	Real-time (Single)	0.522 ± 0.142	0.458 ± 0.180	0.447 ± 0.167	0.412 ± 0.161	0.645	0.453
Subjective	Real-time + Offline	Real-time (Dual)	0.500 ± 0.153	0.386 ± 0.134	0.358 ± 0.132	0.332 ± 0.110	0.572	0.382
Subjective	Real-time + Offline	Real-time (Combined)	0.490 ± 0.115	0.417 ± 0.121	0.413 ± 0.138	0.389 ± 0.120	0.617	0.422
Objective	Real-time	Real-time (Single)	0.582 ± 0.140	0.584 ± 0.149	0.600 ± 0.173	0.559 ± 0.159	0.749	0.597
Objective	Real-time	Real-time (Dual)	0.539 ± 0.131	0.522 ± 0.156	0.490 ± 0.154	0.478 ± 0.148	0.695	0.524
Objective	Real-time	Real-time (Combined)	0.547 ± 0.116	0.539 ± 0.126	0.543 ± 0.132	0.526 ± 0.128	0.722	0.555
Objective	Real-time + Offline	Real-time (Single)	0.547 ± 0.123	0.537 ± 0.143	0.556 ± 0.161	0.517 ± 0.147	0.728	0.569
Objective	Real-time + Offline	Real-time (Dual)	0.529 ± 0.113	0.504 ± 0.152	0.461 ± 0.147	0.458 ± 0.138	0.686	0.515
Objective	Real-time + Offline	Real-time (Combined)	0.537 ± 0.101	0.521 ± 0.118	0.524 ± 0.110	0.511 ± 0.111	0.707	0.541
Objective	Offline (all ages combined)	Offline (all ages combined)	0.560 ± 0.131	0.554 ± 0.147	0.573 ± 0.151	0.542 ± 0.145	0.730	0.565
Objective	Offline (aged 18-58)	Offline (aged 18-58)	0.587 ± 0.115	0.576 ± 0.149	0.568 ± 0.146	0.553 ± 0.144	0.744	0.593
Objective	Offline (aged 60+)	Offline (aged 60+)	0.557 ± 0.088	0.534 ± 0.117	0.526 ± 0.110	0.507 ± 0.108	0.699	0.541

Interpretation

When the machine learning pipelines were trained on objective, performance-derived targets, predictive accuracy improved across both the offline and real-time datasets. The models consistently achieved Global AUCs between 0.69 and 0.75, discriminating between objectively derived low and high load states.

Moreover, the F1 score gap between subjective judgments and objective performance is larger in the real-time datasets than in the offline datasets. For younger participants offline, F1 was 49.1 for subjective judgments against 55.3 for objective performance. For the real-time single task, it was 40.3 against 55.9. This difference suggests that the low-contrast environment of the real-time task exacerbates the participant bias as compared to the high-contrast environment which allowed participant judgments to align more accurately with their objective performance.

The SHAP feature importance plots revealed that despite drastic differences in task mechanics (visual-only vs. dual-task) and contrast gradients, the models consistently relied on oculometrics, specifically variance in pupil dilation and fixation duration. This reliance on pupillometry is consistent with the role of the Locus Coeruleus-Norepinephrine (LC-NE) system in effort [11, 179].

7.3.3 Analysis 3: Sample Size Sufficiency and Learning Curves

Question and Method

A frequent limitation in psychophysiological virtual reality research is the reliance on small, potentially underpowered participant cohorts. To empirically validate the statistical power of the models and determine if the collected sample sizes were sufficient to capture the variance of human cognitive load, a comprehensive learning curve analysis was conducted. It tested whether merging datasets, isolating specific tasks, or segregating by age group altered the data volume required for model performance.

The goal of this analysis is to determine the minimum viable participant count required to map physiological signals to objective cognitive load, and to examine whether the current sample size has reached the point of diminishing returns. Therefore, a participant sub-sampling analysis was performed on the best-performing pipeline configuration, starting at 5 participants and randomly adding 5 at a time up to the total participant count, resampling multiple times at each sample size level to obtain confidence intervals. The results are visualised in a plotted curve where the x-axis is 'Number of Participants in Training Set' and the y-axis is 'Global AUC'. The curve shows whether performance flattens as participant number increases, which would indicate that the current number of participants is sufficient, or not, which would indicate that further data collection is likely to improve the results.

Using the optimal hyperparameters and the top selected features from the objective models, a participant-scaling simulation was executed. The training pool was artificially constrained, starting at $N = 5$ participants and incrementally increasing by steps of 5 up to the maximum available cohort size for that specific condition. At each step, participants were randomly sampled, and a strict Leave-One-Participant-Out Cross-Validation

(LOOCV) pipeline was trained and evaluated. To account for sampling variance, this process was iterated multiple times per step (starting at 10 and going down to 2 with increasing number of participants per step) with different random seeds, allowing for the calculation of the mean Global Macro AUC and its standard deviation.

Results

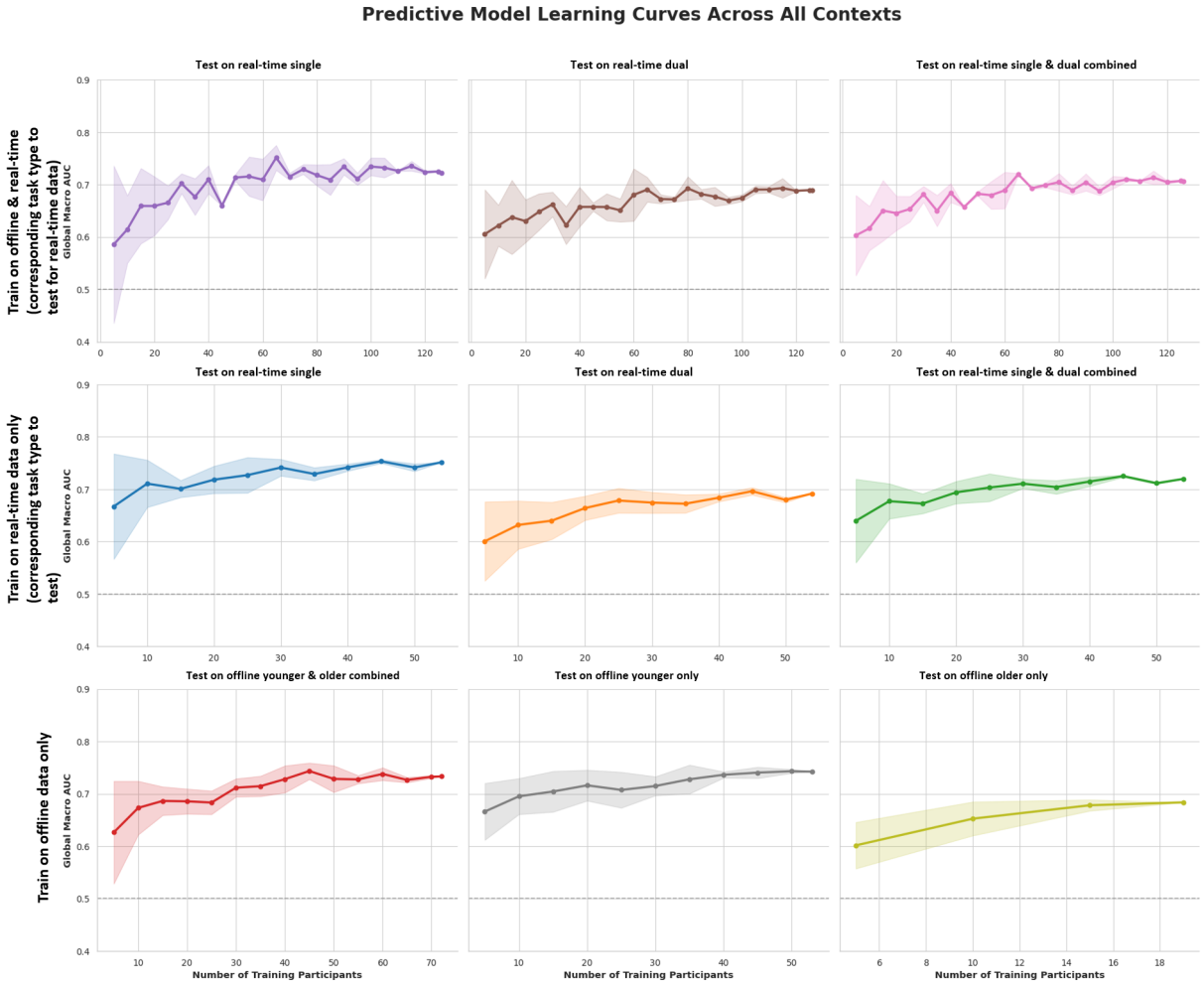


Figure 7.5: A 3x3 matrix of predictive model learning curves demonstrating sample size sufficiency across all experimental contexts. The x-axis represents the number of participants available to the algorithm, while the y-axis shows the resulting LOOCV Global Macro AUC (with shaded regions representing ± 1 standard deviation). Across pure, combined, and age-isolated cohorts, all models exhibit a clear performance plateau, suggesting diminishing returns from further participant recruitment.

The resulting learning curves (Figure 7.5) show the algorithms' learning capacity to predict weighted performance across all permutations. A consistent pattern emerged regardless of the specific dataset or target: the models exhibited steep initial learning trajectories with highly unstable variance bounds when constrained to fewer than 15 participants. However, as sample sizes increased, the performance trajectories flattened, though the point of saturation varied based on environmental complexity.

When models were trained and tested within the same homogeneous environment (the middle and bottom rows), learning saturated relatively early. For the Real-time models, a clear performance asymptote was achieved near $N = 30$ to 40, stabilising at a Global AUC between 0.69 and 0.75. The Offline models exhibited near-identical behaviour; the combined and younger-only (aged 18-58) cohorts plateaued around $N = 30$, reaching AUCs of approximately 0.74. Even the older cohort (aged 60+), limited to a maximum N of 19, appears to plateau.

Conversely, when the algorithm was tasked with finding universal physiological markers across merged datasets (Top Row: Train on Offline & Real-time), the learning curve extended further. Because the algorithm had to map overlapping variance between the offline high-contrast task and the real-time low-contrast task, it required approximately $N = 60$ to 80 participants before the standard deviation bands tightened and the AUC stabilised near 0.71.

The stabilisation of both the mean AUCs and the tightening of the variance bands across all nine contexts suggests that the algorithms had extracted most of the predictive value available in these features. This plateau is evidence that the final participant cohorts were sufficient for these models, and it suggests that the persistent performance ceilings (hovering around 0.70 to 0.75 AUC) are unlikely to be caused by sample size alone. They may also reflect limits of the features, the labels and the model class.

Interpretation

A critical limitation often cited in psychophysiological VR research is sample size. The learning curve analysis (Analysis 3) suggests that the collected cohorts ($N = 51$ for Real-time and $N = 73$ for Offline, after exclusions) were sufficient for the models to reach a plateau.

When trained on homogeneous datasets, the algorithms reached a clear performance plateau around $N = 30$ to 40. Interestingly, when the model was tasked with bridging the high-contrast Offline and low-contrast Real-time environments, the learning curve extended further, requiring 60 to 80 participants to saturate. This is consistent with the computational friction of covariate shift [159, 148]. It may also indicate that the non-stationarity of $P(y|x)$, driven by the user's internal perceptual recalibration, requires a larger data pool to resolve than a simple covariate shift where only the environment $P(x)$ changes [159, 102]. Merging heterogeneous datasets would also be expected to need more data.

The stabilisation of these curves at approximately 0.75 AUC suggests that the models are no longer strongly data-limited. The plateau may reflect the accuracy ceiling of current sensor fidelity and the natural fuzziness of physiological boundaries (such as the continuous transition through medium (*Just Right*) load).

An unexpected but vital insight emerged from isolating the older adult cohort (aged 60+). Despite having a heavily constrained sample size ($N < 20$), the older cohort displayed a distinct, rapid learning curve plateau. Furthermore, the machine learning models trained on this demographic exhibited the lowest variance (± 0.088 Accuracy Std) of any objective configuration.

While older adults may possess different physiological baselines and potentially lower absolute cognitive ceilings due to age-related neurocognitive changes, their autonomic responses to approaching those limits appear to be consistent. This questions the assumption that physiological models for older adults require substantially more data due to age-related noise. It suggests that their physiological boundaries between Low, Medium and High load may be definable and predictable.

7.3.4 Analysis 4: Visualising the Data through t-SNE Plotting

Question and Method

The question is why one target should be predicted better than the other, and whether the physiological data itself shows more structure when its points are coloured by performance than when they are coloured by subjective ratings.

To visualise how the classes are arranged in the feature space used by the classifiers, each 60-second task level was embedded in two dimensions with t-SNE. The input was the same 112-dimensional space used in the modelling, with no feature selection and no dimensionality reduction applied beforehand. There were 28 base signals: 21 eye-tracking measures covering fixations, saccades, pupil and blinks, and 7 physiological measures covering heart rate, heart rate variability and electrodermal activity. These were z-scored within each participant across all of their windows, which removes between-person differences in baseline level. The four windows of each level were then summarised by four statistics per signal (mean, standard deviation, linear slope across the window sequence and range), giving $28 \times 4 = 112$ features per level. Missing values were median-imputed and the features were standardised before embedding. A separate embedding was fitted for each cohort (real-time single-task, real-time dual-task, offline younger and offline older). t-SNE used a perplexity of 30, Euclidean distance, an early exaggeration of 12, an automatically set learning rate, 2,000 iterations and a fixed random seed. Class labels played no role in the embedding and were used only for colouring.

For each cohort, the embedding was plotted twice. The first plot was coloured by the participant's subjective rating (*Too Easy*, *Just Right*, *Too Difficult*). The second was coloured by the objective label derived from the participant's within-person z-scored, difficulty-weighted task performance (below -0.75, between, and above $+0.75$ standard deviations, labelled Low, Medium and High Load). Because the coordinates are identical in the two plots, differences in how the colours are arranged reflect the labels and not the layout. Any clustering visible in the colours may indicate a relationship between the participant's physiology and that label (Figure 7.6).

The settings were not tuned to the labels. A sweep over scaling, perplexity, distance metric and exaggeration did not produce a visibly better arrangement than these defaults, so the defaults are reported. The embeddings are for visual inspection only, since distances between clusters and cluster sizes in t-SNE are not directly interpretable. Predictive performance in Analysis 2 is the quantitative counterpart.

Results

Figure 7.6 shows the embedding of each cohort coloured by the participant's subjective rating (left column) and by the objective label (right column). The subjective ratings *Too Easy*, *Just Right* and *Too Difficult* are shown in the same colours as Low, Medium and High Load.

In the real-time single-task cohort (top row) the objective labels show clearer arrangement across the three classes than subjective ratings. Low Load levels are concentrated in the upper left of the embedding, High Load levels in the centre and lower centre, and Medium Load levels are spread throughout and overlap both. The subjective labels show a weaker version of the same layout, with Low Load levels toward the top and High Load

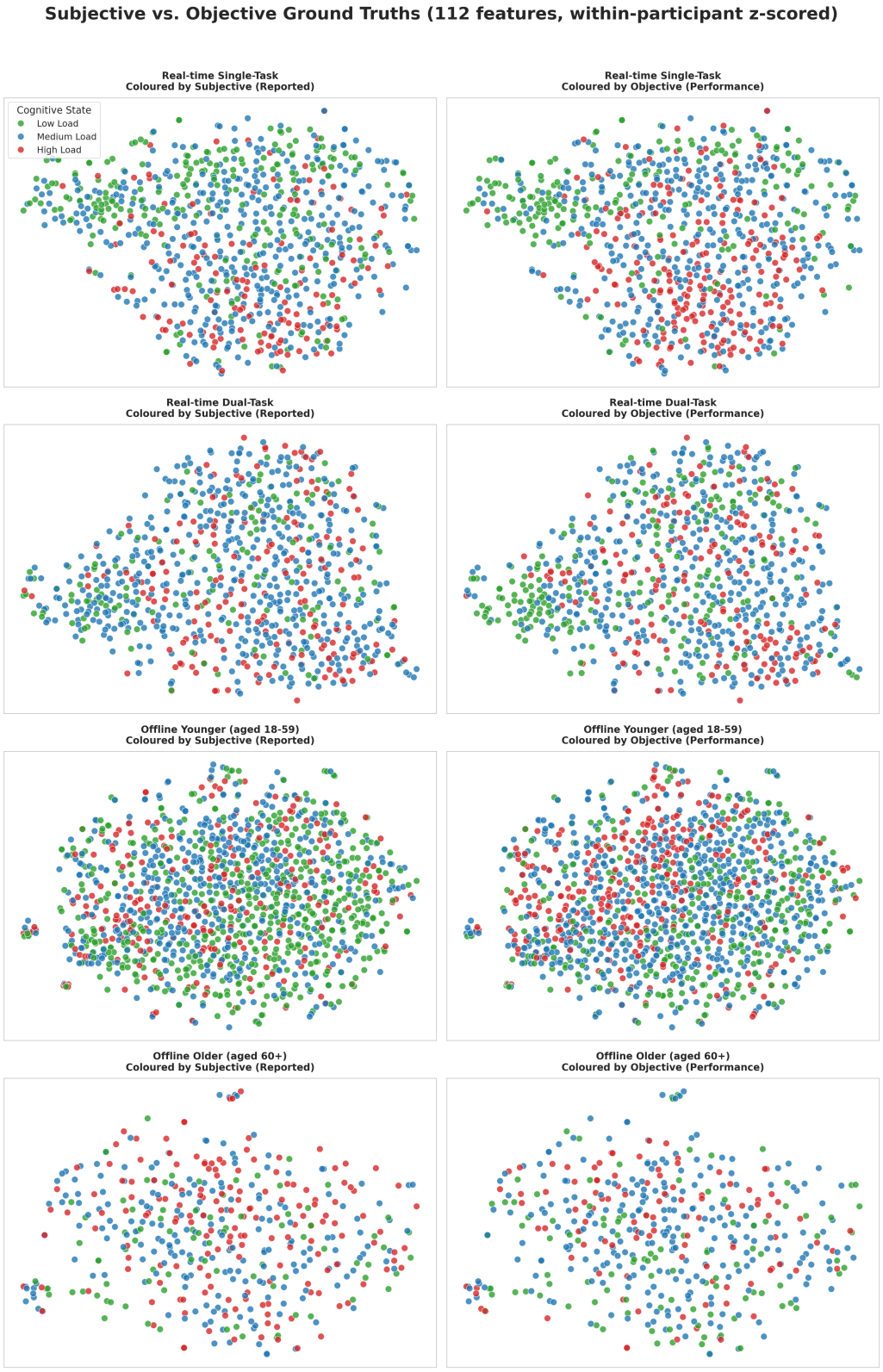


Figure 7.6: t-SNE embeddings of the 112 level-level features (28 signals × 4 statistics), one embedding per cohort (rows). The coordinates in the left and right columns are identical. The left column is coloured by the participant's subjective rating (Too Easy, Just Right and Too Difficult shown as Low, Medium and High Load) and the right column by the objective label derived from difficulty-weighted task performance.

levels toward the bottom but with more overlap. In the real-time dual-task cohort (second row) it is harder to find clear separation between the classes except for the seemingly condensed cluster of Low Load ratings in the bottom left of the graph coloured based on objective labels.

In the offline younger cohort (third row) the performance-based labels show a seemingly better separation of Low and High Load than subjective labels. In the offline older cohort (bottom row) neither labelling shows visible structure, and the projection is sparser because there are fewer levels. Medium Load levels are dispersed across every cohort and do not form a region of their own.

Structure was therefore visible mainly in the real-time cohorts, and there more clearly for the objective labels than for the subjective ratings.

Interpretation

In the real-time cohorts, the features are arranged somewhat more clearly with respect to the objective labels than the subjective ratings, most visibly in the single-task real-time and offline younger cohorts, and there is little structure in the offline cohorts under either labelling. This is consistent with the direction of the predictive results in Analysis 2 but does not establish it. t-SNE is a visual tool. The embeddings show that the features vary with the labels but do not provide a reason for that. The difference between the two labellings in the real-time cohorts is modest at best.

A possible reason for the weaker arrangement of the subjective ratings is that they are judged relative to the range of difficulty experienced (Chapter 2), which would make them a less stable reference when the steps between difficulty levels are small. The embeddings cannot test this account.

7.4 Discussion

This chapter sought to investigate the perceptual and computational differences between subjective experiences of cognitive load and objective task performance, particularly as training environments transition from static, high-contrast paradigms to dynamic, low-contrast adaptive systems. The findings suggest that participants' difficulty ratings and their objective performance diverged in ambiguous, low-contrast environments. They also suggest that eye-tracking and physiological data predicted objective performance more reliably than they predicted those ratings.

Taken together, the four analyses appear to point the same way from four directions. Ratings and performance seem to agree only fairly (Analysis 1). The same features appear to predict performance-derived labels more accurately than ratings (Analysis 2). The difference does not seem to be explained by too few participants (Analysis 3). In the real-time single-task and offline younger cohorts, the physiological data appear to be arranged somewhat more clearly when coloured by performance than by rating, while the remaining cohorts show little structure under either labelling (Analysis 4). These patterns suggest that ratings are a less stable reference than performance when the steps between difficulty levels are small. Two features of the design bear on this reading: the objective target is defined relative to each participant's own performance, and the low-contrast cohorts differ from the offline cohort in more than contrast (different participants, a dual-task condition, adaptive difficulty).

7.5 Summary

This chapter tested RQ7: whether subjective workload assessments remain a reliable ground truth once a system moves from the discrete, high-contrast conditions used to train a model to the continuous, low-contrast conditions of real-time deployment. The evidence is consistent with H7. As the task environment shifted from Chapters 3 and 4's

discrete conditions to Chapter 6's continuous ones, the correspondence between subjective difficulty judgments and objective, physiologically-indexed performance weakened. The same physiological and eye-tracking features also predicted objective performance more reliably than they predicted subjective judgment across every cohort tested. This is consistent with the explanation Chapter 6 offered for its own real-time accuracy drop. Part of the drop may reflect the instability of the subjective label under low-contrast conditions, described here through metacognitive sensitivity, perceptual anchoring and concept shift, rather than a comparable decline in the physiological signal itself.

The findings of this chapter contribute to the design of closed-loop, adaptive cognitive training systems. If a system relies on prompting users to rate their own difficulty, it will adapt to their psychological biases and anchoring effects as well as to their actual neurocognitive limits, which may limit training efficacy in low-contrast, real-time conditions specifically. Training physiologically-driven models against objective performance rather than subjective label, where feasible, is one way to reduce this risk. The Practical Limitations of Performance-Driven Ground Truths discussed earlier in this chapter also noted that objective performance is itself an imperfect target, since it can mask the hidden cost of compensatory effort. The learning-curve analysis further suggests that, for a given task and population, additional data collection past roughly 30-40 participants brought diminishing returns, indicating that a bespoke, moderately-sized model for a given task or population is a feasible approach within this thesis's scope.

Chapter 8 draws together the findings of Chapters 3 through 7 against the seven research questions and thesis statement set out in Chapter 1. It also considers what they jointly imply for the design of physiologically-driven adaptive cognitive training systems.

Chapter 8

Discussion and Limitations

8.1 Introduction

Chapter 1 set out a thesis statement centred on using Virtual Reality to estimate and adapt to cognitive load in real time. It posed seven research questions. RQ1 asked whether eye-tracking and physiological fusion can proxy cognitive load in a single-task VR setting. RQ2 and RQ3 asked whether that model extends to older adults and what that transfer requires. RQ4 asked whether a model trained offline generalises to real-time, dual-task deployment. RQ5 asked whether model-driven control can replace participant-driven control. RQ6 asked whether real-time adjustment changes difficulty judgments relative to a fixed design. RQ7 asked whether subjective workload assessment remains trustworthy under real-time deployment. Chapters 3 through 7 addressed these in turn. Section 8.2 first summarises the outcome of each hypothesis, and the sections that follow discuss the benefits and drawbacks of each finding. This chapter discusses each major finding together with its constraints, rather than discussing what worked first and qualifying it later with a list of limitations. The two are rarely separable. The age-specific markers identified in Chapter 4 are only as trustworthy as the sample they were drawn from. The outcome equivalence reported in Chapter 6 is the measure of success for the model-driven deployment, while the per-trial accuracy of the classifier that produced it is a secondary

indicator. For each finding, this chapter considers what it contributes and what constrains it. It also considers what the finding implies for the research field and for the people such a system would ultimately be built for. Finally, it considers what stands between this research and real-world adoption.

8.2 Outcome of the Hypotheses

The hypotheses set out in Chapter 1 test the central hypothesis of this thesis: that eye-tracking and peripheral physiological signals can estimate how difficult a task feels well enough for a model to adjust difficulty in real time and to produce outcomes comparable to participant-driven adjustment. Table 8.1 summarises what each chapter found. A hypothesis is considered supported when the result agrees with it and no part of it is contradicted. It is considered partly supported when one component is met and another is not addressed by the data. The verdicts describe what the evidence suggests: several comparisons are between cohorts, the outcome comparisons rest on differences that appear small, and the deployment results come from a single sitting per participant.

Read together, the estimation hypotheses (H1 to H3) were supported in their core claim: cognitive load could be estimated above chance in younger and older adults, with different features carrying the signal in each. The deployment hypotheses (H4 to H6) were supported within a single session with younger adults, and H7 was supported in direction but weakly. The central hypothesis is therefore supported in the scope tested: a model driven by eye-tracking and peripheral physiology produced in-session outcomes comparable to those of participant-driven control, even though its per-trial agreement with participants' ratings was moderate offline and low in real time. The evidence does not cover older adults in real time, repeated sessions, or cognitive domains other than sustained attention.

Table 8.1: Outcome of each hypothesis, with the chapter that tested it.

H	Ch.	Key result	Verdict
H1	3	The combined model was above chance under leave-one-participant-out validation (accuracy 0.56, AUC 0.72, against about 0.33 chance accuracy). Eye-tracking carried the signal (F1 0.48); physiological features alone were close to chance (F1 0.34).	Supported, carried by eye-tracking
H2	4	The older-adult model retained moderate predictive validity (accuracy 0.49, AUC 0.66, $N = 19$), lower than in younger adults.	Partly supported
H3	4	The most informative features differed by age (fixation stability in younger adults, saccade amplitude and acceleration in older adults), but the older-adult signal was also weaker (AUC 0.66 against 0.72).	Partly supported
H4	6	Real-time AUC was slightly lower in dual-task than in single-task conditions (0.599 against 0.618) and remained above chance. The difference appears small.	Supported
H5	6	No significant differences between model-driven and participant-driven control in weighted game performance, arithmetic errors or NASA-TLX, including frustration. Raw performance was lower under model-driven control in dual-task conditions. Outcomes appear comparable overall.	Supported
H6	6	Ratings shifted toward *Just Right* (57.1% of real-time ratings against 39.1% of offline ratings). The comparison is between cohorts.	Supported for the shift
H7	7	Agreement between subjective and objective labels was fair in every cohort and slightly lower in real time (κ 0.244 against 0.290). The F1 gap between the two targets appears to widen from about 6 to about 16 points.	Consistent with H7; the effect appears small

8.3 Real-Time Cognitive Load Estimation from Eye-Tracking and Physiology (RQ1)

Chapter 3 established that eye-tracking and physiological fusion can predict subjective cognitive load above chance in a single-task VR paradigm under Leave-One-Participant-Out validation. Oculometric features, particularly fixation duration variability and saccadic dynamics, carried most of the predictive signal. Physiological features contributed comparatively little once genuine cross-participant generalisation was required. Chapter 5 then confirmed, via a purpose-built raw-signal baseline, that this was not an artefact of using simple hand-engineered features when a more expressive model could have learned better representations on its own. A sequence model given only minimally processed signal performed poorly (macro-F1 about 0.42, against about 0.50 for the windowed model), and the full attention-based architecture could not close that gap either. Taken together, these results contribute to methodology as well as to the empirical record. Avoiding pooled-subject cross-validation and task-stacking, as Chapter 2 argued the field as a whole should do, produces a model with an offline accuracy of 52.3% on average (three classes). Pooled validation or task-stacking might have produced higher figures. In Chapter 3, macro-F1 for the combined model fell from 0.57 under pooled validation to 0.49 under leave-one-participant-out validation, a reasonably substantial gap that is in line with the literature. The raw-signal baseline result is also a caution against assuming that architectural sophistication can substitute for feature engineering.

The corresponding drawback is that this accuracy varies considerably across participants (fold accuracy ranged from 16.7% to 83.3% in Chapter 5), although most errors were between *Just Right* and an edge case, which the small-step adaptation rule tolerates. This is a genuine constraint on how confidently this thesis's central sensing pipeline can be deployed for a new, unseen user without some form of calibration, and it compounds through every later chapter that depends on it. For the research field, the implication is a trade-off worth naming explicitly. Participant-independent validation produces a less flattering but more trustworthy headline number, and readers of this literature more

broadly should treat pooled-validation accuracy figures with corresponding scepticism. For end users, the practical implication is that a system built on this pipeline is more likely to serve some users well and others poorly than to perform uniformly. Section 8.5 below returns to what that means for uptake. A further, more mundane barrier is hardware. This thesis relied on a consumer-grade, headset-integrated eye-tracker sampling at 90 Hz, and Chapter 3 discussed at length why this suits the window-level aggregates used here but limits the resolution of individual saccades. More capable consumer hardware would likely improve on the figures reported here without requiring a different modelling approach.

A further consideration concerns the deployed model itself. The real-time system used a bidirectional LSTM with self-attention which, on the same 73 participants, reached a higher macro-F1 than the Mixed XGBoost model of Chapter 4 (0.497 against 0.466), with slightly lower accuracy and AUC (Section 5.3.3). The margin was slightly larger when the real-time experiment had to be run, because the better XGBoost models were trained later, with improved feature selection. The LSTM also discriminated the *Too Easy* and *Too Difficult* classes well (one-versus-rest AUC of 0.78 and 0.74), which is the property the adaptation rule relies on most. The network has 56,708 parameters, which is large relative to the 1,736 training sequences, although dropout, weight decay, early stopping and participant-independent evaluation limit its effective capacity. A simpler model might offer a comparable trade-off at lower cost, and the size and complexity of the deployed model are acknowledged as a shortcoming for sustainability and reproducibility. For the field, a reasonable default is to start from the simplest model that meets the requirement and to add complexity where it earns its cost. This is the principle followed here to arrive at the final deployed model. Admittedly, the jump in size did not justify the jump in performance, despite it being our best performing model with best AUC scores for the two edge classes of particular interest (*Too Easy, Too Difficult*).

8.4 Extending the Model to Older Adults (RQ2, RQ3)

Chapter 4 found that eye-tracking retained meaningful predictive validity for cognitive load in an older cohort (AUC = 0.66 under LOPO validation), but only once the model was allowed to weight different features than it does for younger adults. Younger adults' load is best captured by fixation duration variability, whereas older adults' load is better captured by oculomotor mechanics such as saccade amplitude. The chapter related this pattern to the Compensation-Related Utilisation of Neural Circuits Hypothesis and to age-related contraction of the Useful Field of View. This is a substantive contribution in its own right, since it suggests that age-related differences in this domain are not simply noise to be averaged away. It also suggests that a Mixed Population model may provide a workable route to a single, more inclusive system, without requiring the field to build and maintain fully age-segregated models.

The corresponding drawback, discussed at length in Chapter 4, is that this finding rests on a small (N=19), non-representative older-adult sample. The sample was recruited through a University of the Third Age, a channel independently documented to skew toward more educated, more socially engaged and disproportionately female members, and was screened for the absence of Mild Cognitive Impairment. The sample is informative about relatively healthy, socially engaged, technology-willing older adults specifically. The true extent of the younger/older divergence in a broader, more representative older population remains an open question. Such a population includes individuals with emerging cognitive impairment, greater frailty, or less willingness to engage with VR hardware at all. Compounding this, the real-time system evaluated in Chapters 5 and 6 was never deployed with an older cohort. The offline age-specific markers identified in Chapter 4 are therefore yet to be examined under the real-time, closed-loop, dual-task conditions where Chapter 6 showed the younger-adult model itself lost accuracy.

For the research field, this argues for treating age-diverse cohorts as a standard inclusion in physiological computing research rather than a follow-up study. It also argues for reporting recruitment channel and any associated selection effects as routinely as demographic summary statistics. For end users, the implication cuts both ways. This thesis provides evidence that VR-based cognitive training can in principle be made more inclusive of older adults, who are among the intended beneficiaries of this technology. However, deploying the current system as though this question were settled for older adults generally, rather than for the specific, healthier-than-average slice sampled here, would be premature and could produce a false sense of confidence. The clearest barrier to uptake is therefore empirical. Real-world older-adult populations are more heterogeneous, physically and cognitively, than the sample studied here. The motoric scaffolding this thesis's findings imply is needed (larger UI targets, reduced saccadic demand, accommodations for a contracted field of view) has been proposed but not built or tested in a live system. Chapter 9 returns to this as a concrete direction for future work.

8.5 Real-Time Deployment: Model- versus Participant-Driven Control (RQ4, RQ5, RQ6)

The clearest positive result in this thesis is reported in Chapter 6. Model-driven and participant-driven control produced statistically equivalent outcomes on weighted game performance, subjective workload and learning rate. In dual-task conditions, the model-driven system also sustained participants at modestly harder difficulty levels than they chose for themselves, without a measurable performance cost. This matters because it is measured against the outcomes a training system actually exists to produce, rather than against the model's agreement with a self-report label that Chapter 7 went on to show is itself an unstable target under real-time conditions. On this evidence, it is possible to remove the explicit, immersion-breaking step of asking users to rate their own difficulty without a measurable cost to the outcomes measured here, at least over the

single-session lengths tested. The dual-task result suggests that physiologically-driven adaptation may be able to do something a purely self-regulated design does not: move users past a conservative self-assessment without a performance penalty. This was a small and borderline effect.

The corresponding drawback is that this outcome-level success was achieved by a classifier whose real-time, per-trial accuracy against participants' own difficulty ratings was 36.7%, against a 33.3% chance baseline for three classes and below the 52.3% pre-deployment figure. Most errors were between *Just Right* and an edge case, which the small-step adaptation rule tolerates, whereas errors between the two edge cases (10.4% of predictions) are the ones that push difficulty in the wrong direction. Chapters 6 and 7 together make a reasonable case that a substantial part of this drop reflects the instability of the subjective label under low-contrast, real-time conditions rather than a comparable decline in the underlying physiological signal. This thesis cannot fully separate the two, and some genuine real-time signal degradation likely remains. Individual variability compounds this. Per-participant accuracy in Chapter 6 ranged from 6.2% to 59.4%, meaning the system that produced equivalent outcomes on average is likely to have performed considerably better for some users than others. This thesis's single-session design cannot speak to whether the observed equivalence would hold over the longer, multi-session training programmes that cognitive training is ultimately meant to support.

For the research field, this is evidence that outcome-level equivalence, rather than per-trial classification accuracy, may be the more meaningful benchmark for adaptive training systems, since a system can be practically useful without matching the offline accuracy of the model it is built on. For end users, a system of this kind could mean a less obtrusive experience, with no interruption to ask how a session feels. For at least some users, it could also be better calibrated to their actual capacity than self-regulation would be. For the subset of users for whom the model's alignment with their physiology is poor, however, the same system could produce a worse experience without any indication to the user that this is happening. The barriers to uptake follow directly from this. User trust in a system that occasionally overrides how a session feels, particularly without transparency about

when and why, is not something this thesis's short, single-session, hidden-loop design was positioned to test. A real deployment would need a calibration phase, a way to communicate model confidence to the user, or both. The practical hardware and software pipeline used here is a dedicated eye-tracker-equipped headset feeding a local inference service with sub-500ms latency. This is a meaningfully higher bar than the fixed-difficulty designs this thesis argues against. Its cost and technical complexity are a real barrier to the home-based deployment this research is ultimately motivated by.

8.6 Subjective and Objective References and the Choice of Ground Truth (RQ7)

Chapter 7 tested, rather than assumed, the explanation Chapter 6 offered for its own accuracy shortfall. It found that the same physiological and eye-tracking features predicted objective, difficulty-weighted performance more reliably than they predicted participants' own subjective ratings. The gap between the two appeared to widen specifically in the low-contrast, real-time conditions where the accuracy drop was observed. This is a useful diagnostic result for this thesis and a design principle for the field. It identifies subjective self-report as a specific point of fragility in real-time adaptive systems, grounded in metacognitive sensitivity, perceptual anchoring and machine learning concept shift. It also offers an account of the offline-to-real-time accuracy drop, which would otherwise remain an unexplained fact about deployment in general.

The drawback, which Chapter 7 itself flagged before this thesis had reason to overrely on the finding, is that objective performance is not a clean substitute ground truth either. It can mask the hidden cost of compensatory effort in exactly the way CRUNCH describes for older adults in Chapter 4. It also risks a ceiling effect, in which a system reads high performance as evidence of comfortable capacity when it may instead reflect a user working at their limit. The *Just Right* or flow state, the state the system is built to sustain, was also the hardest class to identify regardless of which ground truth was used.

In Chapter 7's t-SNE analysis it does not form a distinct region but is spread through the embedding between and around the two extremes. This matters less for a small-step, closed-loop controller than for a one-shot classifier, since misclassifying a *Just Right* level as an edge case moves difficulty by one small step that later estimates can offset.

For the research field, the implication is that the choice of ground truth in cognitive load estimation actively determines what a model can be shown to do. This thesis argues that the choice should be reported and justified as deliberately as any other methodological decision, and that subjective report should not be the default simply because it is the most direct available measure of the construct of interest. For end users, a system designed to weight objective performance over self-report is, in effect, a system built to override how a person feels about their own effort when the two disagree. This raises a real question of user autonomy and consent. This thesis's short, hidden-loop laboratory design did not need to resolve that question, but any live deployment would. This is also the clearest ethical barrier to uptake identified in this thesis. Acceptance of a system that can act against a user's stated preference is a materially bigger ask outside a supervised research setting. It would warrant closer scrutiny in any deployment involving vulnerable users, including the older adults and, potentially, clinical populations this research is ultimately aimed at, echoing the Human-in-the-Loop discussion in Chapter 2.

8.7 Implications for Design Practice

Drawing the discussion above together, several design implications follow directly from specific findings in this thesis rather than from general principle.

1. **Sensorimotor accommodation over cognitive simplification for older users.** Following the age-specific marker shift discussed in Section 8.4, adaptation for older users should prioritise motoric scaffolding, such as centralising stimuli to accommodate a reduced Useful Field of View and slowing target movement to align with saccadic velocity caps, rather than assuming the same cognitive-difficulty levers used for younger adults will transfer unchanged.

2. **Temporal context matters for real-time architectures.** The architectural choices in Chapter 5, a windowed, sequence-aware model, were made so that estimates could be computed from a live data stream, and this thesis does not show an accuracy advantage over session-level aggregates; a traditional model fitted to session-level aggregates, of the kind used in Chapters 3 and 4, remains a viable and simpler alternative where real-time adaptation is not required.

3. **Objective-led training, held to appropriate ethical account.** Following Section 8.6, systems aiming to maximise training efficacy have some merit to weight objective performance over a user's momentary self-assessment, but that section also set out why this needs to be an explicit, accountable design decision rather than a default.

4. **Favour frequent, small-step adjustment over large-amplitude, discrete designs.** The task-stacking and large-difficulty-jump designs common in the wider literature, critiqued throughout this thesis, are in tension with the finding that low-contrast, incremental adjustment both stabilises the user experience (Chapter 6) and is where subjective judgment becomes least reliable (Chapter 7); designers should treat this trade-off as a known cost of incremental adaptive design rather than an unexpected side effect.

5. **Design and validate for the population actually being served.** This thesis's own older-adult sample (Section 8.4) is a caution here as much as a recommendation: whatever population a system is intended for should be represented, and its recruitment channel reported, in the data used to build and validate it, rather than assuming a model trained on a convenience sample will generalise to the deployment population.

Chapter 9

Conclusion and Future Work

9.1 Thesis Summary

Chapter 1 set out to use Virtual Reality to estimate and adapt to cognitive load in real time. It argued that adaptive cognitive training requires a way to measure load while a task is still under way. It also argued that neither subjective self-report nor any single physiological modality is an uncontested ground truth for doing so. Read individually, Chapters 3 through 7 each answer one part of this problem; read together, they form a single line of argument rather than four separate studies (counting the real-time model development and deployment of Chapters 5 and 6 as a single study).

Chapter 1 framed the thesis around one central hypothesis: that eye-tracking and peripheral physiological signals can estimate how difficult a task feels well enough for a model to set difficulty in real time, with outcomes comparable to participant-driven control. Within the scope tested, which for the real-time part is a single sitting with younger adults, this hypothesis appears to be supported. Table 8.1 in Chapter 8 gives the outcome of each of the seven supporting hypotheses, with H2, H3 and H7 only partly or weakly supported.

Chapter 3 established that eye-tracking and physiological fusion can estimate cognitive load in a single-task VR paradigm without relying on the task-stacking and pooled-subject validation this thesis identified as central weaknesses in the field, with an offline accuracy obtained under participant-independent validation. Chapter 4 tested whether that model transfers to older adults, one of the populations adaptive cognitive training is most clearly intended to serve. It found that cognitive load is estimable in older adults but that the physiological markers carrying the signal differ with age, so the model may not transfer unchanged. This finding would have been less visible had the thesis treated age as noise to be averaged over rather than a variable to be modelled. Chapter 5 then confronted a different problem: a model capable of offline analysis is not automatically capable of real-time use. It built and validated an architecture designed for exactly this. A raw-signal baseline, which the field rarely tests for, also showed that this thesis's feature engineering was not made redundant by a more expressive model. Chapter 6 deployed that model live. Its clearest result was the equivalence between model-driven and participant-driven outcomes, rather than the classifier's per-trial accuracy, which was lower than its pre-deployment figure. There was also a small behavioural effect: the model-driven condition reached harder final difficulty levels under dual-task load without a performance cost. Chapter 7 then returned to explain that accuracy shortfall rather than leaving it as an unresolved loose end. It suggested that the same physiological signals track objective performance more reliably than subjective judgment specifically once the task becomes continuous and low-contrast. This is consistent with the modelling choices made earlier in this thesis and prompts a rethink of what "accuracy" should mean for a system judged against a moving, self-reported target.

Taken as a whole, this thesis supports a qualified version of its opening thesis statement. A real-time, physiologically-driven adaptive VR system is achievable using consumer-grade, integrated sensing, and, on the outcome measures that matter for a training system's purpose, performs comparably to a person regulating their own difficulty. This achievement is qualified in three specific ways. The first is the per-trial accuracy of the underlying classifier, which was moderate offline, lower in real time, and variable between individuals (Chapters 5 and 6). The second is the narrow, non-representative slice of the

older-adult population in which age-inclusive claims have so far been tested (Chapter 4). The third is that every result in this thesis comes from a single session, so nothing here speaks directly to whether these effects would hold, strengthen or fade over a genuine multi-session training programme. Chapter 8 discussed each of these qualifications alongside the finding it attaches to; this chapter turns from summarising what was found to setting out what should be done next.

9.2 Contributions

This thesis makes the following contributions, spanning methodology, empirical findings, and engineering.

- **A methodological approach**: avoiding task-stacking and pooled-subject cross-validation throughout, and reporting the resulting accuracy figures under participant-independent validation (Chapters 3, 5).
- **A direct empirical test of feature engineering versus architectural sophistication**: a raw-signal baseline showing that a sequence model given only minimally processed sensor output, including the full attention-based architecture used for deployment, could not match a model built on hand-engineered temporal features (Chapter 5).
- **Age-specific markers of cognitive load**: evidence that the physiological and oculomotor features most predictive of load differ systematically between younger and older adults, with a Mixed Population model offered as a workable route to a single, more inclusive system (Chapter 4).
- **An empirical test of real-time, closed-loop deployment against a participant-driven baseline**, rather than an offline benchmark alone, showing outcome equivalence and, under dual-task load, harder final difficulty levels without a performance cost (Chapter 6).

- **Direct evidence on the choice of ground truth for adaptive systems**: evidence that physiological signals track objective performance more reliably than subjective judgment specifically under the real-time, low-contrast conditions where adaptive systems actually operate (Chapter 7).

- **An open dataset and codebase**: the raw physiological and eye-tracking data collected across this thesis's experiments, together with the software used to run them offline and in real time, were designed and built by the author and are made publicly available so that other researchers can build on, replicate, or challenge the findings reported here without re-collecting the underlying data. The dataset is deposited on the Open Science Framework (Szczepaniak, Harvey, & Deligianni, 2026; https://doi.org/10.17605/OSF.IO/VJSNF), and the experimental software is available on GitHub (https://github.com/EyeDAPT).

9.3 Future Work

The limitations discussed in Chapter 8 point to specific, actionable directions for future work, rather than general calls for "more research."

Per-user calibration for individually variable models. Chapter 6 found per-participant real-time accuracy ranging from 6.2% to 59.4%, and Chapter 8 argued this is a genuine barrier to uptake rather than a footnote. Transfer learning, allowing a short, individual warm-up calibration to adjust a pre-trained global model's decision boundaries to a new user's baseline physiology, is a natural next step before any system built on this work is deployed to a general user base.

Real-time validation with older adults. Chapter 4 identified age-specific load markers offline, but the real-time system in Chapters 5 and 6 was tested exclusively with younger adults. Extending real-time deployment to an older cohort is necessary before this thesis's inclusivity claims can be extended to real-time use. The adaptation logic should incorporate the motoric scaffolding implied by Chapter 4's findings, such as adjusting target size or spatial density to accommodate a contracted Useful Field of View, rather than only modulating task speed.

Broader, more representative recruitment. Chapter 4's older-adult sample was recruited through a University of the Third Age and is not representative of the general older population. Future work should recruit through a wider range of channels. It should also test whether the age-specific markers reported here hold in a more demographically varied older-adult sample that is less uniformly healthy and socially engaged.

Generalisability across cognitive domains. This thesis focused on sustained attention, with a mental arithmetic dual task. Whether the oculometric markers identified here transfer to other forms of cognitive work, such as working memory, task-switching, or cognitive flexibility tasks, is untested, and a multi-task paradigm would establish whether a single model can serve multiple training domains or whether a more modular, task-specific approach is required.

Simpler, more sustainable models. The deployed BiLSTM reached a slightly higher macro-F1 than XGBoost on the same data (Chapter 5) but is large relative to the data. Future real-time deployments could test whether a simpler model, such as XGBoost with the improved feature set, offers a comparable trade-off at lower cost. A repeat of the real-time experiment with such a model would show whether the outcome equivalence reported here depends on the model.

Continuous rather than categorical load estimation. The three-class structure used throughout this thesis fared worst at the *Just Right* state specifically (Chapter 7), which behaves as a transitional gradient rather than a distinct class. A regression-based, continuous estimate of cognitive load could in principle provide smoother difficulty transitions and avoid the boundary oscillation a discrete classifier is prone to, at the cost of a more complex evaluation and control framework than the one used here.

Multi-session and longitudinal validation. Every result in this thesis comes from a single experimental session. Whether the outcome equivalence and dual-task advantage reported in Chapter 6 persist, strengthen or fade over a genuine multi-session cognitive training programme is a necessary test before any efficacy claims beyond a single session are warranted. The same programme would show whether real training gains, rather than within-session learning effects, differ between model-driven and participant-driven designs.

Hardware evolution. This thesis relied on consumer-grade, headset-integrated sensing, sampling eye-tracking at 90 Hz and physiological signals at 51 Hz. As Chapter 3 discussed, this suits the window-level aggregates this thesis relies on but not finer-grained markers such as the peak velocity of individual saccades or microsaccades. Future consumer hardware with higher sampling rates and additional sensing modalities (e.g. EEG integrated into a headset at a practical weight and cost) may extend what a real-time system of this kind can detect, without requiring a different modelling approach.

Testing beyond healthy adults. Chapter 1 explicitly scoped this thesis to healthy adults across the lifespan, leaving clinical populations, such as individuals with Mild Cognitive Impairment or dementia, for future work. Adaptive cognitive training is frequently proposed for exactly these groups. The ethical considerations raised in Chapter 8 around a system overriding a user's subjective comfort also apply. This population would therefore require its own dedicated validation, very likely under a Human-in-the-Loop design of the kind discussed in Chapter 2, rather than a direct extension of the system evaluated here.

9.4 Concluding Remarks

This thesis has indicated that a real-time, physiologically-driven adaptive VR system, built from consumer-grade eye-tracking and peripheral physiology alone, can produce in-session outcomes equivalent to a person managing their own difficulty. It may also do so while avoiding some of the self-assessment biases that can affect purely self-regulated designs. It has also shown, at each stage, where that claim needs qualifying. The qualifications are the per-trial accuracy of the classifier and its variability between individuals, a narrow and non-representative older-adult sample, and the single-session scope of every experiment reported here. This thesis names these qualifications precisely as part of its contribution, on the view that a research field advances more reliably by being honest about the size and boundaries of an effect than by describing every result as a validated solution. On that basis, the evidence gathered here supports cautious optimism that physiologically-driven adaptation has a role to play in cognitive training. It also supports a concrete, evidenced list of what would need to be true before that role extends much beyond the specific participants, single sessions and sustained attention task studied in this thesis.

Chapter 10

Supplementary Material

10.1 NASA-TLX Descriptive Statistics

This section reports NASA-TLX descriptive statistics as supplementary detail. It is provided for completeness and reference only; the formal group comparisons and significance tests based on this data are reported in the relevant results chapters and are not repeated here.

Subjective workload was assessed via the NASA Task Load Index (NASA-TLX) in both the offline and real-time experiments. Both experiments captured the full six-subscale NASA-TLX (Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, Frustration) alongside its composite total score.

10.1.1 Offline Experiment

NASA-TLX was administered once per session across three sessions per participant ($N = 78$ participants, 234 session-level ratings). Participants were grouped into younger (aged 18–58, $n = 58$) and older (aged 60–87, $n = 20$) cohorts, following the age-grouping convention used throughout this thesis. Table 10.1 reports descriptive statistics overall and by age group.

Table 10.1: NASA-TLX descriptive statistics, offline experiment (mean (SD); median in brackets)

Subscale	Overall ($N = 234$)	Younger ($n = 174$)	Older ($n = 60$)
TLX Total	54.71 (19.58) [57.0]	52.64 (20.13) [53.0]	60.68 (16.64) [61.5]
Mental Demand	10.47 (4.70) [11.0]	10.28 (4.56) [11.0]	11.05 (5.08) [11.5]
Physical Demand	7.39 (4.69) [6.0]	7.44 (4.66) [6.5]	7.23 (4.84) [5.0]
Temporal Demand	10.29 (4.06) [11.0]	9.92 (4.13) [11.0]	11.37 (3.68) [11.0]
Performance	8.59 (4.83) [8.0]	8.15 (4.94) [7.0]	9.87 (4.28) [10.0]
Effort	11.11 (4.50) [12.0]	10.71 (4.34) [12.0]	12.25 (4.80) [13.5]
Frustration	6.85 (4.62) [6.0]	6.14 (4.23) [5.0]	8.92 (5.09) [10.0]

Formal group comparisons for TLX Total, Mental Demand, Physical Demand, and Frustration are reported in Chapter 4 (Section 4.3.2); Temporal Demand, Performance, and Effort were not part of that test battery and are presented here descriptively only.

10.1.2 Real-Time Experiment

NASA-TLX was administered once per experimental condition ($N = 54$ participants, mean age 28.5 years, SD 8.4, range 18–59; 30 female, 22 male, 2 non-binary), yielding 216 condition-level ratings across four conditions in a 2×2 design crossing difficulty-adaptation strategy (Model-driven vs. Self-driven) with the presence of the concurrent mental arithmetic dual-task (dual vs. single task): MA (Model, dual-task), MNA (Model, single-task), SA (Self, dual-task), SNA (Self, single-task). Table 10.2 reports overall descriptive statistics; Table 10.3 breaks these down by condition.

Table 10.2: NASA-TLX descriptive statistics, real-time experiment, overall ($N = 216$)

Subscale	Mean	SD	Min	Median	Max
TLX Total	59.78	20.06	8	62	110
Mental Demand	12.28	4.49	1	13	21
Physical Demand	8.58	4.89	1	8	20
Temporal Demand	11.38	4.37	2	12	20
Performance	6.44	3.62	1	6	18
Effort	12.93	4.66	2	14	21
Frustration	8.16	5.04	1	8	21

Table 10.3: NASA-TLX Total and Mental Demand by condition, real-time experiment (mean (SD), $n = 54$ per condition)

	MA	MNA	SA	SNA
TLX Total	67.28 (15.34)	55.81 (23.25)	63.57 (19.09)	52.46 (18.70)
Mental Demand	14.74 (3.22)	10.69 (4.62)	13.87 (4.23)	9.83 (3.83)
Effort	14.56 (3.57)	11.63 (5.17)	13.81 (4.20)	11.72 (4.94)
Frustration	9.46 (5.11)	7.15 (5.28)	9.15 (4.92)	6.87 (4.40)

Formal comparisons between conditions, including the dual-task manipulation check and the Model-driven versus Self-paced comparison, are reported in Chapter 6.

Bibliography

[1] Diane E Adamo, Brian J Martin and Susan H Brown. 'Age-related differences in upper limb joint position sense and mental workload'. In: *Experimental Brain Research* 216 (2012), pp. 571–580.

[2] Muneeb Imtiaz Ahmad et al. 'A framework to estimate cognitive load using physiological data'. In: *Personal and Ubiquitous Computing* 27 (2023), pp. 2027–2041. doi: 10.1007/s00779-020-01455-7.

[3] Ju Ah Ahn and Sung Yul Chung. 'Detection of Cognitive Workload Using Heart Rate Variability'. In: *Intelligent Systems in Science and Engineering* 23.1 (2018), pp. 55–62.

[4] N. Ali et al. 'The Effects of Dual-Task Training on Cognitive and Physical Functions in Older Adults with Cognitive Impairment: A Systematic Review and Meta-Analysis'. In: *The Journal of Prevention of Alzheimer's Disease* 9.3 (2022), pp. 359–370. doi: 10.14283/jpad.2022.16.

[5] Richard Andersson, Marcus Nyström and Kenneth Holmqvist. 'Sampling frequency and eye-tracking measures: how speed affects durations, latencies, and more'. In: *Journal of Eye Movement Research* 3.3 (2010), pp. 1–12. doi: 10.16910/jemr.3.3.6.

[6] Pavlo Antonenko et al. 'Using Electroencephalography to Measure Cognitive Load'. In: *Educational Psychology Review* 22.4 (2010), pp. 425–438. doi: 10.1007/s10648-010-9130-y.

[7] Lora Appel et al. 'Older adults with cognitive and/or physical impairments can benefit from immersive virtual reality experiences: A feasibility study'. In: *Frontiers in medicine* 6 (2020), p. 329.

[8] R. V. Aranha, C. G. Correa and F. L. S. Nunes. 'Adapting Software with Affective Computing: A Systematic Review'. In: *IEEE Transactions on Affective Computing* 12.4 (2021), pp. 883–899.

[9] Vladimir Araujo, Alejandra Gonzalez and Diego Mendez. 'Dynamic Difficulty Adjustment for a Memory Game'. In: *Technology Trends*. Ed. by Miguel Botto-Tobar et al. Cham: Springer International Publishing, 2019, pp. 605–616.

[10] U. Asgher et al. 'Enhanced Accuracy for Multiclass Mental Workload Detection Using Long Short-Term Memory for Brain–Computer Interface'. In: *Frontiers in Neuroscience* 14 (2020), p. 584. doi: 10.3389/fnins.2020.00584.

[11] Gary Aston-Jones and Jonathan D. Cohen. 'An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance'. In: *Annual Review of Neuroscience* 28 (2005), pp. 403–450.

[12] A. Aygun et al. 'Cognitive Workload Assessment via Eye Gaze and EEG in an Interactive Multi-Modal Driving Task'. In: *Proceedings of the 2022 International Conference on Multimodal Interaction* (2022), pp. 337–345.

[13] Paul Ayres et al. 'The Validity of Physiological Measures to Identify Differences in Intrinsic Cognitive Load'. In: *Frontiers in Psychology* 12 (2021), p. 702538. doi: 10.3389/fpsyg.2021.702538. url: https://pubmed.ncbi.nlm.nih.gov/34566780/.

[14] Ebrahim Babaei et al. 'Should we use the NASA-TLX in HCI? A review of theoretical and methodological issues around Mental Workload Measurement'. In: *International Journal of Human-Computer Studies* 201 (2025), p. 103515. doi: 10.1016/j.ijhcs.2025.103515.

[15] Valentina Bachurina and Marie Arsalidou. 'Multiple levels of mental attentional demand modulate peak saccade velocity and blink rate'. In: *Heliyon* 8.1 (2022), e08826. doi: 10.1016/j.heliyon.2022.e08826. url: https://www.sciencedirect.com/science/article/pii/S2405844022001141.

[16] Valentina Bachurina et al. 'A machine learning investigation of factors that contribute to predicting cognitive performance: Difficulty level, reaction time and eye-movements'. In: *Decision Support Systems* 155 (2022), p. 113713. doi: 10.1016/j.dss.2021.113713. url: https://www.sciencedirect.com/science/article/pii/S0167923621002232.

[17] Karlene Ball et al. 'Effects of Cognitive Training Interventions with Older Adults: A Randomized Controlled Trial'. In: *JAMA* 288.18 (2002), pp. 2271–2281. doi: 10.1001/jama.288.18.2271.

[18] Karlene K Ball et al. 'Age and visual search: expanding the useful field of view'. In: *JOSA A* 5.12 (1988), pp. 2210–2219.

[19] Paul B Baltes and Margaret M Baltes. 'Psychological perspectives on successful aging: The model of selective optimization with compensation'. In: *Successful aging: Perspectives from the behavioral sciences* 1.1 (1990), pp. 1–34.

[20] Jackson Beatty. 'Task-evoked pupillary responses, processing load, and the structure of processing resources'. In: *Psychological Bulletin* 91.2 (1982), pp. 276–292.

[21] Jackson Beatty and Daniel Kahneman. 'Pupillary response in a pitch-discrimination task'. In: *Perception & Psychophysics* 1.1 (1966), pp. 53–54.

[22] Friederike Behrens et al. 'Quantifying Physiological Synchrony through Windowed Cross-Correlation Analysis: Statistical and Theoretical Considerations'. In: *bioRxiv* (2020). doi: 10.1101/2020.08.27.269746.

[23] Yoshua Bengio, Aaron Courville and Pascal Vincent. 'Representation learning: A review and new perspectives'. In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.

[24] Annamaria Berardi, Raja Parasuraman and James V. Haxby. 'Overall Vigilance and Sustained Attention Decrements in Healthy Aging'. In: *Experimental Aging Research* 27.1 (2001), pp. 19–39. doi: 10.1080/036107301750046124.

[25] Anubhav Bhatti et al. 'CLARE: Cognitive Load Assessment in Real-time with Multimodal Data'. In: *IEEE Transactions on Cognitive and Developmental Systems* PP (2025), pp. 1–13. doi: 10.1109/TCDS.2025.3555517.

[26] Olga Vl Bitkina, Jaehyun Park and Hyun K Kim. 'The ability of eye-tracking metrics to classify and predict the perceived driving workload'. In: *International Journal of Industrial Ergonomics* 86 (2021), p. 103193.

[27] Katherine Boere et al. 'Measuring cognitive load in multitasking using mobile fNIRS'. In: *NeuroImage: Reports* 4.4 (2024), p. 100228. doi: 10.1016/j.ynirp.2024.100228.

[28] Wolfram Boucsein. *Electrodermal activity*. New York: Springer Science & Business Media, 2012.

[29] Jack W. Brehm and Elizabeth A. Self. 'The intensity of motivation'. In: *Annual Review of Psychology* 40 (1989), pp. 109–131.

[30] Leo Breiman. 'Random Forests'. In: *Machine Learning* 45.1 (2001), pp. 5–32.

[31] Pradeep Kumar Brito, G. R. Sinha and Varun Bajaj. 'Automatic muscle artifacts identification and removal from single-channel EEG using wavelet transform with meta-heuristically optimized non-local means filter'. In: *Sensors* 20.10 (2020), p. 2895.

[32] Victòria Brugada-Ramentol, Amir Bozorgzadeh and Hossein Jalali. 'Enhance VR: A Multisensory Approach to Cognitive Training and Monitoring'. In: *Frontiers in Digital Health* Volume 4 - 2022 (2022). doi: 10.3389/fdgth.2022.916052.

[33] Peter Brusilovsky. 'Adaptive systems and personalized systems'. In: *The Adaptive Web*. Berlin, Heidelberg: Springer, 2007, pp. 1–28.

[34] Tianqi Chen and Carlos Guestrin. 'XGBoost: A Scalable Tree Boosting System'. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 785–794.

[35] B. Cheng, Y. Liu and K. Gramann. *EEG-Based Cognitive Load Classification During Landmark-Based VR Navigation*. 2025. arXiv: 2509.14056.

[36] Francesco Chiossi et al. 'Adapting Visual Complexity Based on Electrodermal Activity Improves Working Memory Performance in Virtual Reality'. In: *Proceedings of the ACM on Human-Computer Interaction* 7.MHCI (2023), pp. 1–26.

[37] Aaron Cochrane, Vanessa Simmering and C Shawn Green. 'Load effects in attention: Comparing tasks and age groups'. In: *Attention, Perception, & Psychophysics* 82 (2020), pp. 3072–3084.

[38] Ronald A. Cohen. 'Continuous Performance Tests'. In: *Encyclopedia of Clinical Neuropsychology*. Ed. by Jeffrey S. Kreutzer, John DeLuca and Bruce Caplan. New York, NY: Springer New York, 2011, pp. 699–701. doi: 10.1007/978-0-387-79948-3_1280.

[39] Thomas Constant and Guillaume Levieux. 'Dynamic Difficulty Adjustment Impact on Players' Confidence'. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Glasgow, Scotland UK: ACM, 2019, pp. 1–12. doi: 10.1145/3290605.3300693.

[40] Thomas Constant et al. 'From Objective to Subjective Difficulty Evaluation in Video Games'. In: *Entertainment Computing – ICEC 2017*. Vol. 10507. Lecture Notes in Computer Science. Springer, 2017, pp. 3–10. doi: 10.1007/978-3-319-67684-5_1.

[41] Corinna Cortes and Vladimir Vapnik. 'Support-vector networks'. In: *Machine Learning* 20.3 (1995), pp. 273–297.

[42] A. Coutrot, J. H. Hsiao and A. B. Chan. 'Scanpath modeling and classification with hidden Markov models'. In: *Behavior Research Methods* 50 (2018), pp. 362–379. doi: 10.3758/s13428-017-0876-8.

[43] Mihaly Csikszentmihalyi. *Flow: The psychology of optimal experience*. Harper & Row, 1990.

[44] Yann N Dauphin et al. 'Identifying and attacking the saddle point problem in high-dimensional non-convex optimization'. In: *Advances in neural information processing systems*. 2014, pp. 2933–2941.

[45] Fred D. Davis. 'Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology'. In: *MIS Quarterly* 13.3 (1989), pp. 319–340. doi: 10.2307/249008.

[46] Nicolas Debue and Cécile van de Leemput. 'What does germane load mean? An empirical contribution to the cognitive load theory'. In: *Frontiers in Psychology* 5 (2014), p. 1099. doi: 10.3389/fpsyg.2014.01099.

[47] L. L. Di Stasi et al. 'Saccadic peak velocity sensitivity to variations in mental workload'. In: *Aviation, Space, and Environmental Medicine* 81.4 (2010), pp. 413–417. doi: 10.3357/asem.2579.2010.

[48] Tobias Drey et al. 'Towards Progress Assessment for Adaptive Hints in Educational Virtual Reality Games'. In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI EA '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–9. doi: 10.1145/3334480.3382789.

[49] Andrew T Duchowski and Andrew T Duchowski. *Eye tracking methodology: Theory and practice*. Springer, 2017.

[50] Mirjam Eckert et al. 'Pupillary Light Reflex Correction for Robust Pupillometry in Virtual Reality'. In: *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 5.2 (2022). doi: 10.1145/3530798.

[51] Stephen H. Fairclough. 'Psychophysiological inference and physiological computer interaction'. In: *International Journal of Human-Computer Studies* 67.3 (2009), pp. 226–240.

[52] Kristin E Flegal, J Daniel Ragland and Charan Ranganath. 'Adaptive task difficulty influences neural plasticity and transfer of training'. In: *NeuroImage* 188 (2019), pp. 111–121.

[53] Stephen M Fleming and H C Lau. 'How to measure metacognition'. In: *Frontiers in Human Neuroscience* 8 (2014), p. 443. doi: 10.3389/fnhum.2014.00443.

[54] Marvin Formosa. 'Universities of the Third Age: A rationale for transformative education in later life'. In: *Journal of Transformative Education* 8.3 (2010), pp. 197–219. doi: 10.1177/1541344611419857.

[55] Edith Galy, Maude Cariou and Claudine Mélan. 'The contribution of mental workload models to an understanding of age-related cognitive decline'. In: *International Journal of Psychophysiology* 83.1 (2012), pp. 1–8.

[56] Ian Goodfellow, Yoshua Bengio and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[57] David M. Green and John A. Swets. *Signal detection theory and psychophysics*. Wiley New York, 1966.

[58] Rebecca A. Grier. 'How High is High? A Meta-Analysis of NASA-TLX Global Workload Scores'. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 59. 1. SAGE Publications, 2015, pp. 1727–1731. doi: 10.1177/1541931215591373.

[59] Léo Grinsztajn, Edouard Oyallon and Gaël Varoquaux. 'Why do tree-based models still outperform deep learning on typical tabular data?' In: *Advances in Neural Information Processing Systems*. 2022.

[60] Zhixing Guo and Xiangshi Ren. 'An exploratory study of the relationship between objective game difficulty and subjective game difficulty'. In: *International Journal of Human-Computer Studies* 199 (2025), p. 103502. doi: 10.1016/j.ijhcs.2024.103502.

[61] Zhixing Guo, Ruck Thawonmas and Xiangshi Ren. 'Rethinking dynamic difficulty adjustment for video game design'. In: *Entertainment Computing* 50 (2024), p. 100663. doi: 10.1016/j.entcom.2024.100663.

[62] Calum A Hamilton et al. 'Sustained attention in mild cognitive impairment with Lewy bodies and Alzheimer's disease'. In: *Journal of the International Neuropsychological Society* (2023), pp. 1–7.

[63] Peter A. Hancock. 'The effect of performance failure and task demand on the perception of mental workload'. In: *Applied Ergonomics* 20.3 (1989), pp. 197–205.

[64] Simon Hanzal, Lucie Tvrda and Monika Harvey. 'An investigation into discomfort and fatigue related to the wearing of an EEG neurofeedback headset'. In: *medRxiv* (2023). Preprint.

[65] Sandra G Hart. 'NASA-task load index (NASA-TLX); 20 years later'. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 50. 9. Sage Publications Sage CA: Los Angeles, CA. 2006, pp. 904–908.

[66] Sandra G Hart and Lowell E Staveland. 'Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research'. In: *Advances in psychology*. Vol. 52. Elsevier, 1988, pp. 139–183.

[67] Joshua K. Hartshorne and Laura T. Germine. 'When Does Cognitive Functioning Peak? The Asynchronous Rise and Fall of Different Cognitive Abilities Across the Life Span'. In: *Psychological Science* 26.4 (2015), pp. 433–443. doi: 10.1177/0956797614567339.

[68] Dengbo He et al. 'Classification of driver cognitive load: Exploring the benefits of fusing eye-tracking and physiological measures'. In: *Transportation research record* 2676.10 (2022), pp. 670–681.

[69] Keith C Hendy, Kevin M Hamilton and Lois N Landry. 'Measuring subjective workload: When is one scale better than many?' In: *Human Factors* 35.4 (1993), pp. 579–601. doi: 10.1177/001872089303500401.

[70] Joseph Henrich, Steven J. Heine and Ara Norenzayan. 'The weirdest people in the world?' In: *Behavioral and Brain Sciences* 33.2-3 (2010), pp. 61–83. doi: 10.1017/S0140525X0999152X.

[71] Christopher Hertzog et al. 'Enrichment Effects on Adult Cognitive Development: Can the Functional Capacity of Older Adults Be Preserved and Enhanced?' In: *Psychological Science in the Public Interest* 9.1 (2008), pp. 1–65. doi: 10.1111/j.1539-6053.2009.01034.x.

[72] Joel Hestness et al. *Deep learning scaling is predictable, empirically*. 2017. arXiv: 1712.00409.

[73] Sepp Hochreiter and Jürgen Schmidhuber. 'Flat Minima'. In: *Neural Computation* 9.1 (1997), pp. 1–42.

[74] Sepp Hochreiter and Jürgen Schmidhuber. 'Long short-term memory'. In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[75] Nina Hollender et al. 'Integrating cognitive load theory and concepts of human-computer interaction'. In: *Computers in Human Behavior* 26.6 (2010), pp. 1278–1288. doi: 10.1016/j.chb.2010.05.031.

[76] Stephen Hutt et al. 'Breaking Out of the Lab: Mitigating Mind Wandering with Gaze-Based Attention-Aware Technology in Classrooms'. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, 2021, pp. 1–14. doi: 10.1145/3411764.3445269.

[77] Toshiharu Igarashi et al. 'Research participant selection bias in the workshop using socially assistive robots for older adults and its effect on population representativeness'. In: *International Journal of Environmental Research and Public Health* 20.10 (2023), p. 5915. doi: 10.3390/ijerph20105915.

[78] Sergey Ioffe and Christian Szegedy. 'Batch normalization: Accelerating deep network training by reducing internal covariate shift'. In: *International conference on machine learning*. 2015, pp. 448–456.

[79] Antony William Joseph, J Sharmila Vaiz and Ramaswami Murugesh. 'Modeling cognitive load in mobile human computer interaction using eye tracking metrics'. In: *Advances in Artificial Intelligence, Software and Systems Engineering (AHFE 2021)*. Springer. 2021, pp. 99–106.

[80] Daniel Kahneman. *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall, 1973.

[81] Slava Kalyuga et al. 'The Expertise Reversal Effect'. In: *Educational Psychologist* 38.1 (2003), pp. 23–31. doi: 10.1207/S15326985EP3801_4.

[82] Jared Kaplan et al. *Scaling laws for neural language models*. 2020. arXiv: 2001.08361.

[83] Bsher Karbouj, Baha Eddin Gaaloul and Jörg Krüger. *GAZELOAD: A Multimodal Eye-Tracking Dataset for Mental Workload in Industrial Human–Robot Collaboration*. 2026. arXiv: 2601.21829.

[84] Robert S Kennedy et al. 'Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness'. In: *The international journal of aviation psychology* 3.3 (1993), pp. 203–220.

[85] Nitish Shirish Keskar et al. 'On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima'. In: *International Conference on Learning Representations*. 2017.

[86] J. Kim et al. 'A hidden Markov model for analyzing eye-tracking of moving objects'. In: *Behavior Research Methods* 52 (2020), pp. 1225–1243. doi: 10.3758/s13428-019-01313-2.

[87] Diederik P Kingma and Jimmy Ba. *Adam: A method for stochastic optimization*. 2014. arXiv: 1412.6980.

[88] Panagiotis Kourtesis et al. 'VR-EASY: A unified framework for the development of virtual reality-based cognitive assessment and training software'. In: *Frontiers in Psychology* 10 (2019), p. 2026.

[89] Krzysztof Krejtz et al. 'Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze'. In: *PLOS ONE* 13.9 (2018), pp. 1–23. doi: 10.1371/journal.pone.0203629.

[90] Emmanouil Ktistakis et al. 'COLET: A dataset for COgnitive workLoad estimation based on Eye-Tracking'. In: *Computer Methods and Programs in Biomedicine* 224 (2022), p. 106989.

[91] S. Kunjan et al. 'The Necessity of Leave One Subject Out (LOSO) Cross Validation for EEG Disease Diagnosis'. In: *Brain Informatics*. Ed. by M. Mahmud et al. Vol. 12960. Lecture Notes in Computer Science. Cham: Springer, 2021, pp. 558–567. doi: 10.1007/978-3-030-86993-9_50.

[92] Sylvain Laborde, Emma Mosley and Julian F. Thayer. 'Heart rate variability and cardiac vagal tone in psychophysiological research - Recommendations for experiment planning, data analysis, and data reporting'. In: *Frontiers in Psychology* 8 (2017), p. 213.

[93] Marta Lagomarsino et al. 'An Online Framework for Cognitive Load Assessment in Industrial Tasks'. In: *Robotics and Computer-Integrated Manufacturing* 78 (2022), p. 102380. doi: 10.1016/j.rcim.2022.102380.

[94] D. G. Lapitan, D. A. Rogatkin, E. A. Molchanova et al. 'Estimation of phase distortions of the photoplethysmographic signal in digital IIR filtering'. In: *Scientific Reports* 14 (2024), p. 6546. doi: 10.1038/s41598-024-57297-3.

[95] O. F. P. Larsen et al. 'A method for synchronized use of EEG and eye tracking in fully immersive VR'. In: *Frontiers in Human Neuroscience* 18 (2024), p. 1347974. doi: 10.3389/fnhum.2024.1347974.

[96] Nilli Lavie. 'Distracted and confused?: Selective attention under load'. In: *Trends in Cognitive Sciences* 9.2 (2005), pp. 75–82. doi: 10.1016/j.tics.2004.12.004.

[97] Yann LeCun, Yoshua Bengio and Geoffrey Hinton. 'Deep learning'. In: *Nature* 521.7553 (2015), pp. 436–444. doi: 10.1038/nature14539.

[98] Alexander Leube and Katharina Rifai. 'Sampling rate influences saccade detection in mobile eye tracking of a reading task'. In: *Journal of Eye Movement Research* 10.3 (2017), pp. 1–11. doi: 10.16910/jemr.10.3.3.

[99] Junde Li et al. 'Health monitoring through wearable technologies for older adults: Smart wearables acceptance model'. In: *Applied Ergonomics* 75 (2019), pp. 162–169. doi: 10.1016/j.apergo.2018.10.006.

[100] Ulman Lindenberger, Elisabeth Wenger and Martin Lövdén. 'Towards a stronger science of human plasticity'. In: *Nature Reviews Neuroscience* 18.5 (2017), pp. 261–262. doi: 10.1038/nrn.2017.44.

[101] David Lindlbauer, Anna Maria Feit and Otmar Hilliges. 'Context-Aware Online Adaptation of Mixed Reality Interfaces'. In: *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. ACM, 2019, pp. 147–160.

[102] Anqi Liu, Hong Wang and Brian D. Ziebart. *Robust Covariate Shift Prediction with General Losses and Feature Views*. 2017. arXiv: 1712.10043.

[103] Yichuan Liu, Hasan Ayaz and Patricia A Shewokis. 'Multisubject "learning" for mental workload classification using concurrent EEG, fNIRS, and physiological measures'. In: *Frontiers in Human Neuroscience* 11 (2017), p. 389.

[104] Pedro López, Ana M. Bernardos and José R. Casar. 'Cognitive Load in Mixed Reality: Technology Order and Prior Experience in Eye-Tracking'. In: *Proceedings of the 2024 ACM Symposium on Spatial User Interaction*. SUI '24. New York, NY, USA: Association for Computing Machinery, 2024. doi: 10.1145/3630616.3630628.

[105] Malte Lüken, Šimon Kucharsý and Ingmar Visser. 'Characterising eye movement events with an unsupervised hidden Markov model'. In: *Journal of Eye Movement Research* 15.1 (2022), p. 4. doi: 10.16910/jemr.15.1.4.

[106] Scott M Lundberg and Su-In Lee. 'A unified approach to interpreting model predictions'. In: *Advances in Neural Information Processing Systems 30*. 2017.

[107] Norman H Mackworth. 'The breakdown of vigilance during prolonged visual search'. In: *Quarterly Journal of Experimental Psychology* 1.1 (1948), pp. 6–21.

[108] Neil A. Macmillan and C. Douglas Creelman. *Detection theory: A user's guide*. 2nd. Mahwah, NJ: Lawrence Erlbaum Associates, 2005.

[109] Lukas Magel. *PyShimmer*. 2023. url: https://github.com/seemoo-lab/pyshimmer.

[110] Nicolas Mascret et al. 'Acceptance of a Virtual Reality Headset Designed for Fall Prevention in Older Adults: Questionnaire Study'. In: *Journal of Medical Internet Research* 22.12 (2020), e20691. doi: 10.2196/20691.

[111] Ashley M McPhee, Tsz Ching Karen Cheung and Mark A Schmuckler. 'Dual-task interference as a function of varying motor and cognitive demands'. In: *Frontiers in Psychology* 13 (2022), p. 952245. doi: 10.3389/fpsyg.2022.952245.

[112] Barbara A. Mellers and Michael H. Birnbaum. 'Loci of contextual effects in judgment'. In: *Journal of Experimental Psychology: Human Perception and Performance* 8.4 (1982), p. 582.

[113] Luis Mendoza, Kevin Schmidt and Michael Werner. 'The pitfalls of random cross-validation in physiological computing: A case for Leave-One-Subject-Out protocols'. In: *Frontiers in Neuroscience* 14 (2020), p. 580.

[114] Gabriella Miles et al. 'EM-COGLOAD: An investigation into age and cognitive load detection using eye tracking and deep learning'. In: *Computational and Structural Biotechnology Journal* 24 (2024), pp. 264–280.

[115] Fatemeh Mortazavi, Hadi Moradi and Abdol-Hossein Vahabie. 'Dynamic difficulty adjustment approaches in video games: a systematic literature review'. In: *Multimedia Tools and Applications* 83 (2024), pp. 83227–83274. doi: 10.1007/s11042-024-18768-x.

[116] E. Mosqueira-Rey et al. 'Human-in-the-loop machine learning: a state of the art'. In: *Artificial Intelligence Review* 56.4 (2023), pp. 3005–3054. doi: 10.1007/s10462-022-10246-w.

[117] Khadijeh Moulaei et al. 'Efficacy of virtual reality-based training programs and games on the improvement of cognitive disorders in patients: a systematic review and meta-analysis'. In: *BMC Psychiatry* 24.1 (2024), p. 116. doi: 10.1186/s12888-024-05563-z.

[118] Christian Mühl, Camille Jeunet and Fabien Lotte. 'EEG-based workload estimation across affective contexts'. In: *Frontiers in Neuroscience* 8 (2014), p. 114. doi: 10.3389/fnins.2014.00114.

[119] Douglas P Munoz et al. 'Age-related performance of human subjects on saccadic eye movement tasks'. In: *Experimental Brain Research* 121 (1998), pp. 391–400.

[120] M. Nyström, R. Andersson, D. C. Niehorster et al. 'What is a blink? Classifying and characterizing blinks in eye openness signals'. In: *Behavior Research Methods* 56 (2024), pp. 3280–3299. doi: 10.3758/s13428-023-02333-9.

[121] Maximilian P. Oppelt et al. 'ADABase: A multimodal dataset for cognitive load estimation'. In: *Sensors* 23.1 (2023), p. 340. doi: 10.3390/s23010340.

[122] Allen Parducci. 'Category judgment: A range-frequency model'. In: *Psychological Review* 72.6 (1965), p. 407.

[123] Allen Parducci. 'Category rating and the relational character of judgment'. In: *Directional Psychology*. Springer, 1983, pp. 262–282. doi: 10.1007/978-1-4612-5389-1_13.

[124] Denise C. Park and Patricia Reuter-Lorenz. 'The Adaptive Brain: Aging and Neurocognitive Scaffolding'. In: *Annual Review of Psychology* 60 (2009), pp. 173–196. doi: 10.1146/annurev.psych.59.103006.093656.

[125] Thomas D Parsons. 'Virtual reality and the rehabilitation of cognitive impairments'. In: *Neuropsychology Review* 25.2 (2015), pp. 1–20.

[126] Ronald C. Petersen. 'Mild Cognitive Impairment as a Diagnostic Entity'. In: *Journal of Internal Medicine* 256.3 (2004), pp. 183–194. doi: 10.1111/j.1365-2796.2004.01388.x.

[127] T. D. Pham. 'Time–frequency time–space LSTM for robust classification of physiological signals'. In: *Scientific Reports* 11 (2021), p. 6936. doi: 10.1038/s41598-021-86432-7.

[128] V Pichot et al. 'Heart rate variability and cognitive function'. In: *Clinical Autonomic Research* 15.1 (2005), pp. 17–25.

[129] A. Pinilla et al. 'Real-time affect detection in virtual reality: a technique based on a three-dimensional model of affect and EEG signals'. In: *Frontiers in Virtual Reality* 3 (2023), p. 1088647.

[130] T. Piquado, D. Isaacowitz and A. Wingfield. 'Pupillometry as a measure of cognitive effort in younger and older adults'. In: *Psychophysiology* 47.3 (2010), pp. 560–569. doi: 10.1111/j.1469-8986.2009.00947.x.

[131] Han Qiao et al. 'Exploring the Peak-End Effects in Air Traffic Controllers' Mental Workload Ratings'. In: *Human Factors* (2021). doi: 10.1177/0018720821994355.

[132] L. Quintero, U. Fors and P. Papapetrou. 'Personalized Feature Importance Ranking for Affect Recognition From Behavioral and Physiological Data'. In: *IEEE Transactions on Games* 17.3 (2025), pp. 594–603.

[133] Dobromir Rahnev and Rachel N Denison. 'Suboptimality in perceptual decision making'. In: *Behavioral and Brain Sciences* 41 (2018), e224. doi: 10.1017/S0140525X180004

[134] M. H. Raju et al. 'Filtering Eye-Tracking Data From an EyeLink 1000: Comparing Heuristic, Savitzky-Golay, IIR and FIR Digital Filters'. In: *Journal of Eye Movement Research* 14.3 (2023), pp. 1–16. doi: 10.16910/jemr.14.3.6.

[135] Hrishikesh M Rao et al. 'Predicting cognitive load and operational performance in a simulated marksmanship task'. In: *Frontiers in Human Neuroscience* 14 (2020), p. 222.

[136] Naftali Raz et al. 'Regional Brain Changes in Aging Healthy Adults: General Trends, Individual Differences and Modifiers'. In: *Cerebral Cortex* 15.11 (2005), pp. 1676–1689. doi: 10.1093/cercor/bhi044.

[137] George W Rebok et al. 'Ten-year effects of the advanced cognitive training for independent and vital elderly cognitive training trial on cognition and everyday functioning in older adults'. In: *Journal of the American Geriatrics Society* 62.1 (2014), pp. 16–24. doi: 10.1111/jgs.12607.

[138] Patricia A Reuter-Lorenz and Katherine A Cappell. 'Selective recruitment of frontal cortex in older adults during work-memory: an fMRI study'. In: *Neurobiology of Aging* 25 (2004), S151.

[139] Patricia A Reuter-Lorenz and Denise C Park. 'The creative brain: how aging affects the neural mechanisms of cognition'. In: *Trends in Cognitive Sciences* 14.4 (2010), pp. 170–176.

[140] Michael Richter. 'Three decades of research on motivational intensity theory: What we have learned about effort and what we still don't know'. In: *Advances in Motivation Science* 3 (2016), pp. 149–186.

[141] Ian H. Robertson et al. '"Oops!": Performance correlates of everyday attentional failures in traumatic brain injured and normal subjects'. In: *Neuropsychologia* 35.6 (1997), pp. 747–758. doi: 10.1016/S0028-3932(97)00015-8.

[142] Susana Rubio et al. 'Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods'. In: *Applied Psychology* 53.1 (2004), pp. 61–86. doi: 10.1111/j.1464-0597.2004.00161.x.

[143] Cynthia Rudin. 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead'. In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215.

[144] Mikko Salminen et al. 'Meditating in a neurofeedback virtual reality: effects on sense of presence, meditation depth and brain oscillations'. In: *Behaviour & Information Technology* 43.12 (2024), pp. 2750–2764. doi: 10.1080/0144929X.2023.2258231. eprint: https://doi.org/10.1080/0144929X.2023.2258231.

[145] Dario D. Salvucci and Joseph H. Goldberg. 'Identifying fixations and saccades in eye-tracking protocols'. In: *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*. ACM, 2000, pp. 71–78. doi: 10.1145/355017.355028.

[146] R. Sarailoo et al. 'Assessment of instantaneous cognitive load imposed by educational multimedia using electroencephalography signals'. In: *Frontiers in Neuroscience* 16 (2022), p. 744737. doi: 10.3389/fnins.2022.744737.

[147] Dimitrios Saredakis et al. 'Factors Associated With Virtual Reality Sickness in Head-Mounted Displays: A Systematic Review and Meta-Analysis'. In: *Frontiers in Human Neuroscience* 14 (2020), p. 96. doi: 10.3389/fnhum.2020.00096.

[148] Hidetoshi Shimodaira. 'Improving predictive inference under covariate shift by weighting the log-likelihood function'. In: *Journal of Statistical Planning and Inference* 90.2 (2000), pp. 227–244.

[149] Mina Shojaeizadeh et al. 'Detecting task demand via an eye tracking machine learning system'. In: *Decision Support Systems* 116 (2019), pp. 91–101.

[150] Nicola Silvestrini et al. 'The curve of control: Nonmonotonic effects of task difficulty on cognitive control'. In: *Psychological Research* (2023). Integrative review of nonmonotonic task difficulty effects.

[151] Vasileios Skaramagkas et al. 'Cognitive workload level estimation based on eye tracking: A machine learning approach'. In: *IEEE International Conference on Bioinformatics and Bioengineering*. 2021, pp. 1–5.

[152] Miranda D Skurla et al. 'Virtual reality and mental health in older adults: a systematic review'. In: *International Psychogeriatrics* 34.2 (2022), pp. 143–155.

[153] Lisa G Smithers et al. 'A systematic review and meta-analysis of effects of early life non-cognitive skills on academic, psychosocial, cognitive and health outcomes'. In: *Nature human behaviour* 2.11 (2018), pp. 867–880.

[154] Nicholas C. Soderstrom and Robert A. Bjork. 'Learning versus Performance: An Integrative Review'. In: *Perspectives on Psychological Science* 10.2 (2015), pp. 176–199. doi: 10.1177/1745691615569000.

[155] S. Solhjoo, M. C. Haigney, E. McBee et al. 'Heart Rate and Heart Rate Variability Correlate with Clinical Reasoning Performance and Self-Reported Measures of Cognitive Load'. In: *Scientific Reports* 9 (2019), p. 14668. doi: 10.1038/s41598-019-50280-3.

[156] Nitish Srivastava et al. 'Dropout: A simple way to prevent neural networks from overfitting'. In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 1929–1958.

[157] Tristan Stenner. *Lab Streaming Layer (LSL) - A software framework for synchronizing a large array of data collection and stimulation devices.* 2022. url: https://github.com/sccn/labstreaminglayer.

[158] Daniel Stuhlmacher, R. Kockord and O. Bodensiek. 'Relationship Between Eye-Tracking Metrics and Cognitive Load in Mixed Reality'. In: *Proceedings of the 17th International Conference on Computer Supported Education - Volume 2: CSEDU.* INSTICC. SciTePress, 2025, pp. 216–223. doi: 10.5220/0013501400003932.

[159] Masashi Sugiyama, Matthias Krauledat and Klaus-Robert Müller. 'Covariate shift adaptation by importance weighted cross validation'. In: *Journal of Machine Learning Research* 8 (2007), pp. 985–1005.

[160] Chen Sun et al. 'Revisiting Unreasonable Effectiveness of Data in Deep Learning Era'. In: *Proceedings of the IEEE International Conference on Computer Vision.* 2017, pp. 843–852.

[161] John Sweller. 'Cognitive Load During Problem Solving Effects on Learning'. In: *Cognitive Science* 12.2 (1988), pp. 257–285.

[162] John Sweller, Jeroen J. G. van Merriënboer and Fred Paas. 'Cognitive Architecture and Instructional Design: 20 Years Later'. In: *Educational Psychology Review* 31.2 (2019), pp. 261–292. doi: 10.1007/s10648-019-09465-5.

[163] John Sweller, Jeroen J. G. van Merriënboer and Fred G. W. C. Paas. 'Cognitive Architecture and Instructional Design'. In: *Educational Psychology Review* 10.3 (1998), pp. 251–296. doi: 10.1023/A:1022193728205.

[164] Ewa Szczepocka et al. 'Virtual reality-based training may improve visual memory and some aspects of sustained attention among healthy older adults–preliminary results of a randomized controlled study'. In: *BMC Psychiatry* 24.1 (2024), p. 347.

[165] Christoph Tremmel et al. 'Estimating Cognitive Workload in an Interactive Virtual Reality Environment Using EEG'. In: *Frontiers in Human Neuroscience* 13 (2019). doi: 10.3389/fnhum.2019.00401.

[166] Merita Turunen et al. 'Computer-based cognitive training for older adults: Determinants of adherence'. In: *PLOS ONE* 14.7 (2019), e0219541. doi: 10.1371/journal.pone.0219541.

[167] Ken Umetani et al. 'Twenty-four hour time domain heart rate variability and heart rate: relations to age and gender over nine decades'. In: *Journal of the American College of Cardiology* 31.3 (1998), pp. 593–601. doi: 10.1016/S0735-1097(97)00554-8.

[168] Anne E Urai, Anke Braun and Tobias H Donner. 'Pupil-linked arousal predicts over-precision and adaptive belief updating'. In: *Nature Communications* 8.1 (2017), pp. 1–13. doi: 10.1038/s41467-017-01084-3.

[169] Antonino Vallesi et al. 'Age Differences in Sustained Attention Tasks: A Meta-Analysis'. In: *Psychonomic Bulletin & Review* 28.6 (2021), pp. 1755–1775. doi: 10.3758/s13423-021-01908-x.

[170] Paul Van Gent et al. 'HeartPy: A novel heart rate algorithm for the analysis of noisy signals'. In: *Transportation research part F: traffic psychology and behaviour* 66 (2019), pp. 368–378.

[171] Pascal W. M. Van Gerven et al. 'Cognitive Load Theory and Aging: Effects of Worked Examples on Training Efficiency'. In: *Learning and Instruction* 12.1 (2002), pp. 87–105. doi: 10.1016/S0959-4752(01)00017-2.

[172] Viswanath Venkatesh et al. 'User Acceptance of Information Technology: Toward a Unified View'. In: *MIS Quarterly* 27.3 (2003), pp. 425–478. doi: 10.2307/30036540.

[173] Michael A. Vidulich and Christopher D. Wickens. 'Causes of dissociation between subjective workload measures and performance'. In: *Applied Ergonomics* 17.4 (1986), pp. 291–296.

[174] Alexandra Vulpe and Gabriel Grigorasi. 'Saccadic velocity and duration as biomarkers for mental effort in immersive virtual environments'. In: *Scientific Reports* 13.1 (2023), p. 14201.

[175] Lev Semenovich Vygotsky. *Mind in society: The development of higher psychological processes*. Ed. by Michael Cole et al. Cambridge, Mass: Harvard University Press, 1978.

[176] Q. Wang et al. 'Characterisation of Cognitive Load Using Machine Learning Classifiers of Electroencephalogram Data'. In: *Sensors* 23 (2023), p. 8528. doi: 10.3390/s23208528.

[177] Douglas H. Wedell, Allen Parducci and R. Edward Geiselman. 'A compromise between range and frequency in judgments of physical attractiveness'. In: *Journal of Experimental Psychology: General* 116.1 (1987), pp. 52–65.

[178] Jishang Wei et al. 'Cognitive Load Inference Using Physiological Markers in Virtual Reality'. In: *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. 2025, pp. 759–769. doi: 10.1109/VR59515.2025.00098.

[179] Robert P. van der Wel and Henk van Steenbergen. 'Pupil dilation as an index of effort in cognitive control tasks: A review'. In: *Psychonomic Bulletin & Review* 25 (2018), pp. 2005–2015.

[180] Kathleen Wild et al. 'Automated cognitive monitoring in rehabilitation via the internet'. In: *International Journal of Telemedicine and Applications* 2012 (2012).

[181] Alison Williamson. 'Gender issues in older adults' participation in learning: Viewpoints and experiences of learners in the University of the Third Age (U3A)'. In: *Educational Gerontology* 26.1 (2000), pp. 49–66. doi: 10.1080/036012700267394.

[182] Sherry L Willis et al. 'Long-term effects of cognitive training on everyday functional outcomes in older adults'. In: *JAMA* 296.23 (2006), pp. 2805–2814. doi: 10.1001/jama.296.23.2805.

[183] Christopher J. Wilson and Alessandro Soranzo. 'The Use of Virtual Reality in Psychology: A Case Study in Visual Perception'. In: *Computational and Mathematical Methods in Medicine* 2015 (2015), p. 151702. doi: 10.1155/2015/151702.

[184] Barry Winn et al. 'Factors affecting light-adapted pupil size in normal human subjects'. In: *Investigative Ophthalmology & Visual Science* 35.3 (1994), pp. 1132–1137.

[185] Jonathan R Wolpaw et al. 'Brain-computer interface technology: a review of the current state of the art'. In: *IEEE Transactions on Rehabilitation Engineering* 10.2 (2002), pp. 75–84.

[186] Yei-Yu Yeh and Christopher D Wickens. 'Dissociation of performance and subjective measures of workload'. In: *Human Factors* 30.1 (1988), pp. 111–120. doi: 10.1177/001872088803000110.

[187] Dan Yu, Xun Li and Frank H.-Y. Lai. 'The Effect of Virtual Reality on Executive Function in Older Adults with Mild Cognitive Impairment: A Systematic Review and Meta-Analysis'. In: *Aging & Mental Health* 27.4 (2023), pp. 663–673. doi: 10.1080/13607863.2022.2076202.

[188] M. Zahabi and A. M. Abdul Razak. 'Adaptive virtual reality-based training: a systematic literature review and framework'. In: *Virtual Reality* 24.4 (2020), pp. 725–752.

[189] Thorsten O. Zander and Christian Kothe. 'Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general'. In: *Journal of Neural Engineering* 8.2 (2011), p. 025005.

[190] R. Zanetti et al. 'Real-Time EEG-Based Cognitive Workload Monitoring on Wearable Devices'. In: *IEEE Transactions on Biomedical Engineering* 69.1 (2022), pp. 265–277. doi: 10.1109/TBME.2021.3092206.

[191] Roberto Zangróniz et al. 'Electrodermal Activity Sensor for Classification of Calm/Distress Condition'. In: *Sensors* 17.10 (2017), p. 2324. doi: 10.3390/s17102324.

[192] Chiyuan Zhang et al. 'Understanding deep learning requires rethinking generalization'. In: *International Conference on Learning Representations*. 2017.

[193] Hongning Zhang. 'A Mental Workload Evaluation Model Based on Improved Multibranch LSTM Network with Attention Mechanism'. In: *Advances in Multimedia* 2022 (2022), p. 9601946. doi: 10.1155/2022/9601946.