UNIVERSITY OF PATRAS

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

DIVISION OF ELECTRONICS AND COMPUTER
INTERACTIVE TECHNOLOGIES LABORATORY

Design and evaluation of an intelligent system for estimating the user's cognitive load through electroencephalographic signal analysis

DIPLOMA THESIS

LYDIA EXARCHOU

SUPERVISOR: Dr. CHRISTOS FIDAS

PATRAS – FEBRUARY 2025

University of Patras, Department of Electrical and Computer Engineering.

LYDIA EXARCHOU

© 2025 – All rights reserved

The whole work is an original work, produced by Lydia Exarchou, and does not violate the rights of third parties in any way. If the work contains material which has not been produced by him/her, this is clearly visible and is explicitly mentioned in the text of the work as a product of a third party, noting in a similarly clear way his/her identification data, while at the same time confirming that in case of using original graphics representations, images, graphs, etc., has obtained the unrestricted permission of the copyright holder for the inclusion and subsequent publication of this material.

CERTIFICATION

It is certified that the Diploma Thesis titled

Design and evaluation of an intelligent system for estimating the user's cognitive load through electroencephalographic signal analysis.

of the Department of Electrical and Computer Engineering student

LYDIA EXARCHOU

Registration Number: 1072721

was presented publicly at the Department of Electrical and Computer Engineering at

25/02/2025

and was examined by the following examining committee:

Christos Feidas, Associate Professor, ECE (supervisor)

Konstantinos Moustakas, Professor, ECE (committee member)

Michalis Xenos, Professor, CEID (committee member)

The Supervisor	The Director of the Division
Christos Feidas Associate Professor	George Theodoridis Professor

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ (Η.&Υ.)
ΕΡΓΑΣΤΗΡΙΟ ΔΙΑΔΡΑΣΤΙΚΩΝ ΤΕΧΝΟΛΟΓΙΩΝ

Σχεδιασμός και αξιολόγηση ευφυούς συστήματος εκτίμησης του γνωστικού φόρτου του χρήστη μέσω ανάλυσης ηλεκτροεγκεφαλικών σημάτων

Δ Ι Π Λ Ω Μ Α Τ Ι Κ Η Ε Ρ Γ Α Σ Ι Α

ΛΥΔΙΑ ΕΞΑΡΧΟΥ

ΕΠΙΒΛΕΠΩΝ: Δρ. ΧΡΗΣΤΟΣ ΦΕΙΔΑΣ

ΠΑΤΡΑ – ΦΕΒΡΟΥΑΡΙΟΣ 2025

Πανεπιστήμιο Πατρών, Τμήμα Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών.

ΛΥΔΙΑ ΕΞΑΡΧΟΥ

© 2025 – Με την επιφύλαξη παντός δικαιώματος

Το σύνολο της εργασίας αποτελεί πρωτότυπο έργο, παραχθέν από την ΛΥΔΙΑ ΕΞΑΡΧΟΥ, και δεν παραβιάζει δικαιώματα τρίτων καθ' οιονδήποτε τρόπο. Αν η εργασία περιέχει υλικό, το οποίο δεν έχει παραχθεί από τον/την ίδιο/α, αυτό είναι ευδιάκριτο και αναφέρεται ρητώς εντός του κειμένου της εργασίας ως προϊόν εργασίας τρίτου, σημειώνοντας με παρομοίως σαφή τρόπο τα στοιχεία ταυτοποίησής του, ενώ παράλληλα βεβαιώνει πως στην περίπτωση χρήσης αυτούσιων γραφικών αναπαραστάσεων, εικόνων, γραφημάτων κ.λπ., έχει λάβει τη χωρίς περιορισμούς άδεια του κατόχου των πνευματικών δικαιωμάτων για την συμπερίληψη και επακόλουθη δημοσίευση του υλικού αυτού.

ΠΙΣΤΟΠΟΙΗΣΗ

Πιστοποιείται ότι η Διπλωματική Εργασία με τίτλο

Σχεδιασμός και αξιολόγηση ευφυούς συστήματος εκτίμησης του γνωστικού φόρτου του χρήστη μέσω ανάλυσης ηλεκτροεγκεφαλικών σημάτων

της φοιτήτριας του Τμήματος Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών

ΛΥΔΙΑ ΕΞΑΡΧΟΥ ΤΟΥ **ΠΑΥΛΟΥ**

Αριθμός Μητρώου: 1072721

Παρουσιάστηκε δημόσια στο Τμήμα Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών στις

25/02/2025

και εξετάστηκε από την ακόλουθη εξεταστική επιτροπή:

Χρήστος Φείδας, Αναπληρωτής Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών & Τεχνολογίας Η/Υ (επιβλέπων)

Κωνσταντίνος Μουστάκας, Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών & Τεχνολογίας Η/Υ (μέλος επιτροπής)

Μιχαήλ Ξένος, Καθηγητής, Τμήμα Μηχανικών Η/Υ και Πληροφορικής (μέλος επιτροπής)

Ο Επιβλέπων

Χρήστος Φείδας
Αναπληρωτής Καθηγητής

Ο Διευθυντής του Τομέα

Γεώργιος Θεοδωρίδης
Καθηγητής

Acknowledgments

I would like to thank Professor Fidas for his guidance and insightful feedback throughout this research. His expertise played a crucial role in shaping this work.

I am also thankful to my family for their unwavering support and encouragement throughout these years. Their belief in me has been a constant source of strength and motivation.

ABSTRACT

Design and evaluation of an intelligent system for estimating the user's cognitive load through electroencephalographic signal analysis.

STUDENT NAME, SURNAME: **LYDIA EXARCHOU**

SUPERVISOR NAME, SURNAME: **CHRISTOS FIDAS**

Cognitive load estimation is crucial in understanding human cognitive processes, with applications in neuroscience, education, and human-computer interaction. One of the most prominent non-invasive techniques for assessing cognitive load through neural activity is Electroencephalography (EEG). In this thesis, EEG-based cognitive load estimation is investigated with an SVM classifier while focusing on how well it generalizes across different experimental conditions. A publicly available EEG dataset, which includes an N-back working memory task, has been preprocessed through filtering, ICA, and artifact rejection. Another EEG experiment is performed with an arithmetic task to create additional variation in cognitive load measurements. Different training sets were created from data in both the public dataset and the experiment to evaluate how well the classifier generalizes to new tasks and recording conditions. Cross-task and cross-subject validation were performed to evaluate its robustness. Results indicate that while the SVM classifier achieves high accuracy in subject dependent settings, it can perform well in cross subject and cross task settings, but only with appropriate training data, otherwise, accuracy will suffer. Inter-individual variability and differences between tasks will affect classification performance. The importance of accomplishing feature selection and reduction of channels to boost classifier performance while simplifying computational effort has been studied. These results illustrate the difficulties that exist in attempting to build EEG based cognitive load classifiers and suggest ways in which real world usage can be enhanced.

ΕΚΤΕΤΑΜΕΝΗ ΕΛΛΗΝΙΚΗ ΠΕΡΙΛΗΨΗ

Σχεδιασμός και αξιολόγηση ευφυούς συστήματος εκτίμησης του γνωστικού φόρτου του χρήστη μέσω ανάλυσης ηλεκτροεγκεφαλικών σημάτων.

ΟΝΟΜΑΤΕΠΩΝΥΜΟ ΦΟΙΤΗΤΗ:
ΛΥΔΙΑ ΕΞΑΡΧΟΥ

ΟΝΟΜΑΤΕΠΩΝΥΜΟ ΕΠΙΒΛΕΠΟΝΤΟΣ:
ΧΡΗΣΤΟΣ ΦΕΙΔΑΣ

Η εκτίμηση του γνωστικού φόρτου αποτελεί βασικό ερευνητικό πεδίο στις νευροεπιστήμες, την εκπαίδευση και την αλληλεπίδραση ανθρώπου-υπολογιστή, καθώς σχετίζεται με την απόδοση των ατόμων σε απαιτητικές γνωστικές εργασίες. Η ηλεκτροεγκεφαλογραφία (EEG) προσφέρει μια μη επεμβατική μέθοδο καταγραφής της εγκεφαλικής δραστηριότητας και έχει χρησιμοποιηθεί ευρέως για την εκτίμηση του γνωστικού φόρτου μέσω της ανάλυσης εγκεφαλικών σημάτων. Στην παρούσα διπλωματική εργασία, διερευνάται η δυνατότητα ταξινόμησης του γνωστικού φόρτου με χρήση ενός μοντέλου μηχανικής μάθησης Υποστηρικτικών Διανυσματικών Μηχανών (Support Vector Machine - SVM), εστιάζοντας στη δυνατότητά του να προσαρμόζεται σε διαφορετικές πειραματικές συνθήκες.

Για την υλοποίηση της μελέτης, χρησιμοποιήθηκε ένα δημόσιο σύνολο δεδομένων EEG που περιλαμβάνει καταγραφές κατά την εκτέλεση του επιστημονικά τεκμηριωμένου τεστ N-back, ενώ σχεδιάστηκε και διεξήχθη ένα επιπλέον πείραμα στο οποίο οι συμμετέχοντες εκτέλεσαν αριθμητικές πράξεις, ως μια διαφορετική εργασία γνωστικού φόρτου. Τα σήματα EEG προεπεξεργάστηκαν μέσω τεχνικών φιλτραρίσματος συχνοτήτων, Ανάλυσης Ανεξάρτητων Συνιστωσών (ICA) και απόρριψης θορύβου, προκειμένου να βελτιωθεί η ποιότητα των δεδομένων. Στη συνέχεια, εξήχθησαν χαρακτηριστικά από τον χρονικό τομέα και τον τομέα της συχνότητας, όπως η ισχύς των ζωνών συχνοτήτων, η φασματική πυκνότητα ισχύος(Power Spectral Density - PSD), η περιπλοκότητα σήματος (Lempel-Ziv Complexity - LZC), η φασματική εντροπία και άλλοι δείκτες που σχετίζονται με τη γνωστική επεξεργασία.

Η ταξινόμηση του γνωστικού φόρτου πραγματοποιήθηκε αποκλειστικά μέσω ενός μοντέλου μηχανικής μάθησης SVM. Για την αξιολόγηση της γενίκευσης του μοντέλου, δημιουργήθηκαν διαφορετικά σύνολα εκπαίδευσης, τα οποία περιλάμβαναν δεδομένα τόσο από το N-back πείραμα όσο και από το πείραμα αριθμητικών πράξεων. Διεξήχθη διασταυρούμενη επικύρωση τόσο μεταξύ διαφορετικών συμμετεχόντων (cross-subject) όσο και μεταξύ διαφορετικών εργασιών (cross-task), προκειμένου να εξεταστεί η ικανότητα προσαρμογής του μοντέλου σε νέες συνθήκες και περιβάλλοντα καταγραφής.

Τα αποτελέσματα έδειξαν ότι το SVM μοντέλο επιτυγχάνει ικανοποιητική ακρίβεια στην αναγνώριση του γνωστικού φόρτου όταν εκπαιδεύεται και αξιολογείται σε δεδομένα από την ίδια γνωστική εργασία (task-dependent ταξινόμηση). Ωστόσο, η απόδοσή του μειώνεται σε cross-task και cross-subject σενάρια λόγω των διαφορών στα EEG σήματα μεταξύ ατόμων και της ποικιλομορφίας των γνωστικών διεργασιών που εμπλέκονται σε διαφορετικές εργασίες. Παρόλα αυτά, παρατηρήθηκε ότι η ενσωμάτωση δεδομένων από διαφορετικές εργασίες και συμμετέχοντες στο σύνολο εκπαίδευσης μπορεί να βελτιώσει σημαντικά την ακρίβεια της ταξινόμησης, ενισχύοντας τη γενίκευση του μοντέλου.

Επιπλέον, διερευνήθηκαν στρατηγικές επιλογής χαρακτηριστικών και μείωσης των ηλεκτροδίων, ώστε να βελτιωθεί η απόδοση των μοντέλων και να μειωθεί το υπολογιστικό κόστος. Συγκεκριμένα, εξετάστηκαν διαφορετικές επιλογές ηλεκτροδίων, περιλαμβάνοντας είτε το πλήρες σύνολο 32 καναλιών, είτε υποσύνολα που περιλάμβαναν κυρίως μετωπικά και κεντρικά ηλεκτρόδια, τα οποία σχετίζονται περισσότερο με τη γνωστική επεξεργασία. Παράλληλα, μελετήθηκε η μείωση του αριθμού των εξαγόμενων χαρακτηριστικών, ιδιαίτερα στο cross-task σενάριο, όπου η διατήρηση μεμονωμένων των σημαντικών χαρακτηριστικών κρίθηκε απαραίτητη για τη βελτίωση της γενίκευσης του μοντέλου. Τα αποτελέσματα έδειξαν ότι η χρήση ενός μειωμένου αριθμού καναλιών και χαρακτηριστικών μπορεί να προσφέρει παρόμοια ή ελαφρώς βελτιωμένη ακρίβεια σε σχέση με το πλήρες σύνολο, καθιστώντας δυνατή την ανάπτυξη μοντέλων με χαμηλότερες απαιτήσεις σε υπολογιστικούς πόρους και καλύτερη προσαρμοστικότητα σε νέα δεδομένα.

Συνοψίζοντας, η παρούσα διπλωματική εργασία αναδεικνύει τις προκλήσεις της γενίκευσης των μοντέλων εκτίμησης γνωστικού φόρτου σε πραγματικές συνθήκες και διερευνά στρατηγικές για τη βελτίωση της απόδοσής τους. Τα ευρήματα υπογραμμίζουν τη σημασία της χρήσης κατάλληλων συνόλων εκπαίδευσης που περιλαμβάνουν δεδομένα από διαφορετικές πηγές, καθώς και της εφαρμογής τεχνικών επιλογής χαρακτηριστικών για την αύξηση της προσαρμοστικότητας των μοντέλων. Μελλοντικές έρευνες θα μπορούσαν να επικεντρωθούν στην περαιτέρω βελτίωση της προσαρμοστικότητας των μοντέλων μηχανικής μάθησης μέσω τεχνικών μεταφοράς μάθησης (transfer learning) και πιο προηγμένων μεθόδων βαθιάς μάθησης (deep learning) για την ενίσχυση της γενίκευσης σε ποικίλες γνωστικές εργασίες και ατομικές διαφορές. Ακόμη μελλοντικά θα μπορούσε να γίνει εμπλουτισμός των δεδομένων, ειδικά από άτομα που προέρχονται από διαφορετικό δημογραφικό περιβάλλον, για παράδειγμα σε μεγαλύτερο ηλικιακό εύρος.

Table of Contents

Acknowledgements ..... Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.

Table of Figures **14**

Table of Tables **15**

2. Cognitive Load **16**

2.1 Cognitive Load and Its Importance **16**

2.2 Neural Activity and Cognitive Load **17**

2.3 Cognitive Load Tasks and Measurement **17**

2.4 Machine Learning in Cognitive Load Estimation **18**

3. Electroencephalography (EEG) **19**

3.1 Brain Physiology and Brain Waves **19**

3.1.1 Brain Waves *19*

3.1.2 Anatomy and Function of the Brain *19*

3.2 Electroencephalography (EEG) **21**

3.2.1 Basics and Techniques *22*

3.2.2 EEG Artifacts and Noise Management *23*

3.2.3 Independent Component Analysis *23*

3.3 EEG and Cognitive Load **23**

3.4 Event-Related Potentials **24**

4. Related work **25**

4.1 Cognitive load recognition and Tasks **25**

4.1.1 General Tasks *25*

4.1.2 Arithmetic Tasks *25*

4.2 Feature Extraction **26**

4.2.1 Common Features *26*

4.2.2 Channel Reduction *26*

4.3 Machine Learning Models in EEG Analysis **26**

4.3.1 Common Approaches *26*

4.3.2 Performance Evaluation *27*

4.4 Identified Research Gaps **27**

4.4.1 Task Variability in Cognitive Load Classification *27*

4.4.2 Real-World Adaptation and Hardware Variability *28*

4.4.3 Intra-Subject Validation and Individual Differences *28*

4.4.4 Cross-Task Validation and Classifier Adaptation *28*

4.4.5 EEG Feature and Channel Selection *28*

4.5 Research Questions **28**
5. Methodology **30**
5.1 Dataset and Task Description **30**
5.1.1 N-Back Task and Cognitive Load Levels *30*
5.1.2 Dataset Description and Characteristics *31*
5.2 EEG Signal Preprocessing **31**
5.2.1 Filtering and Bad Channel Detection *31*
5.2.2 Common Average Referencing *32*
5.2.3 Independent Component Analysis *32*
5.2.4 Epoch Extraction *32*
5.3 Feature Extraction **33**
5.3.1 Time-Domain Features *34*
5.3.2 Frequency-Domain Features *35*
5.3.3 PFD *36*
5.3.4 DFA *36*
5.3.5 LZC *36*
5.3.6 Entropy *36*
5.4 Classification **37**
5.4.1 Feature Selection *37*
5.4.2 SMOTE *38*
5.4.3 Classification Models *38*
5.5 Model Training and Validation **40**
5.5.1 Leave-one-participant-out cross-validation *40*
5.5.2 Performance Metrics *41*
6. Experiment Design **42**
6.1 Experimental Setup **42**
6.1.1 Equipment and Tools Used *42*
6.1.2 Participant Selection Criteria *43*
6.2 Experiment Protocol **43**
6.2.1 N-Back Task Protocol *44*
6.2.2 Arithmetic Task Protocol *45*
6.3 Experiment Procedure **45**
6.3.1 Pre-experiment Setup *45*
6.3.2 Experiment Phase – N-Back Task *48*
6.3.3 Experiment Phase – Arithmetic Task *48*
6.3.4 Triggers during the experiment phase *49*

7. Results and Analysis 51
7.1 Perfomance Results 51
7.1.1 N-back Task 51
7.1.2 Arithmetic Task 51
7.2 NASA Task Load Index Scores 52
7.3 Preprocessing Results 54
7.3.1 Preprocessing N-back dataset 54
7.3.2 Feature Selection Results 57
7.3.3 Preprocessing the experiment's data 57
7.4 Classification Performance on N-Back Dataset 57
7.4.1 Electrode Selection 57
7.4.2 All Features vs Band Power 58
7.4.3 Performance Analysis 59
7.5 Generalization to the Designed Experiment 60
7.5.1 Initial Classifier Performance 60
7.5.2 Training with experiment's N-Back Data 63
7.5.3 Training with Experiment's N-Back and Arithmetic Task Data 67
7.5.4 Training with experiment's Arithmetic Task Data 69
7.6 Comparative Analysis 73
7.6.1 Insights into Cognitive Load Variations 73
7.6.2 Accuracy and Limitations of Predictions 75
8. Discussion 76
8.1 Implications of Findings 76
8.2 Challenges Encountered 76
8.3 Comparison with Related Studies and Future Research 77
9. Citation 78
10. Appendix 87
10.1 Participant Consent Form 87
10.2 Pre-Experiment and Post-Experiment Questionnaires 89
10.3 Task Interface 91
10.4.1 N-Back Task 91
10.3.2 Arithmetic Task 92

Table of Figures

Figure 1: Brain Regions 21
Figure 2: The 10-20 system 22
Figure 3: Electrode Placement 34
Figure 4: N-Back Task 44
Figure 6: Electrode Setup 46
Figure 7: Actiview software indicating misplacement 46
Figure 8: Channels' offset 47
Figure 9: Normal EEG signals in Actiview Software 47
Figure 10: Subject isolated and monitored 47
Figure 11: N-Back Task Process 48
Figure 12: Arithmetic Task Process 48
Figure 13: Subject during the arithmetic task 49
Figure 14: Power Spectrum Before Filtering 54
Figure 15: ICA Components for Subject 3 55
Figure 16: Dropped Epochs for Subject 3 55
Figure 17: Responsible Channels for Dropped Epochs 56
Figure 18: Activity after Visual Stimuli 56
Figure 19: Beta frequencies in Resting State (Left), Low Difficulty Task (Middle) and High Difficulty Task (Right) 57
Figure 20: Accuracy after Channel Selection 58
Figure 21: Per Subject Accuracy in Binary SVM 59
Figure 22: Per Subject Accuracy in Multi-class SVM 60
Figure 23: Old Classifier - Per Subject Accuracy - Multi-class 61
Figure 24: Confusion Matrix - Old Classifier – Multiclass 61
Figure 25: Per Subject Accuracy - Old Classifier - Binary 62
Figure 26: Confusion Matrix - Old Classifier - Binary 62
Figure 27: Per Subject Accuracy - Adapted Classifier 1 - Multiiclass 64
Figure 28: Confusion Matrix - Adapted Classifier 1 - Multiiclass 65
Figure 29: Per Subject Accuracy - Adapted Classifier 1 - Binary 65
Figure 30: Confusion Matrix - Adapted Classifier 1 - Binary 66
Figure 31: Per Subject Accuracy - Adapted Classifier 2 - Multi-class 67
Figure 32: Confusion Matrix - Adapted Classifier 2 - Multi-class 68
Figure 33: Per Subject Accuracy - Adapted Classifier 2 - Binary 68
Figure 34: Confusion Matrix - Adapted Classifier 2 - Binary 69
Figure 35: Per Subject Accuracy - Adapted Classifier 3 - Multi-class 70
Figure 36: Confusion Matrix - Adapted Classifier 3 - Multi-class 71
Figure 37: : Per Subject Accuracy - Adapted Classifier 3 - Binary 72
Figure 38: Confusion Matrix - Adapted Classifier 3 - Binary 72
Figure 39: Accuracy across Multi-class Models 73
Figure 40: Accuracy across Multi-class Models 74

Table of Tables

Table 1: Trigger Codes 50
Table 2: N-Back Task Success Rates 51
Table 3: Arithmetic Task First Session Success Rates 52
Table 4: Arithmetic Task Seconf Session Success Rates 52
Table 5: NASA TLX Ratings 53
Table 6: Perfomance Matrix - Old Classifier - Multiclass 62
Table 7: Performance Metrics - Old Classifier - Binary 63
Table 8: Performance Metrics - Adapted Classifier 1 - Multiiclass 65
Table 9: Perfomance Metrics - Adapted Classifier 1 - Binary 66
Table 10: Performance Metrics - Adapted Classifier 2 - Multi-class 68
Table 11: Perfomance Metrics - Adapted Classifier 2 - Binary 69
Table 12: Performance Metrics - Adapted Classifier 3 - Multi-class 71
Table 13: : Perfomance Metrics - Adapted Classifier 3 - Binary 73

2. Cognitive Load

Cognitive load, a central concept in cognitive psychology and educational research, refers to the mental effort required to process, manipulate and integrate information. It directly influences how individuals acquire, integrate, and utilize knowledge. This concept emerged from understanding how human beings navigate complex informational environments, recognizing that our mental processing capabilities are inherently finite and strategically allocated.

First introduced by Sweller, Cognitive Load Theory (CLT) emphasizes that human cognitive capacity is finite, with excessive demands often resulting in reduced efficiency in learning and task execution [1], [2].

2.1 Cognitive Load and Its Importance

Cognitive Load in Educational Contexts

The study of cognitive load has revolutionized instructional design, emphasizing the need to tailor educational materials to optimize learning. For instance, Mayer and Moreno demonstrated that integrating multimedia elements judiciously—such as pairing visuals with complementary narration—reduces extraneous processing, enhancing comprehension [3]. In medical education, balancing the complexity of clinical scenarios with the learner's expertise is critical to preventing cognitive overload, which can compromise learning outcomes and decision-making accuracy [4].

Innovations in adaptive e-learning platforms leverage cognitive load principles to create personalized learning environments[5]. These systems dynamically adjust instructional difficulty based on a learner's performance and engagement, ensuring sustained attention and motivation [6]. This approach is particularly beneficial for novice learners who may otherwise struggle with the intrinsic complexities of certain subjects.

Role in Performance and Decision-Making

Beyond education, cognitive load significantly impacts professional performance, particularly in high-pressure fields like medicine, aviation, and engineering[7]. For example, minimizing extraneous cognitive load through well-structured workflows and supportive technology can enhance focus and decision-making accuracy under stressful conditions[8]. In healthcare, real-time support tools and user-friendly interfaces for electronic health records can help practitioners manage their cognitive load effectively, leading to better patient outcomes[9].

Applications of EEG in Cognitive Load Measurement

Advances in neuroscience have introduced tools like electroencephalography (EEG) to assess cognitive load objectively. EEG measures brainwave activity, with specific patterns in alpha, beta, and theta waves correlating to mental effort and attention levels. These insights enable the development of intelligent systems that adapt to a user's cognitive state, such as neurofeedback training platforms and brain-computer interfaces[10]. For instance, real-time EEG feedback can help users optimize their mental focus during complex tasks, offering applications in education, rehabilitation, and high-performance environments.

Broader Implications for Design and Technology

Human-computer interaction research has adopted cognitive load principles to improve the usability and accessibility of digital interfaces. Designing systems that minimize unnecessary mental effort enhances user satisfaction and efficiency. Examples include simplified navigation

in apps, voice-controlled virtual assistants, and augmented reality tools that reduce the need for extensive multitasking[11]. Furthermore, wearable devices that monitor cognitive load through physiological signals are paving the way for more responsive and user-centric technology[12].

2.2 Neural Activity and Cognitive Load

Neural activity is directly linked to the brain's response to cognitive workload, providing critical insights into how mental effort is allocated during complex tasks. Cognitive workload, or mental workload, refers to the amount of cognitive resources required to perform a specific task, and it is influenced by both the task demands and the individual's cognitive state. Research has shown that changes in neural activity patterns correlate with variations in mental workload. For example, increased mental workload is often associated with enhanced activity in regions such as the prefrontal cortex, which is involved in executive functions like decision-making, working memory, and attention control. Conversely, under lower workload conditions, brain activity in these regions tends to decrease, reflecting a more passive cognitive state. This dynamic adjustment of neural activity plays a crucial role in helping the brain allocate resources efficiently to meet task demands[13].

One of the most commonly used methods for assessing mental workload is through Electroencephalography (EEG), which measures electrical activity in the brain. Studies have demonstrated that specific EEG patterns, such as increased beta waves and decreased alpha waves, are associated with higher cognitive workload. These findings indicate that the brain's electrical activity provides valuable insights into an individual's mental state, offering a non-invasive approach to workload assessment. Moreover, EEG-based approaches allow for real-time monitoring, which is particularly useful for applications requiring continuous feedback, such as human-computer interactions or tasks demanding high levels of attention[14].

2.3 Cognitive Load Tasks and Measurement

Cognitive load tasks are designed to assess the mental effort and processing demands placed on an individual while performing specific activities. These tasks typically require participants to engage various cognitive processes, such as working memory, attention, and problem-solving, and can be categorized based on the level of complexity and resource demands they impose. Some standardized tasks, such as the N-back task, that will be analyzed later in this thesis, the Sternberg task, the Oddball task and the Stroop task have been scientifically established in measuring Cognitive Load. The measurement of cognitive load aims to quantify the mental effort exerted during task performance, which is crucial for understanding how individuals manage cognitive resources. Different methods are employed to assess cognitive load, including subjective self-report scales (e.g., NASA-TLX), physiological measurements (e.g., heart rate, pupil dilation), and objective neurophysiological indicators (e.g., EEG, fMRI). These measurements provide valuable insights into how cognitive load fluctuates across different task demands and how individuals cope with varying levels of mental workload [70].

Subjective rating scales are the most commonly used in CLT research. Examples include adapted versions of the NASA-Task Load Index[15] or the 9-point symmetrical mental effort scale by Paas[16]. Subjective ratings often involve a single question (e.g., "Please rate the mental effort invested in the task"), with responses ranging from "very very low" to "very very high" effort, as in Paas's scale. Ratings are usually collected after each task, to capture accumulated cognitive load or, less often, after a series of tasks to measure overall cognitive load.

These scales do not capture real-time fluctuations in load unless repeatedly applied during a task. However, even with repeated measures, the large time gaps between ratings may fail to provide a continuous measure of cognitive load changes during task performance[17].

2.4 Machine Learning in Cognitive Load Estimation

In EEG-based MI-BCI systems, machine learning (ML) aims to detect brain patterns associated with a particular activity without the use of conventional statistical techniques. EEG research used to focus only on analyzing brain activity and diagnosing illnesses. However, over time, the area of BCI has seen several breakthroughs due to technical advancements. Applications of BCI have spread into other domains thanks to ML technology, offering a novel approach to researching brain activity[18].

Supervised learning algorithms, such as Support Vector Machines (SVMs), Random Forests, and Neural Networks, have demonstrated significant efficacy in classifying cognitive load levels by learning from labeled EEG datasets[19]. Furthermore, feature extraction techniques—such as time-domain, frequency-domain, and wavelet-based analyses—enhance the performance of these classifiers by providing meaningful inputs for training models[18]. Recent advances in deep learning, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have further improved classification accuracy by automating feature learning from raw EEG data[20]. However, the generalization of machine learning models to real-world conditions remains a significant challenge due to variations in individual physiology, task complexity, and environmental factors.

3. Electroencephalography (EEG)

3.1 Brain Physiology and Brain Waves

3.1.1 Brain Waves

In the past, the primary oscillations' amplitude, form, and duration were described in the first study of an EEG. According to classical theory, this generates four different kinds of waves in the 0–30 Hz frequency range.

The distinctive trace of an awake, aware subject actively thinking is the beta rhythm (β), which is characterized by fast oscillations of 13-30 Hz and around 30 μV amplitude. The cause of this rhythm would be cortico-cortical activation.

With a frequency of 8–13 Hz and an amplitude of 30–50 μV, the first historically described alpha rhythm (α) is that of an individual who is awake, eyes closed, relaxed, or meditating. The interaction of thalamic and cortical pacemakers (cortico-thalamic network) may be the cause of this tracing.

Theta rhythm (θ) has an amplitude of 50–100 μV and a relatively sluggish frequency between 4 and 7 Hz. This rhythm, which is seen during light sleep (stage 2), could be an indication of how GABAergic interneurons influence the cortico-thalamic network. It might have something to do with limbic activity (emotions and memory).

Delta rhythm (δ) has a significant amplitude (100–200 μV) and a slow frequency (0.5–4 Hz). This pattern is seen in coma and deep sleep. Without any sensory input, the cortex produces it (cortico-thalamic dissociation).

The recently described gamma rhythm (γ) is characterized by extremely fast oscillations exceeding 30 Hz. This pattern could be linked to consciousness, which is the process by which information from many parts of the brain is combined to create a cohesive idea. This rhythm may have a role in perceptual processes, reflecting the interdependence of cortico-cortical and cortico-thalamo-cortical networks (loop activity)[14].

3.1.2 Anatomy and Function of the Brain

The brain performs numerous vital tasks. It provides context for events that take place in the world around us. The brain gets messages—often many at once—through the five senses of sight, smell, hearing, touch, and taste. Thoughts, speech, memory, arm and leg motions, and the operation of numerous body organs are all governed by the brain. By controlling respiration and heart rates, it also affects how people react to stressful events. The brain is a well-organized structure made up of numerous parts that each have vital and distinct roles to play.

The nervous system

The central nervous system and the peripheral nervous system are the two common divisions of the nervous system. The brain, its cranial nerves, and the spinal cord comprise the central nervous system. The autonomous nervous system, which is further subdivided into the sympathetic and parasympathetic nervous systems, and the spinal nerves that emerge from the spinal cord make up the peripheral nervous system.

Signals are sent from the brain to the body's internal organs and other parts through the nervous system. Thus, movement, breathing, vision, thought, and other functions are all governed by the activity of the nervous system.

There are two primary components to the nervous system. The brain and spinal cord comprise the central nervous system and the nerves that emerge from the spinal cord and travel throughout the body comprise the peripheral nervous system[21].

Neurons

Neurons are the building blocks of the nervous system, with about 100 billion in the human brain. Each neuron consists of a cell body, axons, and dendrites. The cell body contains the nucleus, where most of the molecules that the neuron needs to survive and function are manufactured. Dendrites extend out from the cell body like the branches of a tree and receive messages from other nerve cells. Signals then pass from the dendrites through the cell body and travel away from the cell body down an axon to another neuron, a muscle cell, or cells in some other organ. Bundles of axons, called nerves, carry signals over long distances. Neurons have specialized roles, such as motor neurons that control movement and sensory neurons that detect light, sound, and other stimuli. The nervous system also regulates involuntary processes like heartbeat, digestion, and hormone release.

Neurons communicate using electrical and chemical signals. Electrical impulses travel along axons and are converted to chemical messages at the synapse, the gap between neurons. The signal triggers the release of microscopic sacs known as vesicles when it reaches the end of the axon. Neurotransmitters are substances released into the synaptic cleft by these vesicles. After passing across the synapse, the neurotransmitters bind to receptors on the adjacent cell. The receiving cell's characteristics can be altered by these receptors[22]. Neurotransmitters released at the synapse carry the signal to the next neuron, which converts it back to an electrical signal. Supporting the neurons are glial cells, which provide structure, insulation, repair, and regulate neurotransmitter activity. Together, neurons and glial cells form intricate networks that control body functions, emotions, behavior, and thought processes[21].

Brain regions

Cerebrum

The greatest part of the brain is the cerebrum. White matter makes up its core, while gray matter makes up its outlying regions. The brain is in charge of starting and directing movement and control of temperature. Speech, analysis, logical reasoning, decision-making, emotional reactions, and knowledge acquisition are all facilitated by other areas of the cerebrum. Additionally, several functions relate to sensory processes, including touch, hearing, vision, and other experiences. There are two hemispheres in the cerebral cortex. The hemispheres' roles in body coordination are inverted; specifically, the left hemisphere controls the right arm, and the right hemisphere the left[23]. Each of the hemispheres is further subdivided into four lobes, which can be seen in Figure 1:

The frontal lobe bears responsibility for consciousness. Sometimes separated into two lobes, the prefrontal lobe and the frontal lobe, the frontal lobe oversees thinking and is also in charge of the brain. The parietal lobe is in charge of processing information from the many senses and interacting with environmental items, the temporal lobe processes complex information stimuli and is in charge of the sensations of smell and sound, and the occipital lobe specializes in the sense of sight[24].

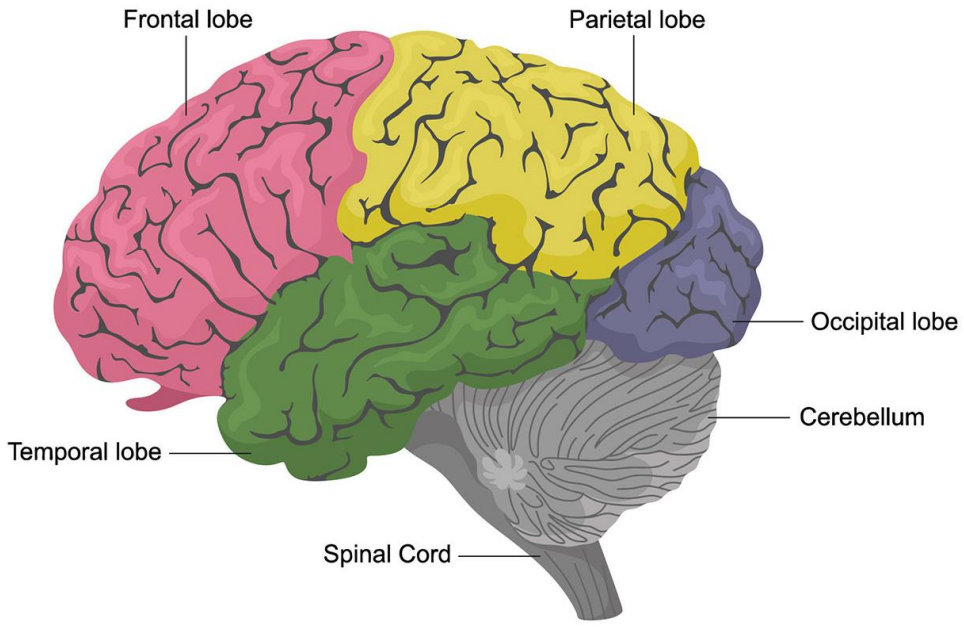


Figure 1: Brain Regions[104]

Cerebellum

The cerebellum is an essential part of the human brain because it regulates motor movement and balance. The cerebellum regulates voluntary muscle activity and tone, maintains posture, and coordinates locomotion, although it cannot start muscular contraction. Located at the back of the brain, is divided into two hemispheres joined by the vermis and consists of three lobes: anterior, posterior, and flocculonodular. Its surface is folded, with gray matter forming three layers: molecular, Purkinje, and granular. Despite its small size, the cerebellum contains 80% of the brain's neurons, crucial for coordination and balance[25].

Brainstem

A crucial connection between the cerebrum and the spinal cord is the brainstem, which is in the middle of the brain. It is made up of the medulla, pons, and midbrain and has a complex role in many body processes. Hearing, movement, and response processing are all influenced by the complex neural structures, neuronal clusters, and colliculi found in the midbrain. The substantia nigra, which is essential for coordination, is likewise housed there. A cranial nerve hub called the pons regulates a variety of functions, including vision, balance, and facial emotions. The medulla, a vital organ that controls respiration, heart rate, blood circulation, and reflexes like swallowing and sneezing, is located below. The spinal cord, which emerges from the medulla and is shielded by vertebrae, transmits signals from the brain to the body[26].

3.2 Electroencephalography (EEG)

The most effective noninvasive methods for examining the electrical dynamics of the brain and connecting them to cognition and illness are electroencephalography (EEG) and magnetoencephalography (MEG). High temporal precision and direct monitoring of human population-level brain activity are only two of the many benefits of EEG. EEG's limitation to large, synchronous populations of neurons may be its worst drawback; measuring small-scale and asynchronous activity is challenging or impossible[27].

Perhaps the first scientist to study brain potentials was Richard Caton (1842-1926). Using a galvanometer, which projects a beam of light reflected from its mirror onto a scale set on a nearby wall, he measured electric currents in the exposed brains of cats and rabbits. "Weak currents of varying directions pass through the multiplier when the electrodes are placed at two points of the external surface, or one electrode on the gray matter and one on the surface of the skull," according to the findings, which were reported in 1875. This finding can be considered an electroencephalogram activity discovery[28].

3.2.1 Basics and Techniques

The flow of neuronal ionic currents captured by a pair of electrodes placed inside or outside the head is known as an electroencephalogram (EEG). Although it will not be covered in this thesis, the intracranial EEG (iEEG), a type of EEG signal captured inside the skull, can be used for surgical planning of intractable epilepsies. Electric potential differences between electrode pairs are measured via an EEG. The electrodes can be put inside a cap (or a net) for easier attachment, or they can be directly affixed to the scalp surface at certain points.

EEG signals are now recorded in computers as sampled numerical data due to advancements in digital engineering and computer technology. We can apply a range of computational EEG analysis approaches, including source imaging, functional connectivity analysis, and time-frequency analysis, when we use a digital EEG.

Because EEG detects the potential difference between two distant scalp regions, it requires the use of at least two electrodes to capture data. EEG signals from many scalp regions can be recorded simultaneously by modern EEG recording devices. EEG recording techniques come in two varieties: unipolar and bipolar. All of the electrodes are paired in the bipolar approach, and the potential differences between each pair are noted. The potential differences between each electrode and a reference electrode are noted in the unipolar (or monopolar) approach[29].

The 10-20 system

To guarantee reproducibility, the 10-20 system was developed to standardize the electrode placement for EEG. The skull is separated into five sections based on the distance between the nasion, which is the junction of the frontal bone and two nasal bones, and the inion, which is the most noticeable projection of the occipital bone. The first area is situated 10% of the estimated distance away, while the other four are spaced 20% apart.

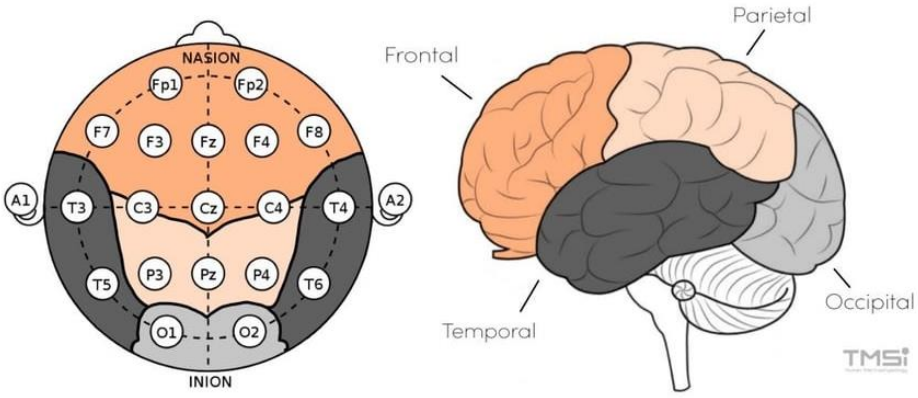


Figure 2: The 10-20 system[105]

There are five separate parts to this measurement. 10% of the total measurement corresponds to where the starting point, Fp, is located. The second, third, fourth, and fifth markers are then written as F, C, P, and O, and are placed 20% apart along the entire measurement. Notably, the O mark is placed 10% above the inion. The frontopolar, frontal, central, parietal, and occipital regions are represented by the letters Fp, F, C, P, and O. Along the geographic latitude and longitude, the remaining electrodes are positioned concerning the original markings[30].

3.2.2 EEG Artifacts and Noise Management

Any electrical potential that originates outside of the brain and shows up on the EEG trace is known as an EEG artifact. Scanning for these signal distortions, or artifacts, is a fundamental aspect of evaluating EEG recordings. Typically, EEG recordings have a larger amplitude and different form than signal sequences without significant contamination. The recorded EEG artifacts could be technical or patient-related. Unwanted physiological signals that have the potential to seriously disrupt the EEG are known as patient-related artifacts. Shorter electrode cables and lower electrode impedance can reduce technical artifacts like AC power line noise. Minor body movements, muscle activity (EMG), and heart activity (ECG) are some of the main patient-related artifacts. Eye movements cause major artifacts; excessive sweating can also affect electrode conductivity. Power line interference at the respective power line frequencies of 50/60 Hz is a very common technical cause of noise in EEG recordings. Variations of electrode impedances, movement of cables, or broken wire contacts can cause even more degradation of the signal's quality. Additionally, improper application of electrode paste either too much or too dried will contribute to poor connectivity. Low battery in the recording equipment could contribute to resorting stability of transmission signal; otherwise, it introduces unreliable data.[31]

3.2.3 Independent Component Analysis

Independent Component Analysis (ICA) is a powerful statistical technique that plays a crucial role in preprocessing EEG data by isolating the independent components that contribute to the mixed signals. After performing ICA, artifact components—such as those originating from ocular movements (EOG)—can be identified and removed, allowing for a cleaner signal that is more suitable for further analysis. In the context of EEG signal processing, ICA is particularly effective for separating brain activity from noise, as it assumes that the mixed signals result from linear combinations of independent sources, which are typically non-Gaussian in nature. By exploiting this non-Gaussianity, ICA can uncover the underlying structure of the data, distinguishing between the relevant brain signals and unwanted artifacts. The application of ICA to EEG data in this study allows for the removal of EOG-related components, improving the quality of the data for subsequent event extraction and classification. Following the ICA preprocessing, the data was resampled to 16 Hz to reduce computational complexity and prepare it for machine learning classification, enhancing the overall accuracy and efficiency of cognitive load estimation in the experiment[32].

3.3 EEG and Cognitive Load

EEG is a physiological index that can be used as a continuous online measure of cognitive load. It can identify subtle changes in instantaneous load, which can help explain the effects of instructional interventions when measures of total cognitive load are unable to account for these variations in cognitive processing[33]. Electroencephalography (EEG) has been widely utilized in cognitive load estimation due to its non-invasive nature and its ability to capture

real-time neural activity. Cognitive load is closely linked to changes in brain activity patterns. EEG provides insights into these patterns by measuring voltage fluctuations across different frequency bands, such as beta (13–30 Hz) and alpha (8–13 Hz). Studies have shown that beta activity often increases with higher cognitive load, particularly in frontal regions, while alpha activity tends to decrease as mental effort intensifies[18]

Frequency-based features derived from EEG signals are commonly used to assess cognitive load. For instance, power spectral density analysis can reveal the distribution of signal power across various frequency bands, providing a quantitative measure of mental effort[34]. Additionally, event-related potentials (ERPs), such as the P300 component, have been employed to evaluate cognitive processes, including attention and working memory, which are integral to cognitive load[18]. The spatial resolution of EEG enables researchers to study region-specific brain activity, such as the prefrontal cortex's involvement in executive functions and task complexity.

3.4 Event-Related Potentials

Multimedia learning environments present learners with rich information and varied experiences but also pose significant cognitive load challenges. Engaging with multimedia content that combines text, images, sound, and interaction often leads to information overload across multiple cognitive and perceptual channels. This complexity demands that the learner's cognitive system process and integrate diverse information, making the learning task more challenging. In addressing these challenges, event-related potentials (ERPs) have proven to be a valuable physiological tool for directly measuring cognitive load. Through EEG recordings, ERPs capture neuroelectrical activity within specific time windows, offering insights into the temporal dynamics of cognitive load during learning[18].

ERP’s high temporal resolution allows for tracking rapid brain responses to stimuli, effectively revealing the temporal characteristics of cognitive load distribution in learning tasks. Additionally, ERP circumvents the limitations of subjective assessments by providing direct physiological measurements. Its sensitivity to neural responses under varying cognitive loads enables ERP to detect brain activity changes at the millisecond level. This makes ERP an excellent tool for exploring cognitive load in multimedia learning environments, facilitating a deeper understanding of the challenges learners face, and offering objective guidance for improving multimedia learning design[18].

Unlike frequency domain EEG analysis, ERP involves averaging signals across EEG channels within a specific time window to minimize unwanted activity. ERP components are categorized based on their latency after a stimulus and are typically named according to their deflection direction and expected latency. One of the ERP's most studied components is the P3 (or P300) component, a positive voltage with a delay of 300–650 ms. According to the prevailing theory, P3 amplitude measures post-perceptual evaluation processes, namely the updating of working memory in reaction to subjectively unexpected, task-relevant events[35].

4. Related work

4.1 Cognitive load recognition and Tasks

4.1.1 General Tasks

Most research on estimating cognitive load through EEG experiment with standardized tests, including the Stroop, Oddball, Sternberg, and N-back tasks, that have been scientifically shown to measure cognitive load. In neuroscientific and cognitive studies, these tasks are frequently used to access working memory, attention, and problem-solving processes. Having them in the experimental protocol ensures standardization and comparability across different studies, which leads to a more reliable cognitive load classification model. The N-back task especially is a widely used working memory (WM) paradigm in literature and is being utilized more and more to assess individual differences[36]. Involvement of tasks as the N-back in the experimental protocols guarantees standardization and hence comparability across studies, which helps more reliable cognitive load classification models to be built. Various aspects of the EEG-based cognitive load assessment, using these tasks or variations[37], have been addressed in literature, including the efficacy of various EEG feature sets (e.g., band power, connectivity metrics)[38], [39], optimal channel selection (e.g., frontal and central regions)[39], and classifier performance[40]. Almost all these studies are carried out in a laboratory, under optimal sound and light conditions and while ensuring the subject is completely isolated from any other type of stimulation. An interesting approach to this issue is explored in [41], where subjects took the same oddball task both in a laboratory and outside under different conditions.

Apart from such tasks, a few studies explore real-world cases of high cognitive load. Some studies mention monitoring cognitive load in the workplace, especially in aviation [[42], [43]], surgery[[44], [45]], manual labor[46] and office workers[[47], [48]]. In gaming and virtual reality (VR) applications, researchers have assessed cognitive load in highly interactive environments, where factors such as immersion, sensory input, and rapid decision-making contribute to mental effort[49], [50]. Some studies have also examined cognitive load in driving simulations and real-world driving, analyzing EEG signals to detect mental fatigue, workload fluctuations, and distraction levels[51], [52]. An interested approach was explored in [53], where subjects performed a driving simulation while also taking an auditory N-back test, to capture the drivers' high cognitive load.

4.1.2 Arithmetic Tasks

An aspect of real-world problems, that require high cognitive load are all activities that include problem solving or studying. Quite a few publications experiment around such conditions, for example [54] studied subjects while they were reading, in [55] and subjects were studied while reading and understanding code and in [56] an experiment where an educational VR task for medicine students was employed. In the context of mathematical tasks, publications explored different task conditions. Mental mathematical tasks were mentioned in [57], although not analyzed, and simple arithmetic calculations with two answer options were mentioned in [58]. Mental additions were also mentioned in [59], with their difficulty classified as easy, medium and hard based on the calculations' Q-value, that is thoroughly explained in chapter 6.2.2. Study [60] also mentioned using Q-value to categorize stimuli's difficulty and in this case the subjects solved a total of 240 addition problems, ranging again in 3 stages of difficulty. Additionally, [61] mentioned auditory stimulation, by having subjects solve as many calculations as

possible in 3 minutes by audio instructions. In [62], two 2-minutes sessions were introduced, on which the subject was given 100 three-digits addition problems to complete without a range in difficulty. Logical mathematical problems were mentioned in [63], where problems stated with 8 words were classified as simple and the ones with more than 25 were considered complex, at which the subject responded with correct or false statements.

4.2 Feature Extraction

4.2.1 Common Features

Feature extraction is a crucial step in EEG-based cognitive workload recognition, as it transforms raw EEG signals into meaningful representations for classification. Traditional methods focus on extracting power spectral densities, coherence measures, and event-related potentials, which capture specific cognitive load markers[39], [46], [49], [64], [65], [66], [67]. However, recent advancements incorporate machine learning-based feature selection techniques, such as Random Forests[64] and Recursive Feature Elimination (RFE)[67], to identify the most relevant features for classification. Additionally, deep learning models are increasingly applied to automatically learn and extract hierarchical features from EEG data[68], improving classification performance while reducing manual preprocessing[69]. Despite these advancements, challenges remain, particularly in ensuring interpretability and optimizing feature selection across different tasks and individuals.

4.2.2 Channel Reduction

Most reviewed studies recorded signals from the frontal lobe. The parietal lobe was also commonly analyzed. As noted in [70], most studies have been concerned with cognitive-load assessment based on the frontal and parietal regions, whereas occipital and temporal lobes are sometimes considered in domains such as visual or auditory processing. Research using low-density EEG headsets frequently omits the central and occipital electrodes, which imposes a limited spatial sampling of cognitive processes. The frontal lobe continues to be the most analyzed and linked to aspects of decision-making, memory workload, and attentional regulation, while parietal activity has been correlated with cognitive load due to learning and fluctuations in workload .These studies highlight the regional specificity of EEG-based cognitive load assessment, with the frontal, parietal, and occipital regions being most frequently analyzed, whereas the temporal and central regions remain underexplored in some research.

4.3 Machine Learning Models in EEG Analysis

4.3.1 Common Approaches

The review article [10] encompassed the evaluation of various machine learning algorithms for classification of cognitive workload from EEG signals. The study found SVMs to be very effective in managing EEG features having high dimensionality for binary as well as multiclass cognitive workload classification. SVM was applied in [18], [19], [66]. LDA was noted to possess good simplicity and high efficiency at differentiating cognitive states from EEG data based on the concept of linear separability of the features. KNN was also discussed, as it is nonparametric and easy to implement, but can become quite cumbersome in terms of computation time when handled on large datasets.

Deep learning architectures such as CNN and RNN, were applied in [7], [20], [57]. Review [10] noted that such techniques can autonomously extract hierarchical features from EEG signals

to provide accurate classification results. However, these frameworks require large datasets and ample computational power to yield generalized results. The final remarks in article [10] stated that the choice of classification technique is a compromise dictated by relative requirements concerning accuracy, interpretability, and real-time processing in EEG-based cognitive workload estimation.

For all models, usually two to four classes were incorporated in training, based on the tasks difficulty. For two classes, either resting state vs cognitive load or low load vs high load was examined. For three or four classes, usually a resting state, and two to three levels of low, medium and high load were used.

4.3.2 Performance Evaluation

In total, three types of performance evaluation were applied. Firstly, subject-dependent cross-validation. For a total of N subjects, the experiment data of all N subjects was combined and then divided into two subsets, one used for training and one for testing, with a ratio of 80%-20% or 70%-30% [10]. Very high accuracy was usually achieved, as in [38], [62], [71], accuracy as high as 90% could be due to overfitting and the classifier generalizing. Then intra-subject types of validation were used, mostly leave-one-participant-out cross validation. In this case a classifier was built on data from N-1 participants and then tested with data from the remaining subject. Due to high inter-subject variability in EEG signals, cross-subject models exhibited lower accuracy compared to intra-subject approaches[66][64]. Lastly, cross-task evaluation was also applied, but few publications conducted experiments involving different tasks. In [66], an n-back task and an arithmetic task were performed by the same subjects, and then a cross-task classification and a within-task classification was attempted. For the cross-task part, 70% accuracy was achieved for two classes. In [64], an N-back task and a Stroop were performed. The study examined four validation scenarios for cognitive load classification pertaining to N-back and Stroop tasks. It was found that the accuracy of the leave-one-out cross-validation scores was relatively low and that this might have been due to the model failing to generalize across subjects better. A systematic bias in prediction trends was exhibited by the cross-task classification. Intra-subject/intra-context validation yielded a much higher accuracy, indicating that the cognitive load patterns were much more distinguishable within individual subjects. The Mann-Whitney U test indicators suggested that there was indeed a significant difference between the predictions coming from both contexts, meaning that although the intra-subject models were successful, the cross-task classification posed a real challenge.

4.4 Identified Research Gaps

4.4.1 Task Variability in Cognitive Load Classification

Most research concentrates on standardized tasks for working memory (WM), like the various N-back paradigms, for assessing cognitive load. However, classifiers trained on one type of cognitive task are often not good at generalizations across different task paradigms. While some attempts have been made to perform cross-task classification across similar WM tasks, it remains an open question how models could adapt across different classes of tasks that are, at least in theory, rather distant from each other, such as N-back vs. mental arithmetic. This knowledge will help us understand the robustness of classifiers to distinguish cognitive processes across task representations.

4.4.2 Real-World Adaptation and Hardware Variability

At most, the EEG-related cognitive load experiments are run under strictly regulated laboratory conditions, with light, sound, and participant movement controlled. In contrast, in real-world environments, several external distractions hinder classification of cognitive load, with variations in cognitive effort and motion artifacts. Furthermore, differences in EEG hardware amplify the complexity of such an endeavor. While most studies rely on high-resolution research-grade EEG systems, more practical applications often employ consumer-grade or portable EEG devices with fewer electrodes and lower signal quality. Thus, understanding the adaptability of cognitive load classification across various recording environments and types of EEG hardware is crucial for bringing EEG-based workload monitoring into applied environments.

4.4.3 Intra-Subject Validation and Individual Differences

The majority of studies on cognitive load have relied on cross-subject validation under which models are trained on a subset of participants while their performance is validated on an unseen subject. Due to the high variance between subjects with respect to EEG signals, however, this mode often results in reduced accuracy. On the other hand, intra-subject validation, wherein the models are trained and tested on data coming from the same subject, is generally associated with relatively higher accuracies. Nevertheless, it is still somewhat neglected in EEG-based cognitive load classification. The optimization of intra-subject modes for tracking cognitive load of an individual over long periods could increase both the personalization and reliability of such classifiers.

4.4.4 Cross-Task Validation and Classifier Adaptation

Training models for new tasks is one major challenge in EEG-based classification. Cross-task classification has been practiced within the realm of similar cognitive tasks (like different WM paradigms); however, only a small number of studies have found ways to investigate how classifiers adjust to rather disparate cognitive tasks (for example, akin to comparing the N-back task to an arithmetic task). To design task-independent cognitive load classifiers, one must grasp how EEG feature distributions might shift across tasks and how to fine-tune or retrain classifiers efficiently for new domains.

4.4.5 EEG Feature and Channel Selection

EEG feature extraction and channel selection for cognitive load classification have been explored extensively, with many studies focusing on frontal and central electrodes as key regions for cognitive load estimation. Very limited studies exist where tasks were classified in the multi-task classification manner, whereby a model is trained across different cognitive tasks. Different cognitive tasks may require the engagement of different neural processes, implying that what constitutes optimal features and electrode configuration for one task may not obtain for another. This therefore triggers the question of how EEG classifiers can adapt to varying tasks without compromising their performance, thereby needing further exploration in terms of task-independent features and channel selection.

4.5 Research Questions

Based on the identified research gaps, five research questions were raised and studied throughout the rest of this thesis.

- Research Question A: Would a classifier trained to distinguish cognitive states using data from a specific working memory task, maintain its accuracy when tested on cross-participant data? Additionally, can it generalize to new subjects without prior training on their individual EEG signals?
- Research Question B: What specific channels offer the most valuable data? Would reducing them be possible?
- Research Question C: Would a classifier trained to distinguish cognitive states using data from a specific working memory task, be able to maintain its accuracy when tested on different task data? What if this task is not a scientifically validated task and differs notably from the one used in training? What if the EEG hardware in each case differs?
- Research Question D: Can a classifier trained to distinguish cognitive states using data from a specific working memory task be adapted to maintain its accuracy when tested on different task data? Would recording participant's EEG while taking the specific working memory task and then adding it to the classifier's initial training data, help the classifier adapt to the new task and the new EEG hardware?
- Research Question E: For this cross-task classifier approach, what features would work best?

5. Methodology

Firstly, a dataset appropriate for the research was needed. Cognitive load estimation still lacks relevant open-access datasets. Although much data sharing occurs with active Brain-Computer Interfaces, datasets related to pBCIs are far less available, and many of those available often have no sufficient documentation or validation. Notably, the OpenBCI platform currently does not provide any dataset based on their open-access EEG research for cognitive load estimation analysis. The use of a well-researched and verified test condition to create a quality dataset had to suffice, so after closely examining the available options, the COG-BCI database was chosen[72], [73].

5.1 Dataset and Task Description

A thorough and authenticated resource that aims to foster research and development in pBCIs, the COG-BCI database, includes recordings from recording sessions for 29 subjects and a nested set of tasks that provoked varying cognitive states. It thus presents over 100 hours of electroencephalography and electrocardiography data, as well as subjective validations and behavioral validations for the study of cognitive load, vigilance, and decision-making. Ethical approval was obtained from the University of Toulouse (CER number 2021-342), ensuring very strict research standards compliance.

The COG-BCI database was chosen for this study because its experimental tasks are scientifically validated and widely recognized in cognitive neuroscience research. These tasks were specifically designed to elicit measurable cognitive states, ensuring their reliability for studying mental workload, vigilance, and decision-making processes. Four distinct tasks are evaluated individually over the course of three sessions (spaced precisely one week apart) in the COG-BCI database. Additionally, at the start and finish of each session, a resting state was recorded, which consisted of one minute with the eyes open and one minute with them closed. The four distinct tasks were selected to evaluate various facets of executive functioning. A variety of mental workload levels were evaluated using the MATB and N-back tasks. Vigilance decrement was examined using the Psychomotor Vigilance Task (PVT), and conflict and decision-making were examined using the Flanker task. As the n-back test was the only one utilized in this thesis, it will be thoroughly analyzed in the next paragraph.

5.1.1 N-Back Task and Cognitive Load Levels

An extensively used test for mental strain and working memory is the N-Back Task. Single numbers are shown to participants on a computer screen for a brief duration[74]. Participants are told to keep in mind the sequence in which the numbers appear and to click a button if the number that is displayed is the same as the N-th number that has been shown previously. Here, N is a variable that establishes how challenging a given block is. The work becomes more challenging as N grows since more numbers must be remembered. A number is displayed for 500ms at the start of the trial, and then the screen is blank for 1500ms. Participants are required to respond by pressing the keyboard's spacebar if the number displayed is a hit number (for example, the same number as the previous number in the 1-back condition). Participants in the 0-back condition are told to react each time a predetermined target number—such as "3"—appeals. It is set at 1/3 (16 hit tries per block) for hit numbers to appear in all three situations. Five conflict trials are added in every block under the 2-back condition. The conflict trials are distinguished by the fact that one number is immediately followed by another.

Although it has been proposed that this could lead to conflict, the participant should not respond in any way[72].

5.1.2 Dataset Description and Characteristics

Electroencephalogram (EEG) data were collected using a 64-channel setup, with electrodes positioned according to the International 10–20 system. The standardized and consistent electrode placement helps in reproducibility and hence comparability of neural recordings for all participants. The 10-20 system was developed and recognized worldwide as a method for describing and applying the positioning of scalp electrodes in the context of an EEG recording session. Besides EEG, ECG recordings were made to supplement heart activity measurements for physiological data. Each session began and ended with resting-state recordings of one minute with eyes open and one minute with eyes closed, thus establishing the baseline neural activity[75].

The dataset has been very well cross-validated and certified by subjective self-report, behavioral performance metrics, and physiological measures of cardiac and cerebral activity to make it credible and useful for the pBCI community. It also follows the Brain Imaging Data Structure (BIDS) format, thus improving usability and compatibility with other analysis tools[75].

5.2 EEG Signal Preprocessing

Significant weight is placed on proper preprocessing and, above all, on effective and sufficient noise removal. Multiple pipelines have been presented in the literature discussing the best technique to manage large data sets because of the complexity and magnitude of the acquired data. The EEG data was cleaned to eliminate noise and other artifacts that might have impacted the findings, using Python's library MNE[76]. The procedure that will be analyzed in this chapter was applied to the signals acquired both from the dataset and the experiment. Before the preprocess starts, a standard 10–20 montage is applied to align the electrode positions with a standardized coordinate system. This step is essential for maintaining spatial consistency in channel locations and enabling accurate downstream analysis. To properly view the electrode location in space, montages allow for both two-dimensional and three-dimensional representations of the electrode placement.

5.2.1 Filtering and Bad Channel Detection

Filtering in the frequency domain is an important step in the pre-processing of brain signals. This process is used to preserve the relevant information within specific frequency ranges. Among the different frequency bands in brain signals, which include delta, theta, alpha, beta, and gamma waves, this research focuses on studying all frequencies relevant to the cognitive states, so from 0.1 Hz to 35 Hz. To isolate this frequency range, a bandpass filter is used. This type of filter allows only the frequencies within the specified range to pass through while rejecting frequencies outside of it. By applying this filter, we can effectively analyze the desired frequency components in brain signals and extract valuable information for our research. Also, after this filter, a notch filter was also applied, to filter out noise due to the power line frequency and its harmonics[77].

After filtering, bad channels need to be addressed. The pipeline identifies noisy channels by using a combination of variance-based detection, invalid signal checks, and flat signal detection. Channels with variances exceeding three standard deviations above the mean are flagged as noisy. This thresholding approach suitably captures outliers with uncommonly high

amplitude fluctuations. It also marks a channel "bad" if it contains NaN or infinity, which usually happens due to hardware problems or corruption of data. Flat channels, defined as those having peak-to-peak amplitudes below 10^{-6}, are also indicated because they often indicate deference or malfunction of the electrodes.

The pipeline also applies Common Average Referencing (CAR), a widely used technique that re-references each channel to the average signal of all channels. This approach reduces the influence of global noise sources and enhances the signal-to-noise ratio by providing a uniform reference frame[78]. CAR is applied iteratively alongside the bad channel interpolation to ensure the data quality improves with each iteration.

5.2.2 Common Average Referencing

Electrodes are often used as references in EEG recordings to minimize noise and enhance signal clarity. These references can include external electrodes, such as those placed on the earlobes or mastoid area, or electrodes already in use, like T7 and T8. However, these references may sometimes degrade the signal quality due to factors like poor contact, perspiration, or electrode damage, which can interfere with accurate brain signal measurements. To avoid introducing errors, a reference should ideally be entirely neutral.

A commonly used alternative is the Common Average Reference (CAR) method. CAR works by subtracting the average activity across all electrodes from each individual electrode, establishing a zero-reference point. This approach helps reduce external noise and noise originating from the scalp. However, noisy or faulty channels can distort the average, spreading noise to all other channels.

To address this issue, the method utilized in our case combines CAR with an iterative process that detects and interpolates bad channels to calculate an accurate average. After each CAR application, noisy channels are identified and corrected, and the process is repeated until no further noisy channels are detected. This ensures a clean, noise-free reference. Once the final reference is established, CAR is reapplied, noisy channels are rechecked and corrected if necessary, and CAR is applied one final time to produce the cleanest possible signal[79].

5.2.3 Independent Component Analysis

The pipeline continues by loading the EEG data as a raw MNE object and creates fixed-length events at 1-second intervals. These events are used to segment the data into epochs, facilitating the application of ICA. The ICA decomposition is configured to retain components explaining 99% of the variance, and the model is trained on the segmented data. To identify and exclude artifacts caused by ocular movements (EOG), the function adaptively detects components associated with EOG activity using prefrontal channels (e.g., Fp1, Fp2, F7, F8). The identified components are excluded from the data, and the ICA model is saved to a specified path for reuse.

5.2.4 Epoch Extraction

The preprocessing pipeline continues by extracting event markers from the associated annotations. These events are used to create epochs, which are time-locked segments of EEG data around each event, spanning a specified window (e.g., from 100 ms before to 2000 ms after the event). A baseline correction can be applied, and epochs containing high-amplitude artifacts exceeding a predefined threshold (e.g., 200 µV) are automatically rejected to improve data quality. If the rejection criteria result in an insufficient number of valid epochs or if no

valid epochs are obtained, fixed-length epochs are created as a fallback. Additional strategies are employed to ensure an adequate number of usable epochs, such as re-epoching the data into shorter, fixed-length segments if necessary.

Artifact removal is further enhanced using Independent Component Analysis (ICA). A new ICA model is computed on the event-related epochs, independent of the ICA model previously applied to the 1-second fixed-length epochs. This ensures that the artifact removal process is tailored specifically to the statistical properties of the event-related data, which may differ from the fixed-length epochs. By computing a fresh ICA decomposition, components specific to artifacts in the event-related data, such as ocular or muscle activity, are identified and excluded. The resulting clean epochs are saved to a file, providing a reliable dataset for downstream neurophysiological analyses, including event-related potential (ERP) studies and machine learning-based classification tasks.

5.3 Feature Extraction

To manage the enormous number of samples in each experiment, it's important to extract common data across all patients and trials. The most popular techniques for data extraction are presented in this chapter and were also applied in both the data from the dataset and the experiment.

The feature extraction pipeline for EEG data begins by selecting relevant channels from preprocessed epochs. Frontal and central channels are appointed, since they are mostly related to cognitive tasks and are expected to contain some neural activities about them. Specifically, all 32 channels of the equipment were used, and three different configurations were tested. For the first setup, all channels were considered so that activity would be captured in a comprehensive manner. The second configuration incorporated only frontal and central channels since these are generally associated with cognitive processing. The third was a more targeted selection: all frontal channels were retained, and from other regions, only midline or key lateral channels were included: C3, C4, T7, T8, Pz, P3, P4, Oz, O1, and O2. This systematic comparison aimed to evaluate the influence of different channel sets on classification performance and determine the optimal spatial coverage for cognitive load estimation. This early step then leaves the rest of the pipeline to work completely and with optimization in mind toward computing task-related signals within the relevant areas of interest. For every channel selected in an epoch, several time- and frequency-domain features are computed for a complete picture of the neural activity to be presented. Among others, these include Petrosian Fractal Dimension (PFD), Lempel-Ziv Complexity (LZC), Sample Entropy, and Detrended Fluctuation Analysis (DFA). Each computes some different, yet important, aspects of the signal including its complexity, regularity, fractal geometry, and so on. Also computed is power spectral density (PSD) using the Welch method which extracts band-specific power information under delta, theta, alpha, beta, and gamma bands. Mean, standard deviation, skewness, and kurtosis are also computed and are statistical measures to describe the part of the distribution and variability of the signal.

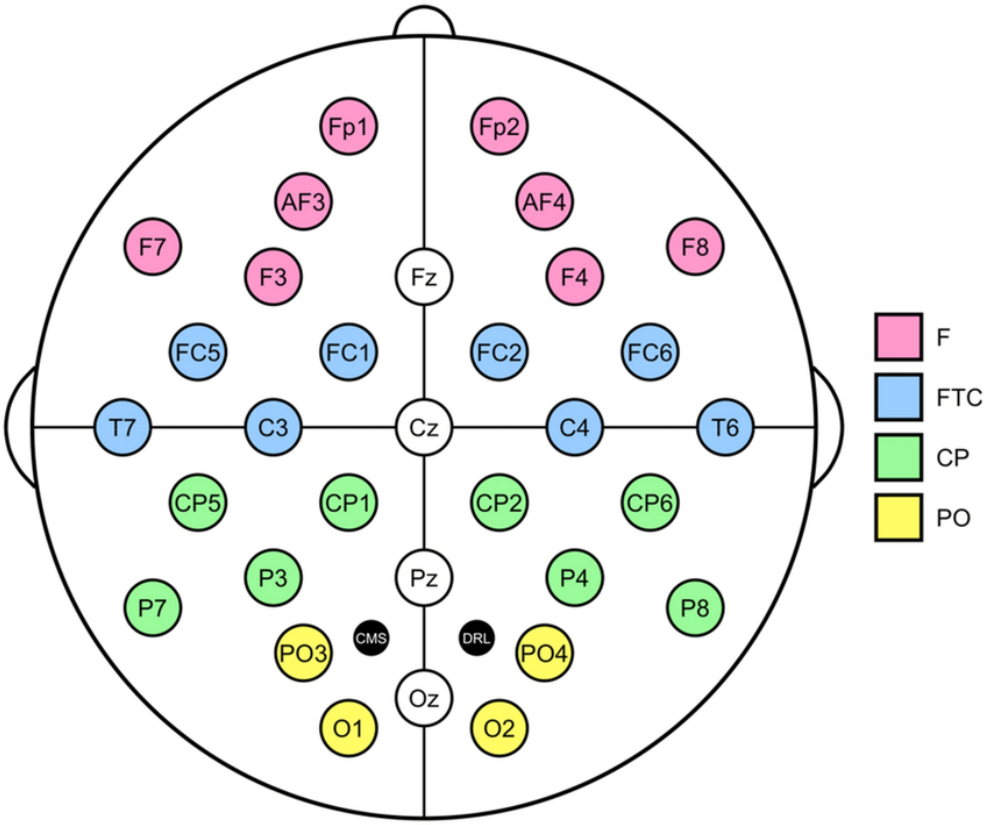


Figure 3: Electrode Placement[106]

Many features were tested while experimenting with the machine learning models that are analyzed in the next chapter, in order to achieve the desired results. Some of the features that were not chosen, as they seemed to deteriorate the classifiers' results, were all of the Hjord parameters (Activity, mobility, and complexity), Higuchi fractal dimension, approximate, fuzzy, and spectral entropy as well as some more statistical features.

After features are extracted, they undergo a normalization process whereby the features are brought to a common range and scale so that no feature can dominate in terms of range or variance. Normalization is done by z-score (standard score), which normalizes the data by subtracting the mean and dividing by the standard deviation. These features are converted under the selected method after normalization to ensure uniformity across the datasets for methods employed in subsequent machine learning tasks. Such steps will boost the performance of the classifier by ensuring that all such features are on a similar scale and do not favor one over the other[80].

5.3.1 Time-Domain Features

Time domain analysis extracts time-domain features from an EEG signal, from either raw or preprocessed EEG recordings. Time domain features are quickly and easily extracted from EEG data. Time domain analysis has lower computing complexity than frequency and time-frequency domain analysis, while these features are widely employed in medical and technical research since they do not require complex transformations and are simple to apply. Their key disadvantage is that they presume the EEG signal is stationary when EEG signals are non-stationary and change with time. Time-domain features that were extracted in this case were mean, the average of the values of all data points of the signal[81]:

$$\mu = \frac{1}{N}\left[\sum_{i=1}^{N} x_i\right]$$

Standard deviation, which measures how the data is spread out from the mean value, and is the square root of variance[81]:

$$\sigma = \sqrt{\frac{1}{N}\left[\sum_{i=1}^{N} (x_i - \mu)^2\right]}$$

Skewness, which measures the degree of asymmetry of a distribution[81]:

$$SKew(x_i) = \sqrt{\frac{\frac{1}{N}\left[\sum_{i=1}^{N}(x_i - \mu)^3\right]}{(\sigma^3)}}$$

Kurtosis, which describes the shape of the distribution of the signal data[81]:

$$kurt(x_i) = \frac{N[\sum_{i=1}^{N}(x_i - \mu)^4]}{[\sum_{i=1}^{N}(x_i - \mu)^2]^2}$$

5.3.2 Frequency-Domain Features

The input signal has been transformed into the frequency domain using the Fast Fourier Transform (FFT) technique. Theta (4 Hz–7 Hz), alpha (8 Hz–14 Hz), beta (14 Hz–30 Hz), gamma (30 Hz–60 Hz), and delta (0.5 Hz–4 Hz) are the frequency bands that are studied in EEG data. One method used to recover information from these frequency bands is by using Fast Fourier transform. By applying Fast Fourier Transform (FFT), we can extract information about the amplitude and phase of the signal for each frequency band. During this analysis, it is common to calculate the square of the amplitude of the FFT in order to obtain an estimate of the Power Spectral Density (PSD)[82].

The power spectral density (PSD) in our case is computed using the Welch method[83], instead of the classic FFT method, to derive the previous frequency band powers. This method could be said to be an improvement of the classical PSD calculation method, since it is based on the separation of the signal into overlapping or non-overlapping frequency windows in which the Fourier transform and then the square of the amplitude of the result (periodogram) are calculated. The average of all the periodograms gives the required spectrum, thus reducing the variance of the classical method. But this has the counterpoint of low frequency resolution, since frequency resolution is defined as

$$\text{Fres} = \frac{Fs}{N} = \frac{Fs}{Fs * t} = \frac{1}{t}$$

where Fs is the sampling frequency, N is the total number of samples and t is the time of each window we define. Therefore, the definition of the appropriate window is particularly important. An empirical rule used is to choose a window time to include at least two complete cycles of the lowest frequency of interest. In our case, we make use of windows of 0.25 seconds.

5.3.3 PFD

Fractal dimension (FD) examines how a pattern's detail varies with scale. A variety of geometric entities, even one-dimensional ones like time series, can be referred to by the term "pattern." It is a measure of complexity and has been used extensively in literature over EEG data[64]. Petrosian FD (PFD) and Higuchi FD (HFD) are two techniques that have been used to extract fractal dimension information from the data.

PFD [33] is a quick FD calculation:

$$PFD = \frac{\log_{10} N}{\log_{10} N + \log_{10}\left(\frac{N}{N + 0.4N_\delta}\right)}$$

where N is the signal's length and Nδ is the number of sign shifts in the first derivative of the signal.

5.3.4 DFA

An indicator of self-affinity appropriate for non-stationary signals is Detrended Fluctuation Analysis (DFA). The signal must first be detrended in order to be calculated:

$$D[i] = \sum_{j=0}^{i-1} (x[j] - \bar{x}), i = 0,1,2,\cdots,N-1$$

where D is the detrended signal, x is the input signal, $\bar{x}$ the mean of x, and N the length of the signal. The signal is split in windows of n elements, and for each window a straight line is fitted (D_n). For each n, its fluctuation is calculated using:

$$F(n) = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1} (D[i] - D_n[i])^2}$$

Given that α is the self-affinity, it may be computed as the slope of log(F(n)) plotted against log(n)[64].

5.3.5 LZC

Lossless compression algorithms have traditionally employed Lempel-Ziv Complexity (LZC) as a metric for assessing the complexity of a sequence. Furthermore, biological signals have also made extensive use of it, frequently for bandwidth estimation[64]. Binarizing a signal is the first step before calculating its LZC.

$$s[i] = \begin{cases} 0, if x[i] \leq threshold \\ 1, if x[i] > threshold \end{cases}$$

The complexity is defined as the number of different patterns found in s. It can be normalized dividing it by b, which is calculated using this formula:

$$b = N/log_2 N$$

5.3.6 Entropy

In the computation of approximate or sample entropy, three key parameters are involved: the window length (m), the similarity threshold (r), and the total length of data N. The process

begins with an EEG time series of N data points. From this, a series of overlapping segments of length m is created. Each segment, denoted as x(i), represents a block of m data points starting at time point i. This segmentation continues from i=1 to i=N−m+1, resulting in N−m+1 such segments (e.g., x(1), x(2),..., x(N−m+1)).

For each segment x(i), we measure its similarity to all other segments by defining the distance between vector $X_m(i)$ and $X_m(j)$, $d[X_m(i), X_m(j)]$, as the absolute maximum difference between their scalar components:

$$d[X_m(i), X_m(j)] = max_{k=0,\dots m-1}(|x(i+k) - x(j+k)|)$$

For each $X_m(i)$, count the number of j(1≤j≤N−m, j≠i), denoted as Bi, such that the distance between $X_m(i)$ and $X_m(j)$ is less than or equal to r. Then, for 1≤i≤N−m

$$B_i^m(r) = \frac{1}{N-m+1} B_i \; and \; B_m(r) = \frac{1}{N-m} \sum_{i=1}^{N-m} B_i^m(r)$$

Increase the dimension to m+1 and calculate A_i as the number of $X_{m+1}(i)$ within r of $X_{m+1}(j)$ where j ranges from 1 to N−m(j≠i), define $A_i^m(r)$

$$A_i^m(r) = \frac{1}{N-m+1} A_i \; and \; A_m(r) = \frac{1}{N-m} \sum_{i=1}^{N-m} A_i^m(r)$$

Thus, $B_m(r)$ is the probability that two sequences will match for m points, whereas $A_m(r)$ is the probability that two sequences will match for m+1 points. Finally, sample entropy can be defined as[84]:

$$SampEn(m, r) = \lim_{N \to \infty} -\ln \left[\frac{A^m(r)}{B^m(r)} \right]$$

5.4 Classification

Classification aims to predict target classes or variables based on supplied input. Classifiers use feature approximation to identify class boundaries and label them accordingly. In both datasets, features where computed on epochs where visual stimuli appeared for the subject to recognize it. The label of each feature is set depending on the difficulty of the problem presented to the subject. The classification aims to estimate the level of cognitive load, which correlates to the level of difficulty of the problem presented. As a resting mode was also recorded, the classifier additionally trained to differ resting states from active states.

5.4.1 Feature Selection

Feature selection was then performed with two approaches regarding the EEG-derived features suitable for cognitive load classification. In the first approach, all extracted features were included and subsequently, they were ranked through the Random Forest classifier -which is thoroughly explained on chapter 5.4.3- based on their importance towards the contribution level in the classification performance. This method allowed for a broad analysis of how different features influenced the model and identified the most effective ones. In the second approach, a narrowed scope was used where only the topmost feature -as ranked by the Random Forest classifier- to assess whether a pinpointed approach could achieve comparable classification performance with reduced computational complexity. Such comparison has made it

possible to know the one from the two approaches between the all-feature set versus small but highly relevant feature being more efficient for EEG-based cognitive load estimation.

5.4.2 SMOTE

The Synthetic Minority Over-sampling Technique (SMOTE) is one of the most recognized techniques to counter class imbalance in classification tasks through the generation of synthetic examples rather than simple duplication of existing ones. SMOTE randomly chooses one instance from the minority class and computes the k nearest neighbors of that instance. Then, synthetic samples are created along the line joining the selected instance and its neighbors. This method helps balance the dataset, thus boosting the classifier's performance while mitigating the chances of overfitting problems that may arise when using duplication. By increasing the diversity of minority class examples, SMOTE enables machine-learning models to learn generalized decision boundaries that can be especially beneficial for applications such as EEG cognitive load classification, where data imbalance directly affects predictive accuracy[85].

5.4.3 Classification Models

Support Vector Machines (SVM)

Vapnik introduced the linear SVM algorithm in 1963, and the nonlinear version in 1992[86]. SVM has grown in favor in BCI research, having originally been presented for data classification. SVM is a binary classifier that is commonly used in the supervised approach. Its goal is to find the best hyperplane for separating the training data while minimizing the predicted risk. In a high-dimensional space, a hyperplane is a decision boundary that divides data points into distinct classes. A hyperplane is a straight line that divides the data points into two classes in two-dimensional space. In three dimensions, a hyperplane is a plane that divides data points into two classes. Likewise, a hyperplane in N-dimensional space has (N-1)-dimensions. The distance between the nearest data points from each class and the decision border (hyperplane) is known as the margin. SVMs aim to minimize classification mistakes while maximizing this margin. When there is a big distance between the decision border and the nearest data points from each class, this larger margin denotes a higher level of classification confidence. The degree of feature space separation between the classes is shown by the margin. SVMs are frequently called maximum-margin classifiers since they are meant to identify the hyperplane that maximizes this margin. The data points that are closest to the hyperplane, or decision border, in a Support Vector Machine (SVM) are called Support Vectors. These data points are crucial because they establish the hyperplane's direction and location, which significantly affects the SVM's classification performance[87].

Nonlinear SVM employs kernel function to transform data into higher space to lower dimensional space, while linear SVM employs linear function as decision boundaries. It is a robust technique that contains several parameters which could be customized to improve classification accuracy as hyperplanes can be used to define linearly separable classes[18].

A polynomial kernel was selected to capture non-linear relationships in the EEG feature set, as it has been thoroughly researched[88]. The hyperparameters of the SVM are carefully configured to enhance its performance: a regularization parameter C=100 balances the trade-off between achieving a low training error and maintaining a small margin, while the kernel degree is set to 3 to allow sufficient flexibility in the decision boundary. The gamma parameter is set to 'auto' to scale inversely with the number of features, ensuring that the model complexity aligns with the dataset's dimensionality. To address potential class imbalance in the EEG

dataset, a 'balanced' class weight is applied, assigning weights to classes inversely proportional to their frequencies in the training data.

Random Forest

Random forest is a versatile machine-learning approach, an algorithm for supervised learning, widely used for classification and regression tasks. It is an extension of decision trees, designed to improve prediction accuracy and reduce overfitting. The algorithm constructs multiple decision trees during training and combines their outputs to make a final decision. Before training, the three primary hyperparameters of random forest algorithms must be established. These consist of the number of trees, the size of the node, and the number of characteristics that were sampled. Regression or classification issues can then be resolved using the random forest classifier[89].

Each tree in the ensemble of decision trees that make up the random forest method is composed of a bootstrap sample, which is a data sample taken from a training set with replacement. The out-of-bag (oob) sample is a third of the training sample that is reserved for use as test data. Feature bagging is then used to introduce another instance of randomness, increasing dataset variety and decreasing decision tree correlation. The forecast will be determined differently depending on the type of difficulty. A majority vote, or the most common categorical variable, will determine the projected class for a classification job, whereas the average of the individual decision trees will be used for regression work[89].

Random Forest algorithms are commonly used to identify the most important features from a set of EEG signals to successfully train machine learning models for cognitive load recognition. Since EEG data typically contains a large number of channels and extracted features, not all of them contribute equally to predicting cognitive load. Random Forest helps address this by computing feature importance scores based on how much each feature contributes to reducing classification error[90].

The Gini impurity measures how mixed the classes are within a node of a decision tree. It ranges from 0 (pure nodes where all samples belong to the same class) to a maximum value when the classes are evenly distributed. In Random Forest, features that result in greater reductions in Gini impurity are considered more important because they contribute more to correctly classifying cognitive load levels.

Another method, permutation-based importance, evaluates feature relevance by randomly shuffling each feature’s values and observing the impact on model accuracy. If accuracy significantly decreases after shuffling a feature, it indicates that the feature plays a crucial role in distinguishing cognitive load states.

If the accuracy significantly drops after shuffling a feature, it indicates that the feature plays a crucial role in distinguishing different cognitive load states. This approach not only helps in selecting the most relevant EEG features but also improves model interpretability and reduces computational complexity by eliminating redundant or less informative features. By leveraging these techniques, Random Forest ensures that only the most discriminative EEG features are used, leading to a more robust and generalizable cognitive load recognition model[90].

5.5 Model Training and Validation

5.5.1 Leave-one-participant-out cross-validation

Before using the recorded data from the experiment procedure, a cross-validation approach was used to validate the accuracy of the classifiers. This was implemented in two ways. In the first approach, the training set (the dataset with the n-back task recordings) was split into training and testing subsets, ensuring a balanced representation of each class across the splits. This stratification is critical for maintaining consistency in class distributions and preventing data imbalances from influencing model performance. The training subset is used to fit the classification models analyzed before, to distinguish between the target classes based on the extracted features. This approach initially scored very high in all performance metrics, but an exceptionally high classification accuracy of 98-99% is indicative of overfitting.

Overfitting happens in machine learning when an algorithm fits too closely, if not perfectly, to its training data, resulting in a model that cannot draw correct predictions or inferences from any additional data. A model overfits the training data when it describes characteristics resulting from noise or variation in the data rather than the underlying distribution from which the data were derived. Overfitting generally results in a decrease in accuracy for out-of-sample data[91].

In the second approach, the evaluation focused on assessing the classifier's generalizability by introducing new data that was not part of the original training set. To achieve this, the dataset was split based on participants, ensuring that the data from specific individuals was exclusively reserved for testing. This approach simulates real-world scenarios where the model encounters data from previously unseen subjects, providing a robust evaluation of its performance beyond the training environment. The initial data composed of 23 participants was separated into 16 participants for training and 7 for testing. The test data consisted of features extracted from various experimental sessions, representing distinct cognitive load levels, such as resting state and memory tasks (e.g., 0-back, 1-back, and 2-back). To ensure consistency with the training data, all features were normalized using a pre-trained z-score scaler, aligning the distributions of the test data with the conditions under which the model was trained. This setup aims to prevent the model from overfitting to participant-specific patterns in the training data, thereby providing a more realistic assessment of its potential deployment performance.

This second approach answers the initial scientific question, of whether the classifier can adjust well to data from new participants. Firstly, an SVM classifier was trained only on data from the resting states and data from the 2-back task, which indicated high cognitive load. In EEG-based cognitive load classification, one common approach is to simplify the problem by focusing on the two most distinct cognitive states rather than attempting to classify multiple intermediate levels. Training a binary classifier using only such data provides a clearer separation between classes, making it easier for the model to learn meaningful differences in brain activity patterns. Resting and 2-back states are at opposite ends of the cognitive load spectrum, reducing ambiguity in the training data and improving the classifier’s ability to generalize. From a practical perspective, models trained on these two states can be highly useful for real-world applications where the goal is to detect cognitive strain, fatigue, or high mental workload in dynamic environments (e.g., aviation, driving, workplace monitoring). A simple binary detection system that identifies whether a person is in a high or low cognitive load state can be more reliable and deployable than a multi-class model, which may struggle with intermediate states due to overlapping features. Moreover, training a binary model first allows for better

interpretability and understanding of EEG signal differences, before moving on to more complex classification tasks. If a binary classifier can effectively separate resting and high-load states, it demonstrates that EEG signals contain useful information for cognitive load estimation. This lays the groundwork for future research since, given the model's accuracy in binary classification, it would make it reasonable to investigate multi-class classification or adaptive models that can manage slow changes in cognitive states.

Additionally, a multi-class model is crucial for capturing the full spectrum of cognitive load, enabling more nuanced real-world applications. It distinguishes between multiple levels of workload, such as resting, 0-back (low load), 1-back (moderate load), and 2-back (high load). This is particularly important for continuous monitoring and adaptive responses, such as optimizing workload in high-risk professions or adjusting difficulty in real-time learning environments. Training a multi-class model introduces additional complexity but provides a more detailed understanding of EEG signal variations across different cognitive states. This helps uncover smooth transitions across cognitive loads or harder-to-distinguish states, guiding future improvements in feature selection and model design. Developing a multi-class model is a natural extension of the binary model, providing valuable insights into the limitations of current models and the need for better feature engineering, adaptive classification techniques, or personalized approaches.

5.5.2 Performance Metrics

Each of the models should then be evaluated in terms of efficiency. To evaluate the selected ML models, metrics commonly used in the ML field we applied, namely accuracy, precision, recall, and f1-score. The final score in each metric is derived by averaging the scores from all folds. The definition of these metrics is based on the confusion matrix consisting of the elements true positive (Tp), true negative (Tn), false positive (Fp), and false-negative (Fn). The metrics are defined as follows[92]:

Accuracy is defined as the percentage of correct predictions, so the ratio of true positives and true negatives to the sum of the total number of predictions

$$A = \frac{TP + TN}{(TP + FP + TN + FN)}$$

Precision indicates the proportion of accurate predictions, when the model predicts true positives. It is calculated by dividing true positives by total number of true positives and false positives

$$P = \frac{TP}{TP + FP}$$

Recall indicates the proportion of true positive predictions among all actual positives.

$$R = \frac{TP}{TP + FN}$$

The efficacy of imbalanced datasets may be measured by the F1 Score, which is a function of precision and recall.

$$F1\ Score = \frac{P * R}{P + R}$$

6. Experiment Design

6.1 Experimental Setup

6.1.1 Equipment and Tools Used

Hardware

The BioSemi ActiveTwo system is a versatile research tool designed to measure potential variations on the human body's surface[93]. It is widely used for recording EEG, ECG, and EMG signals and can also capture data from sensors monitoring factors like body temperature, muscle activity, blood flow, and respiration. While not intended for medical use, its flexibility makes it ideal for research applications. For this study, we utilized the BioSemi ActiveTwo MKII model, which supports up to 256 active electrodes for data collection, along with eight auxiliary channels and one for event triggering. For the experiment, only 32 electrodes were utilized.

The ActiveTwo MKII system has an A/D box that supports up to 280 channels, which can be expanded further by daisy-chaining additional A/D boxes. There are three auxiliary connectors on the front panel: one for the common mode (CMS/DRL) and two for the non-biopotential sensor signals. The box has LED indicators to signal errors or power-related information and a dial to set the speed mode, which determines the channel configuration/sample rate. The ActiveTwo system includes elastic caps with plastic holders to capture data from several electrodes on a variably sized head following the standard 10/20 system. Proper setup means measuring the participant's head with the Cz positioned as accurately as possible. Caps are built ranging from 16 to 128 channels, with variable layouts to configure for diverse research needs. In our case the 32 channel caps were used.

The system includes two types of electrodes: pin type and flat type. Using sintered Ag/AgCl, the pin type is characterized as solid, moisture-resistant, and designed to be used on the scalp for EEG measurements; it may also record other signals such as ECG when placed in specific holders. The flat type of electrobound design electrode is made to be used bare-skin for EOG, ECG, EMG, or EEG signals located, for example, at the mastoids or earlobes, as well as the neck. The ActiveTwo MKII is operated by high-performance rechargeable batteries, with an average duration of operation defined as about 15 hours, varying with the configuration of channels. It has an optical receiver, that connects it to the receiver, which converts the optical data coming from the AD-box to an USB2 output.

The receiver offers external triggering capabilities ideal for ERP setups, which were implemented in the experiment. The USB2 receiver has a trigger port with 16 independent trigger inputs and 15 independent trigger outputs, allowing precise event marking in the recorded data. Each input has an internal 10 kOhm pull-up resistor to 3.3 V, causing unconnected lines to default to a high state. These inputs appear in the data stream as a separate "Status" channel, recorded at the same sampling rate as the electrode channels for accurate synchronization. Because the triggers are managed at the receiver end, participants remain galvanically isolated, enhancing safety and reducing the risk of interference. The ActiView software automatically includes all trigger information in the .BDF file, which can be analyzed offline with third-party tools that support BioSemi's digital trigger format[94]. During the experiment, a total of 30 different triggers were sent through this port, resulting in 350 event markers for each part of the experiment.

Additionally, NEUROSPEC's MBBT Trigger Box was used instead of the USB2 reciever for some subjects. The Neurospec MMBT-S Trigger Box was used as a USB-based trigger interface to send 8-bit TTL trigger signals from the stimulus presentation computer to the EEG system via a virtual COM port connection. This device ensures precise event synchronization with biosignal amplifiers, operating with sub-millisecond latency, and is compatible with various experimental software, including PsychoPy[95].

SignaGel is a specialized conductive gel formulated to maintain reliable signal quality over extended periods, that was applied on the scalp before placing each electrode. It ensures low electrode impedance during EEG recordings and helps avoid artifacts that could distort the collected data. For application, MonoJect syringes with curved plastic tips are utilized to provide precise gel placement, minimizing discomfort for participants. This approach also prevents excessive gel usage and maintains a clean, efficient setup for consistent signal acquisition.[96].

Software

All tasks and parts of the experiment including event handling were developed locally using Python's PsychoPy® v2024.1.1[97]. According to [97], "PsychoPy is an application for the creation of experiments in behavioral science (psychology, neuroscience, linguistics, etc.) with precise spatial control and timing of stimuli". All parts were developed using Python 3.10 scripts.

For signal acquisition, Biosemi's Actiview software was used. ActiView is a software solution that displays real-time data from all ActiveTwo channels and stores it in .BDF format. Its user-friendly interface provides an immediate overview of data quality, making it easy for researchers to monitor signals as they are recorded. Along with EEG, ECG, and EMG data, ActiView supports additional inputs from the AnalogInputBox (AIB), digital triggers via the USB2 receiver, and various sensors for metrics like respiration, GSR, temperature, or plethysmography. It also offers several reference, filtering, and downsampling options. Thanks to a unique single-buffered method—using DMA transfers to a large ring buffer—ActiView remains stable and efficient during multitasking without losing data[98].

After the acquisition, all data was handled and preprocessed using Python 3.10, specifically MNE package, an "Open-source Python package for exploring, visualizing, and analyzing human neurophysiological data: MEG, EEG, sEEG, ECoG, NIRS, and more"[99]. Additionally, for feature extraction, many of the features were extracted using EEGLIB, another Python package for powerful feature extraction oriented toward EEG signals[100].

6.1.2 Participant Selection Criteria

The study's participants consisted of 10 individuals. These participants were healthy adults aged between 20 and 25 of both genders, with an academic background involving calculus to ensure consistent results. All participants were native Greek speakers to facilitate comprehension of instructions. Inclusion criteria require participants to have normal or corrected-to-normal vision, no history of neurological disorders, and to be right-handed. Exclusion criteria include any movement and mental conditions.

6.2 Experiment Protocol

A machine learning classifier was successfully trained to analyze EEG signals during an n-back task, a well-established test for measuring working memory load. The classifier managed to

distinguish between low, medium, and high cognitive load states based on brain wave patterns. However, cognitive tasks vary in nature, and the generalizability of such a classifier to other types of mental activities remains an open question.

Arithmetic tasks, which involve calculation and problem-solving, represent a different cognitive domain characterized by demands on numerical reasoning and time management[101]. These tasks naturally induce varying levels of cognitive load, making them an ideal testbed for evaluating the robustness of the classifier[62], [63], [102]. To bridge the gap between different cognitive tasks, this study introduces an n-back task as a calibration phase before participants engage in arithmetic tasks. The inclusion of this calibration step aims to align the classifier with each participant's individual brain wave patterns, thereby enhancing its accuracy and adaptability across tasks.

By comparing brain wave patterns during arithmetic tasks with those from the n-back task, the study investigates the classifier's ability to assess cognitive load across diverse scenarios. This dual-task approach aims to refine EEG-based cognitive load assessment, supporting its use in adaptive learning systems, workload monitoring, and personalized cognitive training programs.

6.2.1 N-Back Task Protocol

The experimental design calibration will incorporate an n-back test, like the one conducted by participants in the dataset experiment. In the dataset experiment, participants were required to remember one digit number as stimuli, which will also be used here. This n-back test will encompass three phases, each differing in cognitive load, which is determined by the n value. The n-back task will incorporate three phases (0-back, 1-back, and 2-back), each corresponding to increasing levels of cognitive load. This will allow the classifier to learn how brain wave patterns change as cognitive load increases and calibrate itself to the individual brain wave patterns of each participant. In a 0-back test (n=0), participants are required to remember a specific stimulus and identify it within a sequence of stimuli displayed on the screen. For the 1-back test (n=1), participants must recall each stimulus presented on the screen and recognize it if it appears consecutively.

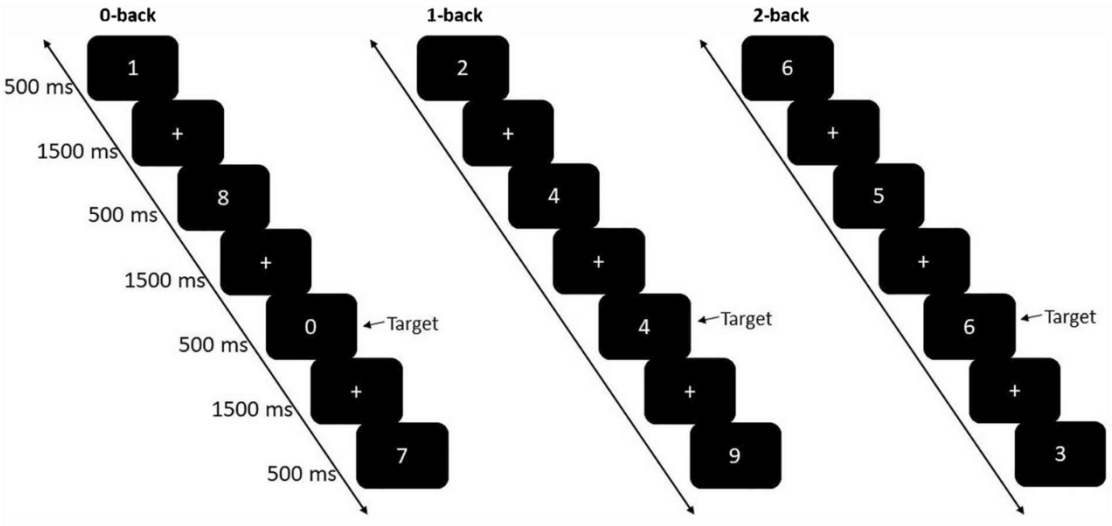


Figure 4: N-Back Task[107]

Similarly, in the 2-back test (n=2), participants need to remember stimuli to ascertain whether the current stimulus matches the one presented two steps prior. The 2-back test includes 3 conflict trials, which were described in chapter 5.1.1, for which no significant response should be noticed.

6.2.2 Arithmetic Task Protocol

To design this part of the experiment, a total of 7 similar experiments were thoroughly studied and taken into consideration.

The experimental design entails participants engaging in arithmetic tasks of varying difficulty levels, determined by the Q-value[59], under time constraints to replicate real-world cognitive stressors. The conditions are categorized as follows: low cognitive load with a Q-value below 2, medium cognitive load with a Q-value ranging from 2 to 4, and high cognitive load with a Q-value exceeding 4. Data collection will be conducted via EEG signals using the ActiveTwo EEG system by Biosemi while participants perform the tasks.

Q-value is explained briefly below[60]:

Q-value, introduced by Thomas in 1963, is a measure used to assess the difficulty or informational complexity of an arithmetic task. It evaluates factors such as the size of individual digits and the presence of carry-over operations. For an arithmetic problem $x_i + y_i$ + y, where x_1 and y_1 are the units digits, x_2 and y_2 are the tens digits, and so on, the Q-value is calculated using the following method:

For single-digit problems that do not require carry-over:

$$Q(x_1 + y_1) = \log[x_1 + y_1 + (x_1 + y_1)]$$

For single-digit problems that require carry-over:

$$Q(x_1 + y_1) = \log[x_1 + y_1 + (x_1 + y_1) + 10 + (x_1 + y_1 - 10)]$$

For multi-digit problems an additional parameter c = 1 indicates the carry-over from one position to the next: For the example problem 27 + 49, the Q-value equals:

$$\begin{gathered} Q(x_1x_2 + y_1y_2) = Q(x_1 + y_1) + Q(x_2 + y_2 + c) \\ = \log[x_1 + y_1 + (x_1 + y_1) + 10 + (x_1 + y_1 - 10)] \\ + \log[x_2 + y_2 + (x_2 + y_2) + c_2 + (x_2 + y_2 + c_2)] \end{gathered}$$

6.3 Experiment Procedure

6.3.1 Pre-experiment Setup

Before starting the experiment, participants were informed about the purpose of the study and the experiment procedure and filled out a consent form ensuring confidentiality, their ability to withdraw from the experiment at any time without penalty, the risks and benefits of the procedure, and contact information. Then, participants filled in a pre-experiment questionnaire, which included demographic questions, as well as lifestyle factors and task familiarity. After ensuring the participants were well-informed, the EEG device setup began.



Figure 5: Electrode Setup

The subject was sat in the experiment room, the cup was placed and the standard procedure started by placing the electrodes on the scalp. After all electrodes were placed, their conductivity had to be checked and which ones seemed to introduce noise, by ensuring that the right amount of gel was used so that all conducted correctly.

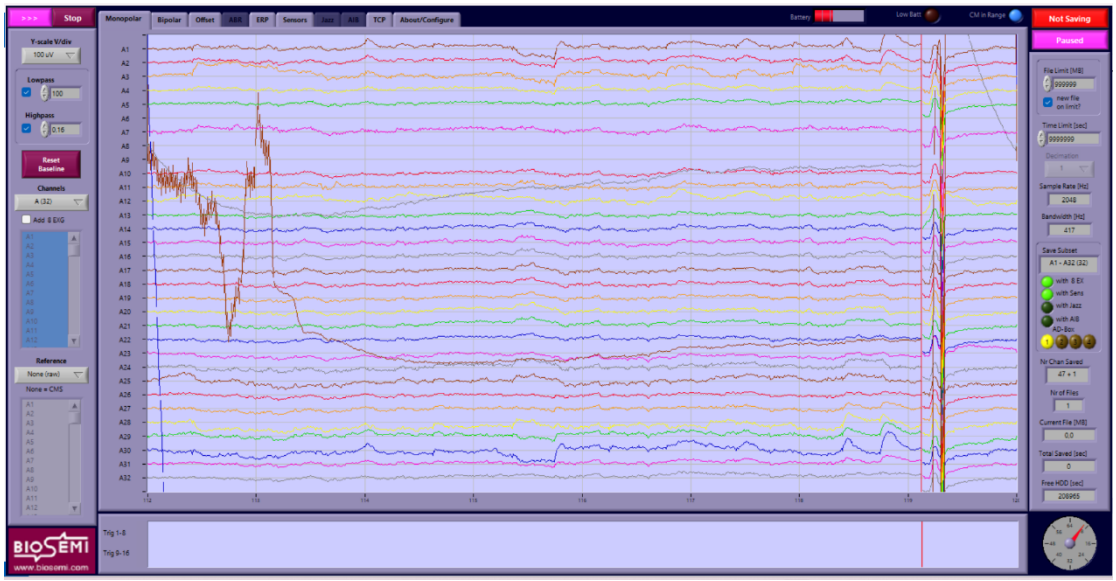


Figure 6: Actiview software indicating misplacement

For example, in Figure 6, the Actiview software showed noise or misplacement for channels A6 and A9, which was verified after checking the channels' offset in Figure 7, and they had to be re-adjusted, until the signals appeared as in Figure 8. Also, each channel's offset was

checked and channels that exceed 50Mv were again re-adjusted, to eliminate noisy channels. After that, a brief training started. The brief training included a small n-back task with 4 stimuli for each of the three levels and a small arithmetic task with 3 additions per difficulty. During that time there was no recording and based on the subjects' results additional instructions and clarification were given to ensure correct comprehension. The participants were also informed to avoid swallowing, moving, and focusing their gaze on anything other than the stimuli.

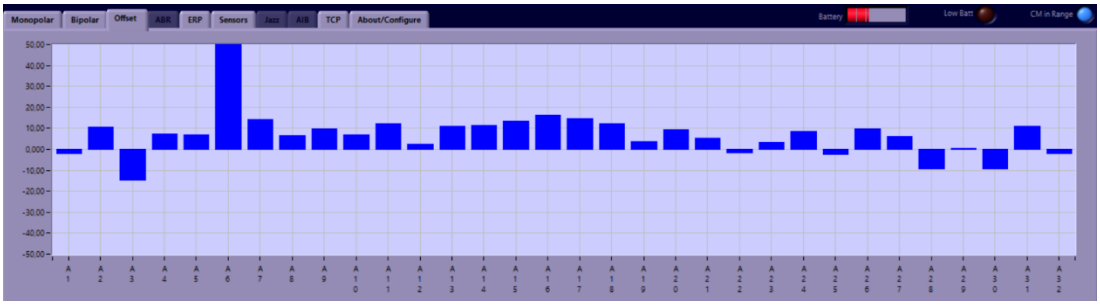


Figure 7: Channels' offset

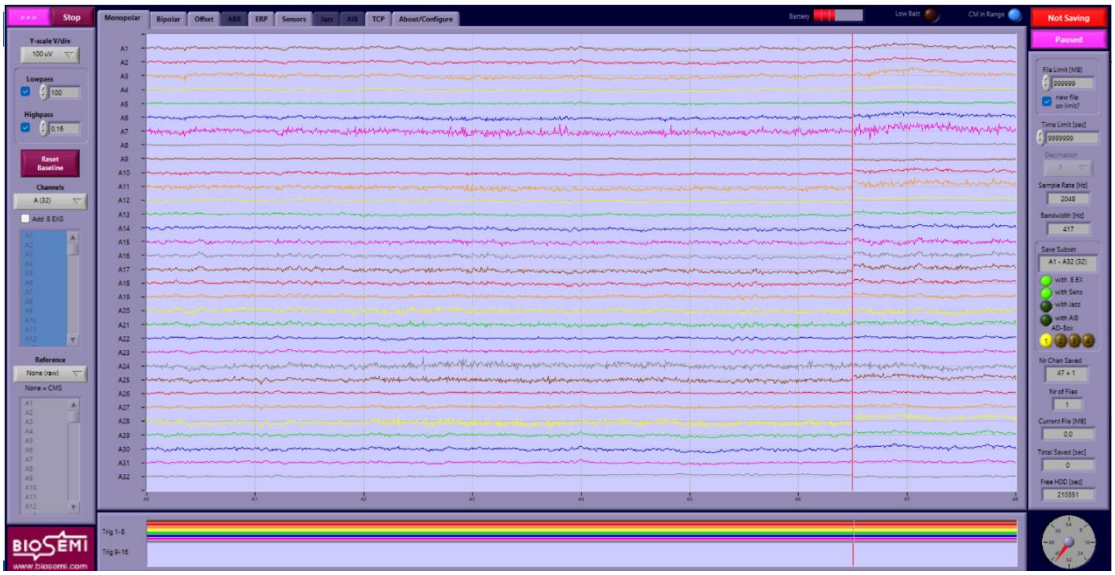


Figure 8: Normal EEG signals in Actiview Software



Figure 9: Subject isolated and monitored

Then the EEG recording started. The subject was instructed to focus on a cross that appeared on the screen so that the resting brain waves could be recorded. This lasted for 2 minutes, following the procedure used in the n-back task dataset. During this part, the subject was isolated in the room with the lights off, and the procedure was monitored from another room, separated by a glass window, as seen in Figure 9.

6.3.2 Experiment Phase – N-Back Task

Each participant then completed one test session that consisted of three phases with varying cognitive loads: low, medium, and high. In each phase, after participants were informed about the instructions corresponding to the certain N level, they encountered 20 visual stimuli and determined if each stimulus matched the one presented zero, one, or two steps prior, depending on the difficulty level. Participants were instructed to press the space bar on their keyboard when they recognized the stimuli displayed on their screen. A short break of 1.5 minutes was included between phases to facilitate recovery. The entire procedure took 10.5 minutes, which included the final break and consisted of 20 stimuli per difficulty resulting in 60 total stimuli.

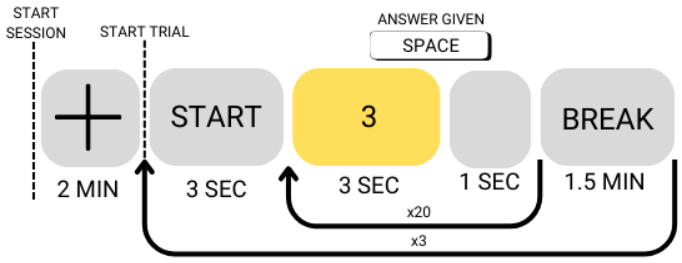


Figure 10: N-Back Task Process

6.3.3 Experiment Phase – Arithmetic Task

When participants felt ready, they pressed the button to start the next phase. Each participant completed two test sessions, each consisting of three phases with 15 addition tasks per phase. The phases differed in cognitive load, categorized as low, medium, or high. The sequence of these phases was low to high, to minimize participant fatigue and reduce error likelihood. Participants were instructed to press the arrow keys on their keyboard to indicate their agreement with either the left or right option. Scheduled breaks between tasks were incorporated to prevent mental fatigue. The whole procedure lasted 25.55 minutes and consisted of 15 stimuli per difficulty per session, resulting in 135 total stimuli.

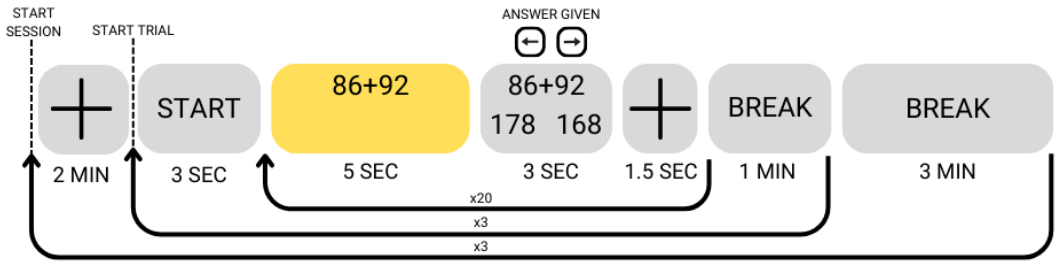


Figure 11: Arithmetic Task Process

On the first test session of the arithmetic task, an additional larger break was introduced between each phase for the participants to fill in a questionnaire about the mental demands of the task. That was the NASA Task Load Index questionnaire, which assesses mental workload based on a weighted average using 6 sub-scales (Mental Demand, Physical Demand, Temporal Demand, Own Performance, Effort, and Frustration)[103]. This resulted in three total questionnaires being filled, one for each difficulty level. A question concerning the participants' opinion on whether the second task reflected real-world problem-solving situations was added to assess the task validity.

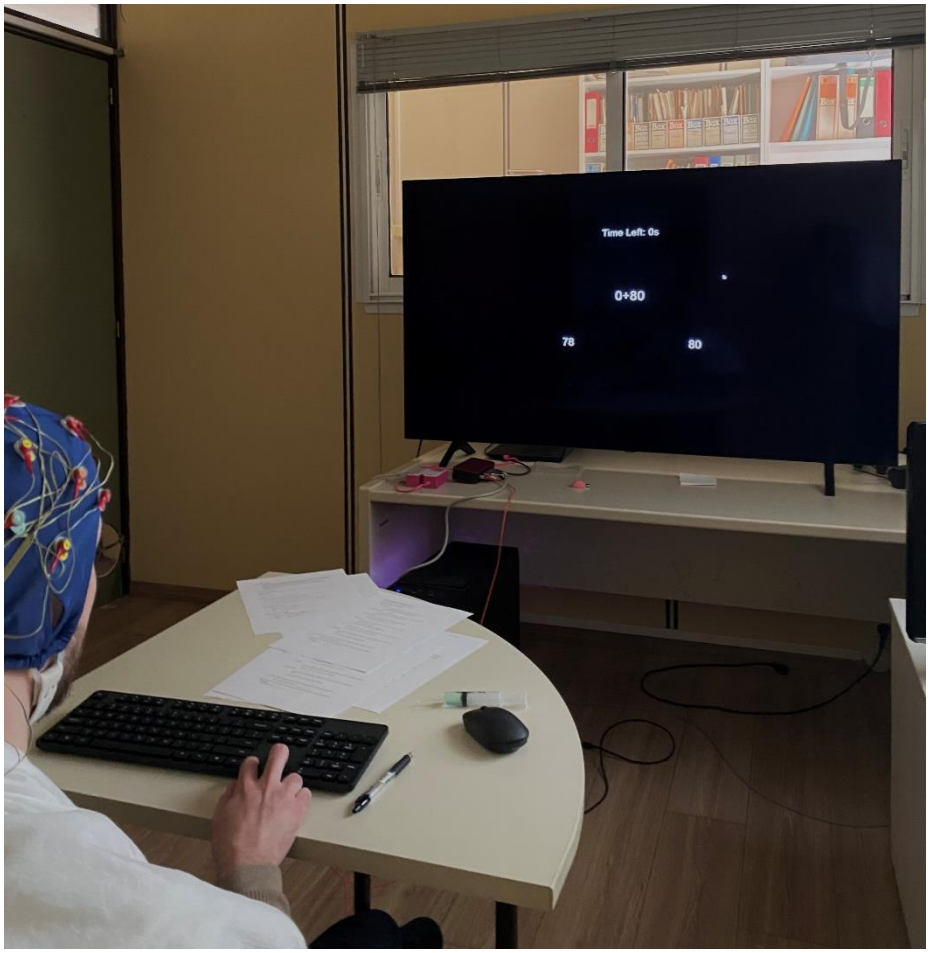


Figure 12: Subject during the arithmetic task

6.3.4 Triggers during the experiment phase

As mentioned earlier, during the signal acquisition event triggers were sent to the BioSemi ActiveTwo USB receiver to mark task-relevant events, for participant 1 to 3, and after a hardware malfunction it was replaced with NEUROSPEC'S MMBT trigger interphase box. To ensure accurate event marking throughout both the N-back and arithmetic tasks, a structured trigger system closely modeled on the logic from the previously acquired dataset was utilized. During the experiment, triggers were sent at specific moments: when the onscreen instructions began and ended, at the start and end of the fixation cross period, at the onset of each problem for every difficulty level, upon the participant's answer submission, and finally when the experiment concluded. In addition, correctness or missed responses were logged with unique codes (e.g., `ANSWER_CORR_1`, `ANSWER_WR_2`). Triggers are noted in Table 1. All triggers appeared in the BioSemi ActiveTwo "Status" channel at the same sampling rate as the EEG

signals, enabling precise alignment of stimulus presentation, participant interaction, and brain activity for subsequent analysis.

Trigger	Code
INSTRUCTION_START	10
INSTRUCTION_END	11
FIXATION_START	20
FIXATION_END	21
PROBLEM_PRESENTED_0	31
PROBLEM_PRESENTED_1	32
PROBLEM_PRESENTED_2	33
ANSWER_PERIOD	40
ANSWER_SUBMITTED	50
ANSWER_CORR_0	51
ANSWER_CORR_1	52
ANSWER_CORR_2	53
ANSWER_WR_0	54
ANSWER_WR_1	55

Table 1: Trigger Codes

7. Results and Analysis

Out of the 10 participants in this experiment, subject 9 will be excluded due to excessive noise in the data, possibly due to misplacement of electrodes. Subject 2 had to be removed from the experiment due to the failure of the USB trigger box, resulting in no recorded triggers in the EEG. Subject 3 reported having slept only 4 hours and performed poorly on all test measures, indicating a lack of focus. Thus, their data was not considered to be useful to the study results.

7.1 Perfomance Results

7.1.1 N-back Task

The N-back results show a marked a clear trend in performance across increasing memory loads. Table 2 shows performance across participants in the N-Back task. All of the subjects in the 0-back and 1-back conditions were almost equally accurate, with most achieving near-perfect scores and an average of 100% for the 0-back session and 97% for the 1-back, indicating that both the tasks with lower cognitive complexity are relatively easy to perform. The 2-back condition was, however, drastically different in performance: one subject even dropped to 50% accuracy while others varied from 60% to even 100%, with average performance dropping to 74%. This certainly goes with the declining performance with greater task difficulty as a higher working memory load means a heavier cognitive burden and increased probability of error.

Success Rate - N-Back Task			
Subject	**0-back**	**1-back**	**2-back**
1	100%	95%	50%
2	excluded	excluded	excluded
3	90%(excluded)	30%(excluded)	0%(excluded)
4	100%	90%	70%
5	100%	100%	95%
6	100%	100%	100%
7	100%	100%	60%
8	100%	100%	80%
9	excluded	excluded	excluded
10	100%	95%	60%
Average	100%	97%	74%

Table 2: N-Back Task Success Rates

7.1.2 Arithmetic Task

The arithmetic task results, shown in table 3, show that performance varies regarding difficulty and between phases. In phase 1, participants were generally successful in the low-difficulty condition; thereafter, accuracy fell as the difficulty of the task increased. Importantly, some of the lower accuracy scores in that phase can be attributed to errors that are not cognitive load related- for example, some subjects reported that they pressed the wrong key by mistake although they knew the correct answer. Also, having been exposed to the task for the first time, there could have been an adaptation period in early performance.

Success Rate - Arithmetic Task - First Session			
Subject	Low Difficulty	Medium Difficulty	High Difficulty
1	100%	95%	50%
2	excluded	excluded	excluded
3	85%(excluded)	65%(excluded)	45%(excluded)
4	100%	95%	90%
5	100%	100%	95%
6	100%	100%	100%
7	100%	95%	75%
8	100%	95%	70%
9	excluded	excluded	Excluded
10	100%	100%	85%
Average	100%	97%	81%

Table 3: Arithmetic Task First Session Success Rates

Phase 2 showed a declining trend in accuracy for most participants, as shown in Table 4, particularly in the high-difficulty task, which implied that mental fatigue affected performance at this level. However, there was an improvement in the medium-difficulty level for subject one, possibly attributable to the fact that the subjects had had time to familiarize themselves with the task structure, as they noted 50% accuracy in the first session and 95% in the second one. All the findings indicate, therefore, the effect of cognitive load and time on arithmetic task performance.

Success Rate - Arithmetic Task - Second Session			
Subject	Low Difficulty	Medium Difficulty	High Difficulty
1	95%	90%	95%
2	excluded	excluded	excluded
3	80%(excluded)	60%(excluded)	45%(excluded)
4	100%	90%	90%
5	100%	100%	85%
6	100%	100%	100%
7	100%	95%	70%
8	100%	90%	70%
9	excluded	excluded	excluded
10	100%	95%	60%
Average	99%	94%	85%

Table 4: Arithmetic Task Second Session Success Rates

Small differences in performance were observed between the N-back and arithmetic tasks, with the arithmetic task almost the same success rates overall. This suggests that the arithmetic task had similar mental demand, requiring equal working memory and problem-solving effort. Disparities in task difficulty and cognitive demands could pose a challenge for a classifier trained to recognize cognitive states across both tasks, as the patterns of neural activity may differ between them, so in this case we should not expect that.

7.2 NASA Task Load Index Scores

For the arithmetic task, participants filled in a NASA-TLX questionnaire at each difficulty level, so that subjective perception of their performance was assessed. In doing so, participants

indicated their self-evaluations for various aspects, including mental demand, effort, and frustration, which provided qualitative data for understanding variance. Subjective ratings were compared against objective performance data to better understand cognitive load differences or inconsistencies between objective accuracy and the perceived difficulty in such tasks. The raw scores were calculated, meaning that the subjects did not rate the dimensions of the questionnaire based on their relativity, as it is quite time-consuming.

The NASA-TLX results, presented as raw scores in Table 5, show significant individual differences in perceived mental workload, indicating that each subject has a unique threshold for what they consider cognitively demanding. For instance, the average success rate in the high-difficulty condition is 81%, yet the corresponding NASA-TLX rating is 56%, indicating that subjects perceived the task as mentally demanding despite maintaining relatively high accuracy. Subject 1 reported a high workload (86%) while experiencing a significant drop in accuracy (50%), suggesting that they found the task highly demanding. In contrast. Subject 6 achieved perfect accuracy (100%) but still reported a relatively high workload (56%), suggesting that they found the task demanding even though they performed flawlessly. In contrast, Subject 8 had a lower accuracy (70%) but reported a much lower workload (26%), implying that they did not perceive the task as particularly taxing despite struggling more than Subject 6. Despite this variability, the overall trend remains consistent: higher difficulty levels lead to increased workload ratings, confirming the difficulty scaling. This pattern is crucial for validating cognitive state classification, as tasks with higher perceived difficulty may introduce greater variability in EEG features.

	NASA TLX Rating		
Subject	**Low Difficulty**	**Medium Difficulty**	**High Difficulty**
1	25%	86%	86%
2	excluded	excluded	excluded
3	excluded	excluded	excluded
4	15%	26%	66%
5	13%	16%	23%
6	26%	38%	56%
7	13%	32%	61%
8	10%	18%	26%
9	excluded	excluded	excluded
10	20%	36%	71%
Average	17%	36%	56%

Table 5: NASA TLX Ratings

Two extra questions were added to asses the difficulty of the task. The first asked the participants whether they found it difficult to stay focused, where almost all answered no except for the one that was excluded and some that noted that the resting state recording for 2 minutes was tiring. Also, the participants were asked if they believed the arithmetic task reflected real-world problems, where 5 out of 7 said yes and the other 2 said no as it was too simple.

7.3 Preprocessing Results

7.3.1 Preprocessing N-back dataset

The procedure described in the previous chapter was applied to all recorded EEG data. For reference, the preprocessing steps are illustrated using data from Subject 3. Before preprocessing, the EEG signals exhibited strong artifacts, including characteristic spikes at 50 Hz, 100 Hz, and 150 Hz, which are AC harmonics caused by electrical interference, as seen in Figure 13 which represents the power spectrum for the subject's 3 EEG recordings.

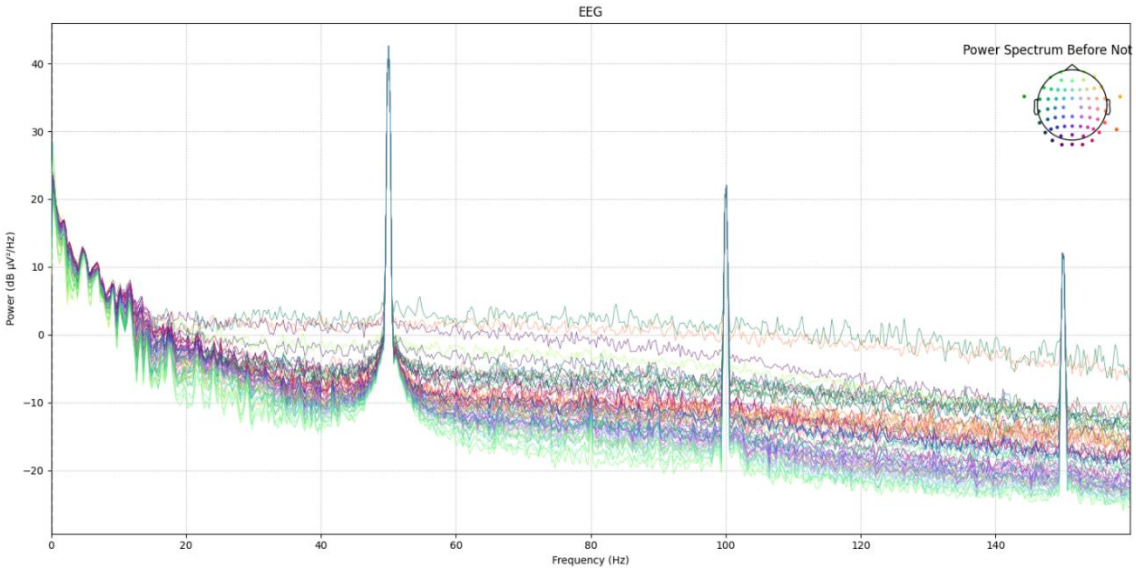


Figure 13: Power Spectrum Before Filtering

These artifacts had to be filtered out using a notch filter. Following this, high-frequency noise beyond the 30-Hz limit was generally caused by muscle activity and external noise and was thus inappropriate for analysis. A band-pass filter was applied to pass only frequencies from 1 to 30 Hz, capture preserving relevant neural activity.

After finding bad channels and interpolating them, Independent Component Analysis (ICA) was performed to identify and remove stereotypical artifacts such as eye blinks and eye movements. The number of ICA components varied across subjects due to differences in signal characteristics and noise levels. In Figure 14, corresponding to the analysis for Subject 3, components 2 and 4 were identified as representing eye blinks, while component 9 corresponded to lateral eye movements. These components were removed before reconstructing the EEG signals. The ICA decomposition was validated visually and by inspecting the power spectra of the components to confirm their association with non-neural artifacts.

The EEG data after removing the artifacts, was epoched around the experimental events, specifically the event corresponding to stimuli. However, some epochs contained a lot of noise and were rejected from the analysis to maintain data quality. For each subject, a dropped channels diagram was made to check how many epochs had been dropped and which channels caused the drop. For subject 3, Figure 15 shows that 9.2% of the epochs were dropped and it was due to the electrodes placed near the eyes, which introduced bad noise.

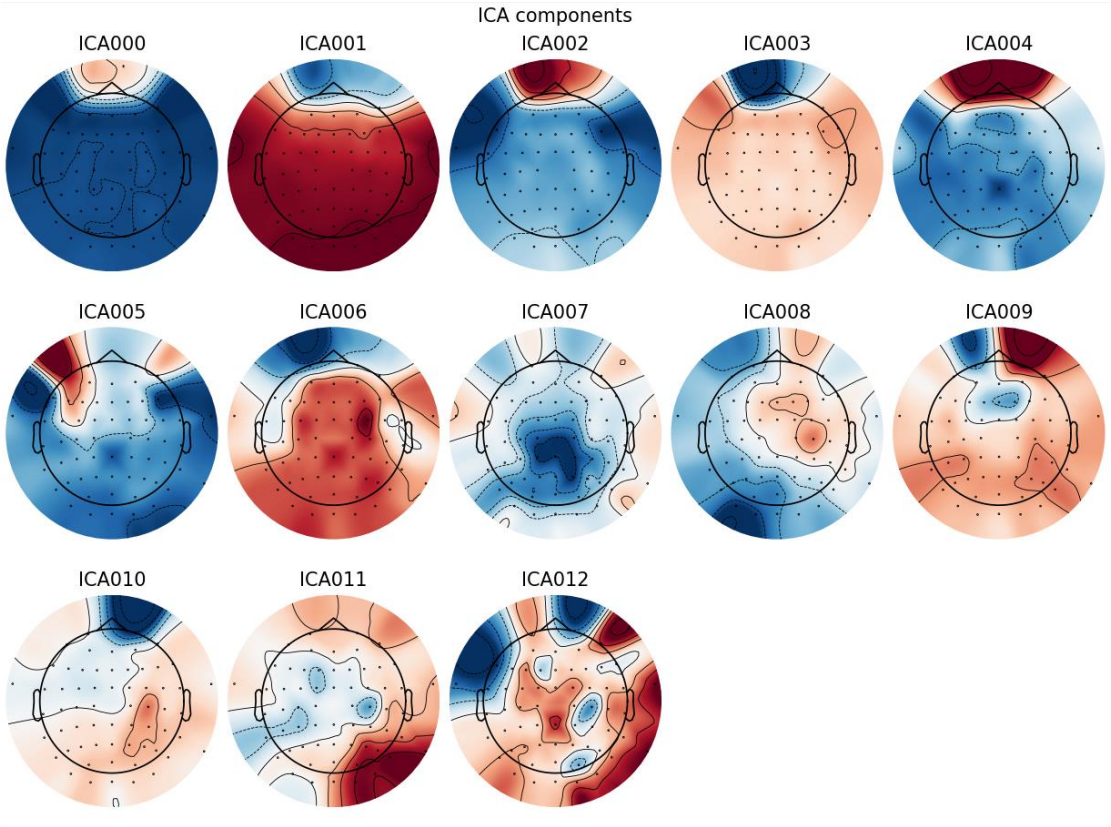


Figure 14: ICA Components for Subject 3



Figure 15: Dropped Epochs for Subject 3

The number of rejected epochs varied across subjects. The majority of epochs were rejected due to the two channels located near the eyes (Fp1, Fp2), where ocular artifacts were predominant most of the time, even after artifact removal. Figure 16 provides a summary of the channels contributing most to epoch rejection, where it becomes evident that frontal electrodes, especially the ones closer to the eyes, caused most contamination.

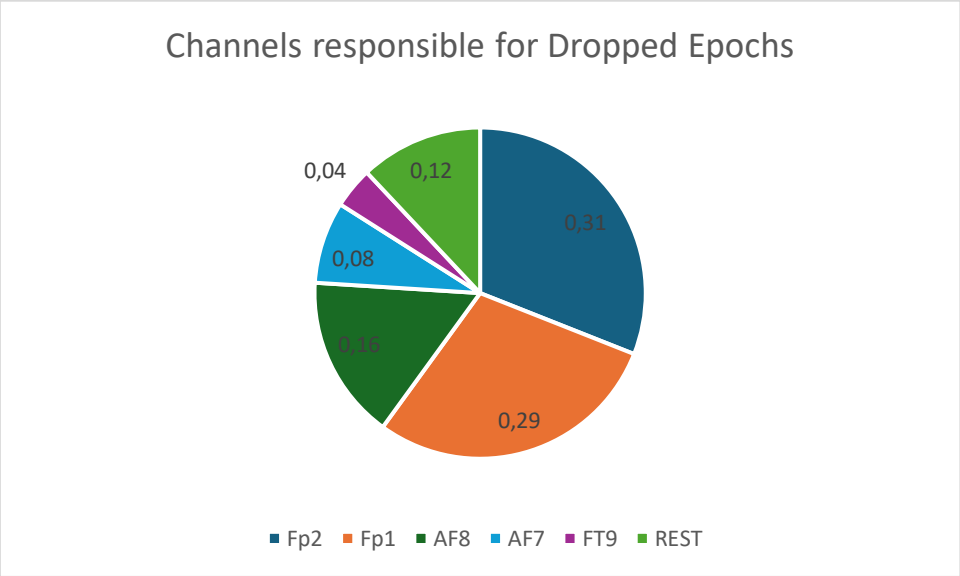


Figure 16: Responsible Channels for Dropped Epochs

Despite these efforts, some subjects exhibited extreme levels of noise, leading to a substantial loss of data. Specifically, Subjects 2, 12, 14, 20, 22, 26, 28, and 29 had more than 40% of their epochs rejected due to excessive noise. Given the extent of data loss, these subjects were excluded entirely from the dataset, as their remaining EEG signals were insufficient for reliable analysis. This exclusion was necessary to maintain the integrity of the dataset and ensure that subsequent machine-learning models were trained on clean, high-quality signals.

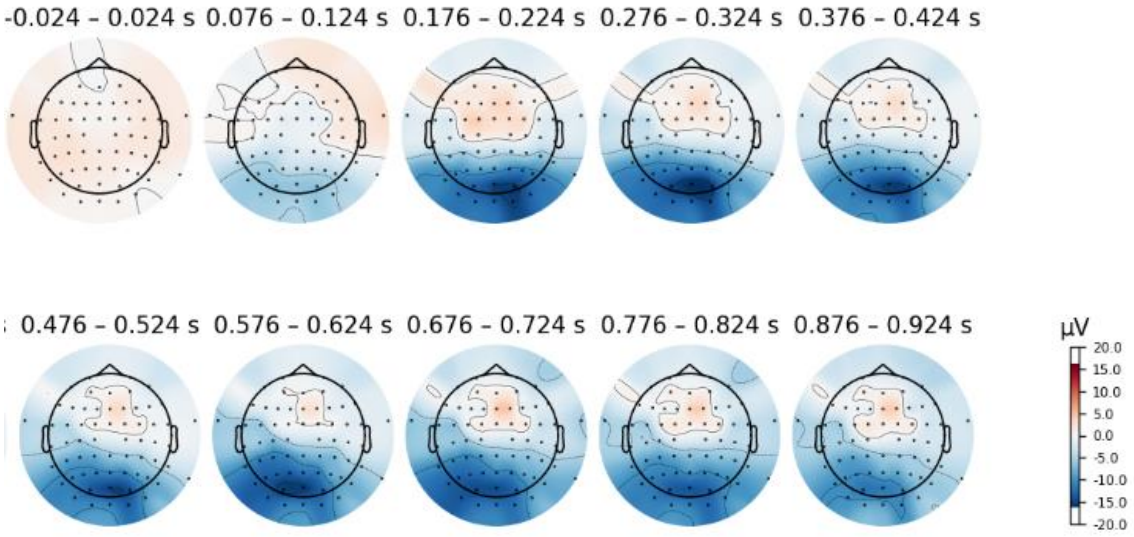


Figure 17: Activity after Visual Stimuli

Figure 17 represents the activity captured by the electrodes in the epoch the second after visual stimuli is presented to the subject, during the 2-back test. After 15ms, activity is noted in the frontal lobe, as the subject is making mental effort to remember the stimuli as well as the previous ones.

7.3.2 Feature Selection Results

For feature selection, the only channels that were picked were the ones corresponding to a 32-channel EEG out of the initial 64 channels, as the experiment phase utilized a 32-channel equipment. All features mentioned above were extracted and then passed on a random forest classifier to evaluate their importance. The most important feature was band power and power spectral density, followed by discrete wavelet transform energy and entropy. The least important features were Hjorth activity and Lempel-Ziv. For subject 3, Figure 18 shows the mean beta frequencies captured during all epochs from resting state, zero back task and two back task. The figure shows changes in beta frequency activity recorded via EEG during different cognitive task conditions. In the resting state, beta activity in the frontal lobe appears at a moderate level. As the task difficulty increases, a noticeable rise in beta activity is observed, with higher levels recorded during the easy task and an even more pronounced activation during the hard task. This observation shows a direct correspondence between cognitive effort and frontal engagement since it seems to demand more and more processing resources as task complexity increases.

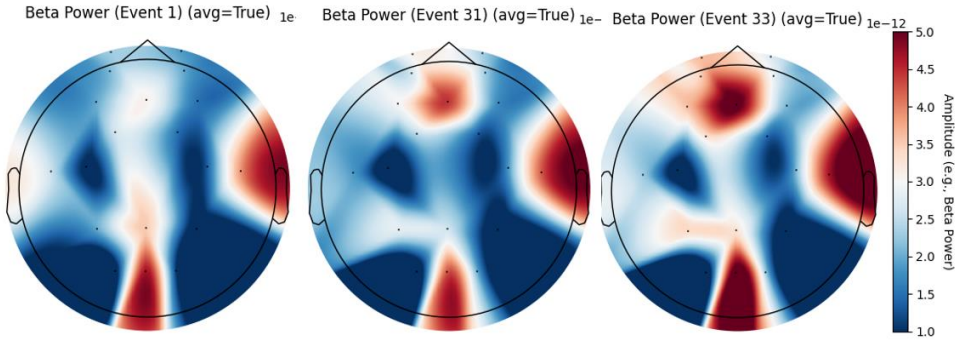


Figure 18: Beta frequencies in Resting State (Left), Low Difficulty Task (Middle) and High Difficulty Task (Right)

7.3.3 Preprocessing the experiment's data

The same procedure was followed for the experiment's data, and no subjects had more than 10% of their epochs dropped, which indicates better quality signals and low noise. Each subject's ICA composition was carefully examined, to ensure noise was identified, and then all data was epoched based on the triggers produced by the MMBT trigger box.

7.4 Classification Performance on N-Back Dataset

7.4.1 Electrode Selection

Research Question B: What specific channels offer the most valuable data? Would reducing them be possible?

Before feature extraction, the right electrodes had to be chosen for both the initial dataset and the experiment data. In the first approach, all 32 channels were selected. This resulted in

a very big and complex dataset, so training the SVM classifier was quite time-consuming. The resulting average accuracy across participants was 30% when training a 4-class model and 61% when training a 2-class model. The second approach only utilized electrodes corresponding to the frontal lobe, so frontal and front central electrodes. This reduced the size of the data by more than 60% and the time needed to train the classifier, but the accuracy of the resulting classifier was not sufficient. For the 4-class model, average accuracy across participants was 26%, surpassing random guess accuracy by only 1%, and for the 2-class model only 52%, again resulting in a not applicable classifier. For the last approach, electrodes from all frontal channels were retained, and from other regions, only midline or key lateral channels were included: C3, C4, T7, T8, Pz, P3, P4, Oz, O1, and O2. This increased the size of the data but still was less than the first approach, and the trained classifier achieved higher accuracy. The 4-class model achieved 33% average accuracy while the 2-class model achieved 63%.

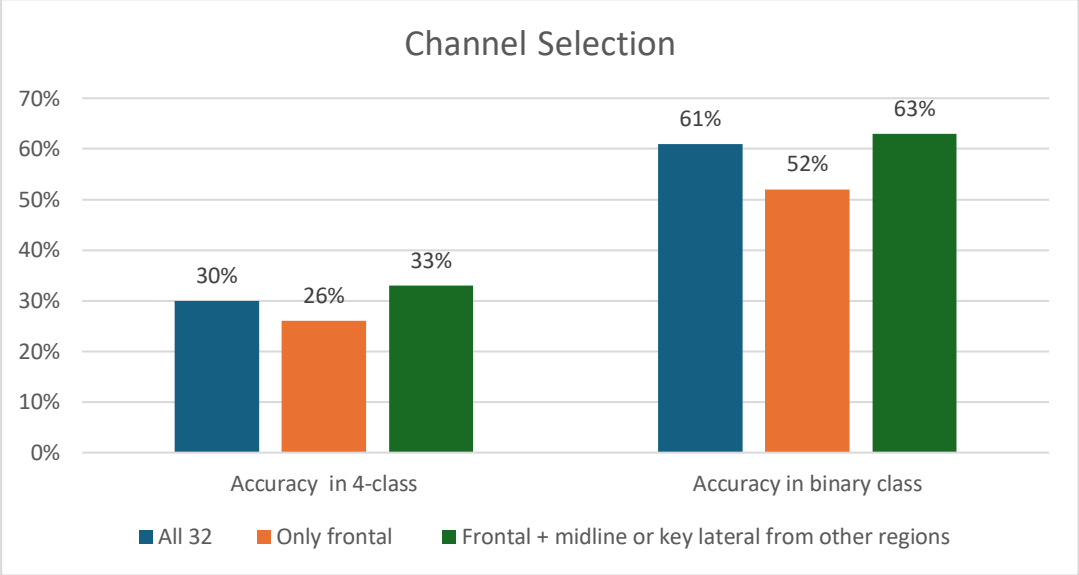


Figure 19: Accuracy after Channel Selection

This suggests that not all electrodes contribute equally to classification, and selecting the most relevant channels—particularly those combining frontal, midline, and lateral regions—enhances model performance. These findings imply that targeted electrode selection can improve classification accuracy by reducing noise and focusing on the most informative signals, rather than relying on all available data. This also suggests that certain midline and lateral electrodes provide crucial information about cognitive load beyond what frontal regions alone can capture.

7.4.2 All Features vs Band Power

To achieve the highest possible accuracy, two separate classifiers were built using the electrodes chosen in the previous chapter, using different sets of features. In the first approach, the best features defined by the Random Forest classifier were used. For the 4-class classifier, accuracy reached an average of 34%, and for the 2-class 64% across participants. In the second approach, only the feature concerning band power was used, which resulted in fewer data for training but achieved similar accuracy, with 33% for 4 classes and 63% for 2 classes. These results suggest that the classifier can achieve comparable accuracy even with a reduced feature set focused on band power, indicating that band power alone captures essential cognitive load information. This also implies that while Random Forest-selected features provide a slight

advantage, a more compact and interpretable feature set can still yield competitive performance. The remaining analysis focuses only on the classifier trained with band power features.

7.4.3 Performance Analysis

Research Question A: Would a classifier trained to distinguish cognitive states using data from a specific working memory task, maintain its accuracy when tested on cross-participant data? Can it generalize to new subjects without prior training on their individual EEG signals?
Results indicate that the classifier's accuracy varies. The accuracy results for a new subject in binary and multi-class models reveal significant inter-subject variability, highlighting the challenges of generalizing across different individuals. In the binary classification model, accuracy ranges from 50% to 74%, with all subjects achieving performance above 50%, suggesting that the model is learning meaningful features rather than relying on chance (which would be 50% for a binary task). An average accuracy of 63% across subjects is achieved, which is in line with current research.

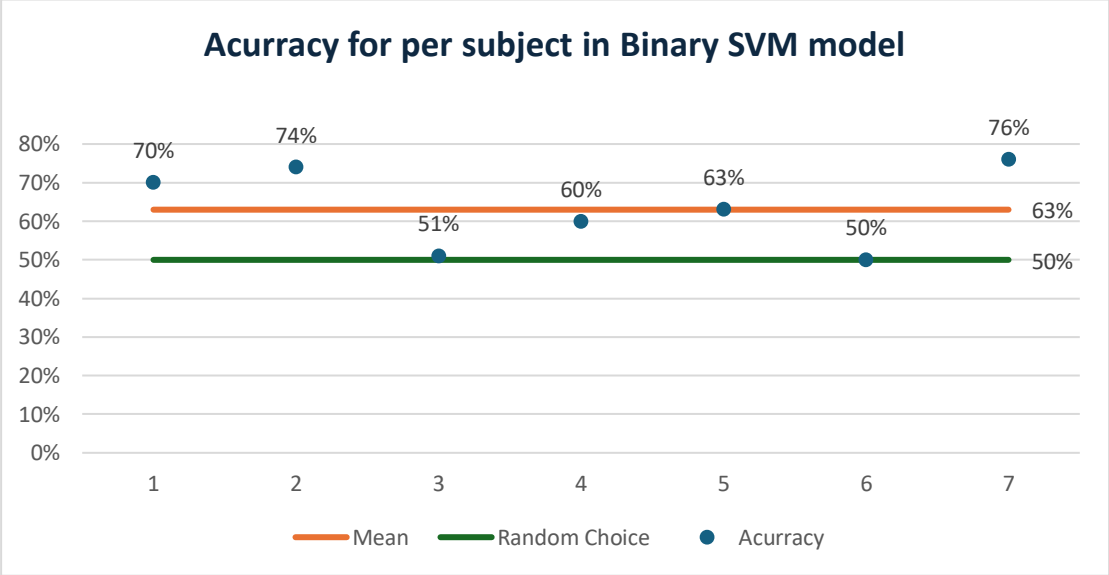


Figure 20: Per Subject Accuracy in Binary SVM

In the multi-class model, where classification is inherently more difficult due to the presence of four categories, accuracy fluctuates between 24% and 48%. The lower values for some subjects suggest that the classifier struggles to differentiate between classes, likely due to high intra-class variability, noisy EEG signals even after pre-processing, or differences in cognitive processing across individuals. Although some subjects achieve performance close to 33%, which is the average in this case and above the random baseline of 25%, others remain closer to chance level, reinforcing the difficulty of multi-class classification in EEG-based cognitive load estimation. These findings emphasize that while binary classification remains relatively stable, multi-class classification is more subject-dependent and poses greater challenges in terms of generalization.

To address the challenges identified during the evaluation phase and enhance the model's performance, the entire pipeline was systematically modified and refined.

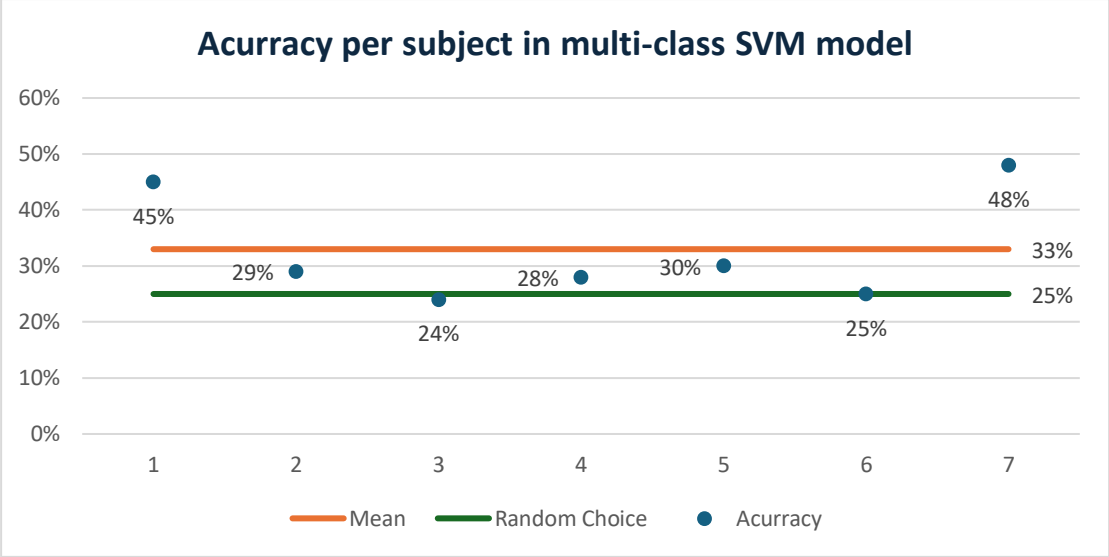


Figure 21: Per Subject Accuracy in Multi-class SVM

These modifications targeted both the data preprocessing steps and the machine learning framework to mitigate overfitting and improve generalization across unseen participants. After these modifications, results reached accuracy that indicated overfitting was reduced. The final SVM algorithm used a polynomial with a 3 degree, a regularization parameter C=100, and gamma set to auto.

7.5 Generalization to the Designed Experiment

7.5.1 Initial Classifier Performance

Research Question C: Would a classifier trained to distinguish cognitive states using data from a specific working memory task, be able to maintain its accuracy when tested on different task data? What if this task is not a scientifically validated task and differs notably from the one used in training? What if the EEG hardware in each case differs?

Both the binary and the multiclass classifier, trained with data from the initial dataset, were used to test data from the arithmetic phase of the experiment. When the old classifier—originally trained on N-back task data—was tested on new arithmetic task data from the experiment conducted in our case, its accuracy fell considerably. This occurred not only when using band power as features, but also when using the features defined by the Random Forest classifier as best. Although both datasets were meant to measure cognitive load, the experimental circumstances surrounding data collection varied in ways that directly impacted the classifier's performance. Differences in hardware setup, such as electrode placement or amplifier quality, and even subtle discrepancies in sampling rate or signal preprocessing can alter the characteristics of the EEG data. Likewise, environmental factors—like noise levels, lab conditions, or participant comfort—all introduce additional variability that can cause a model to fail when confronted with data recorded under different conditions.

Multi-class model

The four-class model achieved 26% accuracy, largely due to classification complexity and model bias. In the four-class setup (Resting State, Low, Medium and High Load), the model must differentiate between multiple cognitive states, each with overlapping EEG patterns, making the classification problem significantly harder. Since random guessing would yield 25% accuracy, the model's performance at 26% suggests that it is barely learning meaningful

distinctions, due to a high bias problem—where the classifier fails to separate classes effectively and instead collapses most predictions into one dominant class. This is obvious after inspecting Figure 22, which indicates per-subject accuracy, where it is obvious that almost all subjects achieve accuracy equal to random choice, except for one that is higher but could be random.

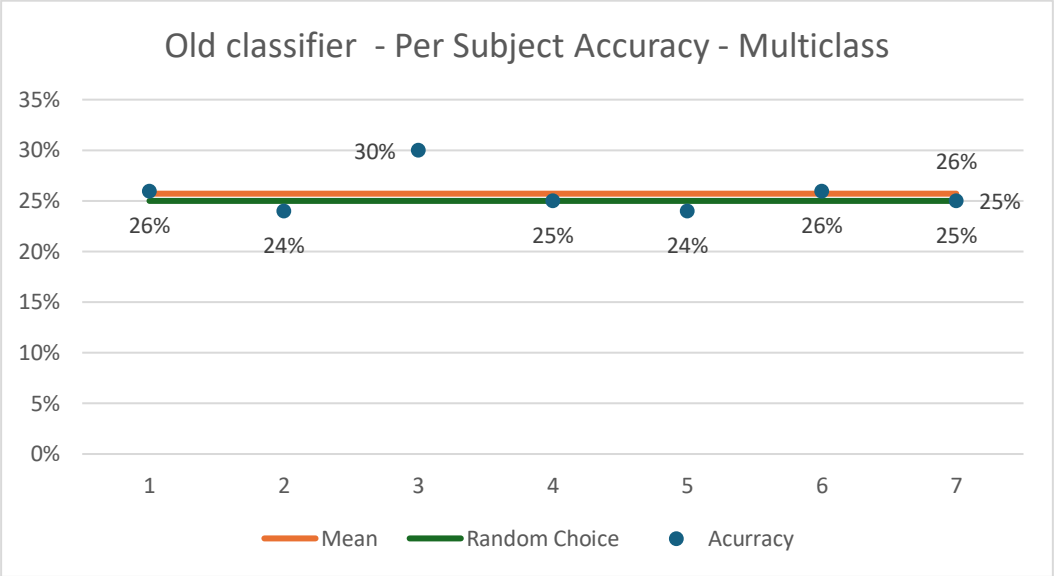


Figure 22: Old Classifier - Per Subject Accuracy - Multi-class

The confusion matrix and the classification report summarize the testing results in Figure 23 and Table 6. The multiclass model considers almost all data Low Load, achieving a score of 100% in Low Load Recall, but 0% in Resting State and High Load, and 5% Recall in Medium Load which could be random. This often happens when a model, struggling to extract robust features, defaults to assigning most samples to the most frequent or easiest-to-predict category. This overfitting resulted in an overall weak model, as seen in the confusion matrix, where 95% of Medium Load samples and 100% of High Load samples were misclassified as Low Load.

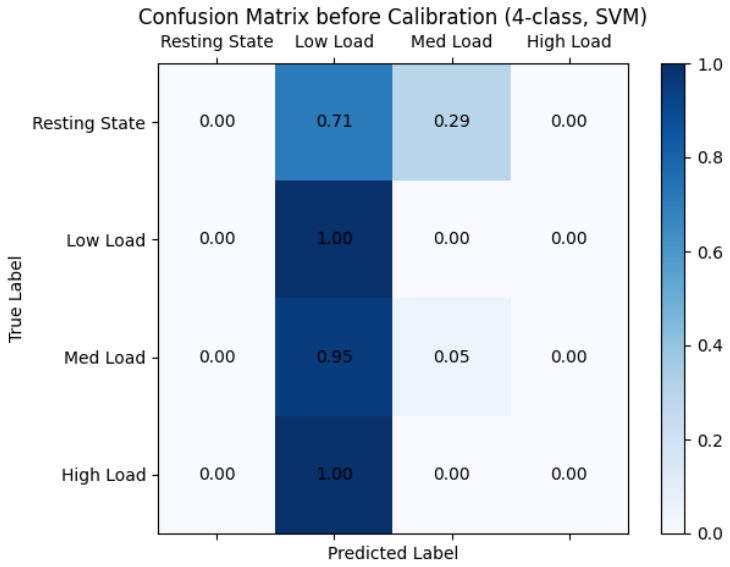


Figure 23: Confusion Matrix - Old Classifier – Multiclass

State	Precision	Recall	F1-score
Resting	0.00	0.00	0.00
Low Load	0.27	1.00	0.21
Medium Load	0.14	0.05	0.04
High Load	0.00	0.00	0.00
Average	0.10	0.26	0.06

Table 6: Perfomance Matrix - Old Classifier - Multiclass

Binary Model

Similarly, the binary model with only Resting State and High Load, so a simpler decision boundary (only two classes), achieves 50% and still suggests a similar bias issue, where the classifier is heavily favoring one class over the other instead of properly distinguishing between them.

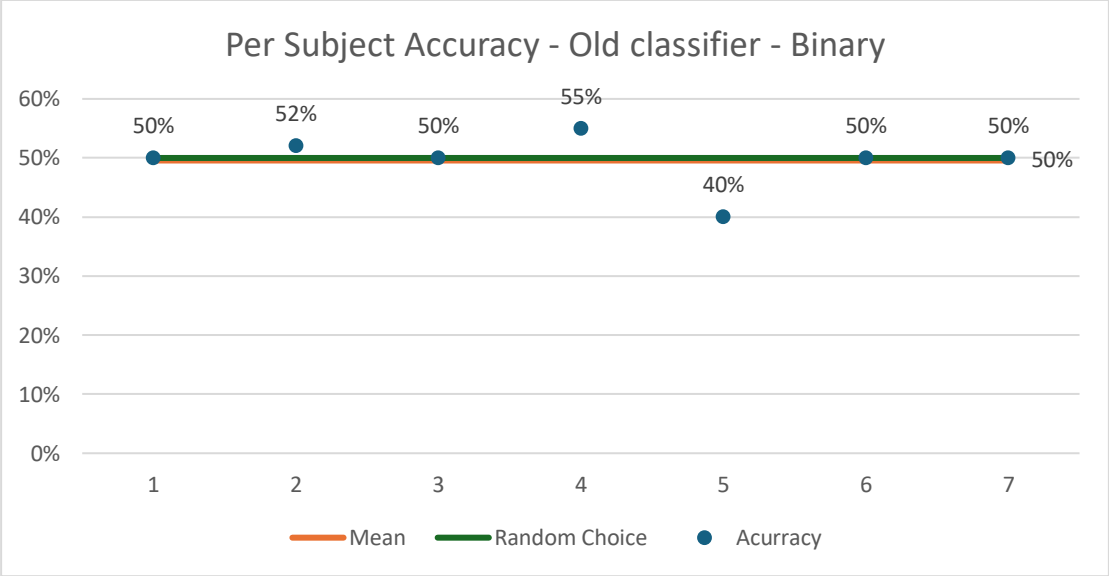


Figure 24: Per Subject Accuracy - Old Classifier - Binary

In this case, the classifier considers all data as High Load and similarly ends up with an average Recall of 50%, exactly as a purely random choice would.

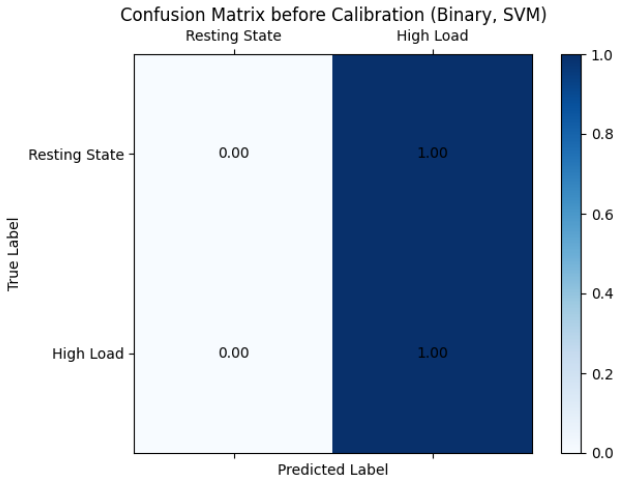


Figure 25: Confusion Matrix - Old Classifier - Binary

State	Precision	Recall	F1-score
Resting	0.00	0.00	0.00
High Load	0.50	1.00	0.66
Average	0.25	0.50	0.33

Table 7: Performance Metrics - Old Classifier - Binary

More specifically, the new arithmetic dataset has been collected with a different participant pool, unique instructions, and a different hardware configuration than the one used in the original N-back experiment. Even minor changes—such as using a newer EEG device model, altering the placement of reference electrodes, or switching to a different lab with slight electromagnetic interference—can substantially shift the underlying signal distributions. Because EEG classifiers are highly sensitive to how data are recorded and processed, the features learned from the old dataset may no longer match the statistical properties of the new signals. This phenomenon, commonly referred to as a domain shift, arises not just from task differences but also from varied recording setups and environments.

These findings highlight and prove the technical and logistical hurdles that come with building broadly applicable EEG-based classifiers. Ensuring consistency in hardware specifications, data acquisition protocols, and environmental conditions is crucial for robust, cross-study performance.

7.5.2 Training with experiment's N-Back Data

Research Question D: Can a classifier trained to distinguish cognitive states using data from a specific working memory task be adapted to maintain its accuracy when tested on different task data? Would recording participant's EEG while taking the specific working memory task and then adding it to the classifier's initial training data, help the classifier adapt to the new task and the new EEG hardware?

One of the key challenges in EEG-based cognitive load classification is ensuring that the model generalizes well across different conditions and participants. To improve the classifier's adaptability, data from the additional N-back task conducted was incorporated into the training process. EEG signals are highly sensitive to variations in recording conditions, participant differences, and task execution, which can lead to poor generalization when a model is trained on a single dataset. By introducing data from an additional N-back session, the classifier is exposed to a broader range of cognitive load patterns and each subject's individual brain waves.

Initially, EEG data were derived from both the dataset's subjects and the first phase of the experiment conducted. Data including the N-back tasks (0-back, 1-back, and 2-back) and resting state trials. For each subject taking part in the experiment, a new classifier was trained, using only their EEG data from the experiment. To ensure consistency across datasets, feature matrices were standardized to a fixed number of features (95 dimensions). Each sample was assigned a cognitive load label based on the task it was associated with, where 0-back, 1-back, and 2-back represented increasing levels of cognitive load, and resting-state trials were labeled separately. A Support Vector Machine (SVM) classifier with a polynomial kernel was selected for classification, and Standardization (Z-score normalization) was applied to the feature set to improve model stability. To address the class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was employed, ensuring that all cognitive load levels were equally represented during training.

Research Question E: For this cross-task classifier approach, what features would work best?
Again, the two sets of features were tested. By utilizing the features that the Random Forest classifier considered best, accuracy across participants averaged 28% for 4 classes and 69% for 2 classes, while it should be noted that for 2 participants, accuracy was at a random choice level. When only using band power features, accuracy increased to 30% for 4 classes and to 73% for 2 classes. These results suggest that band power features alone provide a more robust representation of cognitive load, leading to improved accuracy compared to the broader feature set selected by the Random Forest classifier. The fact that accuracy increased, despite using fewer features, indicates that irrelevant or redundant information in the full feature set may have hindered classification performance. While the improvement is small, it is still noteworthy, as it demonstrates that reducing the feature set does not compromise accuracy and may even enhance it. This makes the approach more applicable for new subjects, as it simplifies the feature extraction process while maintaining strong classification performance. For the rest of the analysis, only band power features were used in all classifiers.

Multiclass Model
Then, the features acquired for the cleaned EEG recordings during the arithmetic task of the experiment were used to test the classifier. For 4 classes (Resting State, Low, Medium, and High Load) the accuracy of the classifier increased to 30%. Classification performance improved across multiple classes, leading to a more balanced distribution of predictions. While the old classifier achieved accuracies around 25–30%, the new classifier consistently outperforms it, with subject-wise accuracies ranging from 28% to 32%. The mean accuracy of the new classifier is approximately 30%, which is notably higher than the previous model's mean of 25%.

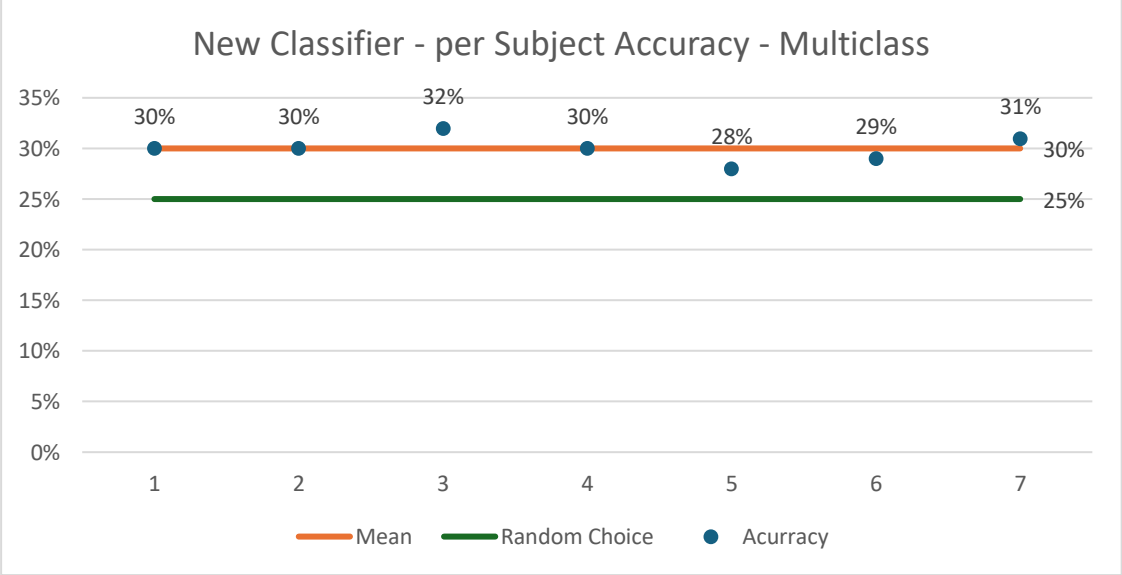


Figure 26: Per Subject Accuracy - Adapted Classifier 1 - Multiiclass

The recall for Resting State increased from 0% to 31%, indicating that the model became more capable of correctly identifying periods of rest. Similarly, Medium Load recall increased from 5% to 20%, while High Load recall improved from 0% to 15%, demonstrating that the classifier became more effective at distinguishing higher cognitive load levels. Precision also improved for Medium Load (0% to 20%) and High Load (0% to 15%), reducing the dominance of Low Load in the predictions. Despite these improvements, misclassifications remain, particularly between adjacent cognitive load levels, suggesting that further enhancements

Overall, the tuning process resulted in a more balanced and generalized model, reducing overfitting and making the classifier more reliable for cognitive load estimation.

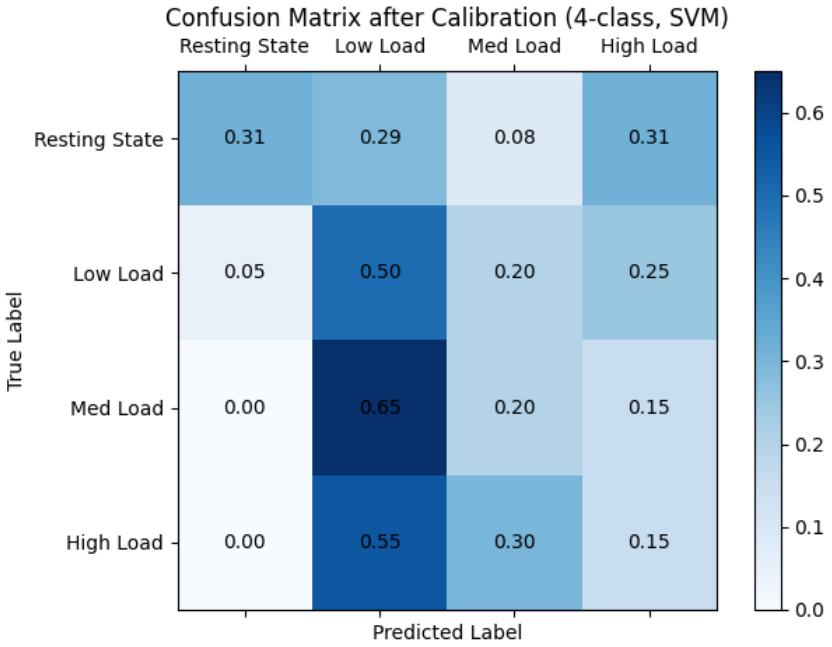


Figure 27: Confusion Matrix - Adapted Classifier 1 - Multiiclass

State	Precision	Recall	F1-score
Resting	0.86	0.31	0.44
Low Load	0.25	0.50	0.34
Medium Load	0.25	0.20	0.22
High Load	0.18	0.15	0.16
Average	0.39	0.30	0.29

Table 8: Performance Metrics - Adapted Classifier 1 - Multiiclass

Binary Model

The binary classification model, distinguishing between Resting State and High Load, exhibited strong overall performance, as indicated by a high recall and precision for both classes, and accuracy increased from 50% to 74%.

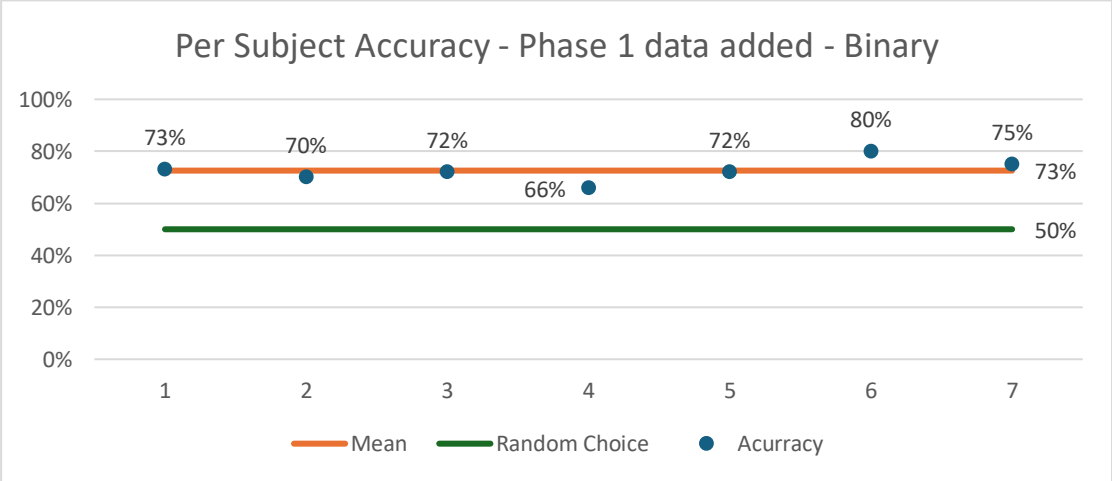


Figure 28: Per Subject Accuracy - Adapted Classifier 1 - Binary

This applies across all subjects, seen in Figure 28, with values ranging from 66% to 80% and a mean accuracy of 73%.

Upon examining the confusion matrix, 52% of Resting State samples were correctly classified, but 48% were misclassified as High Load, indicating some difficulty in distinguishing true resting states. However, High Load samples were classified with 95% recall, meaning that nearly all instances of high cognitive load were correctly identified. The precision for High Load was also perfect (1.0), confirming that the classifier rarely predicts High Load incorrectly. The misclassification of resting samples as high load suggests some overlap in EEG patterns between the two states, possibly due to participant variations or preprocessing artifacts.

Compared to the multi-class model, the binary classifier performed significantly better, likely because it had to distinguish only between two well-separated cognitive states rather than multiple nuanced load levels. These results demonstrate that binary classification of cognitive load states can be highly effective.

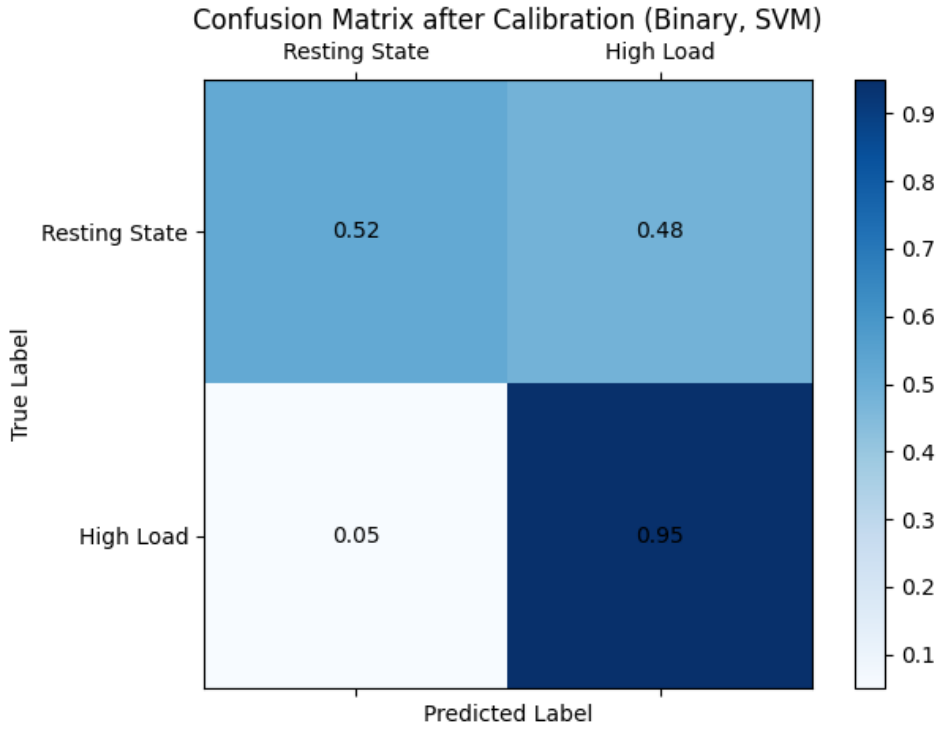


Figure 29: Confusion Matrix - Adapted Classifier 1 - Binary

State	Precision	Recall	F1-score
Resting	0.91	0.52	0.66
High Load	0.66	0.95	0.76
Average	0.79	0.73	0.70

Table 9: Perfomance Metrics - Adapted Classifier 1 - Binary

It was clear from examining the classification data that although the binary classifier did a good job of differentiating between Resting State and High Load, there were some significant misclassifications, especially since 48% of Resting State samples were incorrectly categorized as High Load. This implied that there was some overlap between the EEG properties taken from the two situations, which might have limited the model's capacity to generalize to new, unobserved data.

7.5.3 Training with Experiment's N-Back and Arithmetic Task Data

Considering these findings, it seemed reasonable to investigate whether adding a portion of EEG data from the arithmetic task to the training data may further enhance the classifier. The experiment had the subjects repeat the arithmetic task for two sessions, which resulted in 45 epochs per session, so one session (45 epochs) was added to the training data and the other session was used for testing.

Multiclass Model

The accuracy of the classifier in recognizing states between 4 classes increased from 30% to 34% percent after adding the arithmetic task data to training and overall increased from 26% to 34% accuracy across all subjects, ranging from 29% to 41%, after using additional data to train the initial classifier.

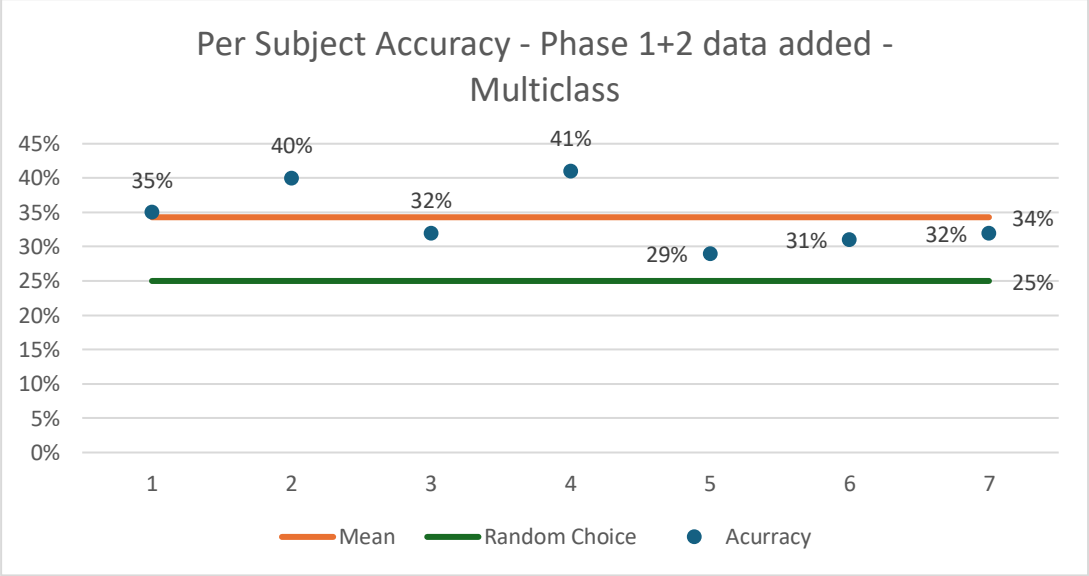


Figure 30: Per Subject Accuracy - Adapted Classifier 2 - Multi-class

The confusion matrix highlights a more balanced distribution of classifications, with reduced overfitting towards the Low Load class and a more effective differentiation of the High Load and Resting State categories. Compared to previous results, the recall and F1-scores for most cognitive states improved, indicating that the additional data helped enhance the generalization of the classifier.

A direct comparison of the classification metrics before and after adding the extra data further supports these observations. The overall recall increased from 30% to 35%, demonstrating that the model became better at identifying true instances of each class. The most substantial improvement is observed in the High Load class, where precision increased from 18% to 42%, recall improved from 15% to 35%, and the F1-score nearly doubled from 16% to 38%. These enhancements suggest that the additional arithmetic task data introduced greater variability in high-load EEG patterns, enabling the classifier to distinguish them more effectively.

Additionally, Medium Load classification exhibited some trade-offs, as precision decreased from 25% to 19% and recall from 20% to 15%, potentially due to increased complexity in separating intermediate cognitive load levels. However, the improvement in Resting State classification (recall from 31% to 33% and F1-score from 44% to 48%) indicates that the model

became less prone to misclassifying rest as a cognitive task, a crucial step toward robust cognitive load estimation. The Low Load category also showed modest gains in precision (25% to 28%) and recall (50% to 55%), reflecting a more balanced classification across load levels.

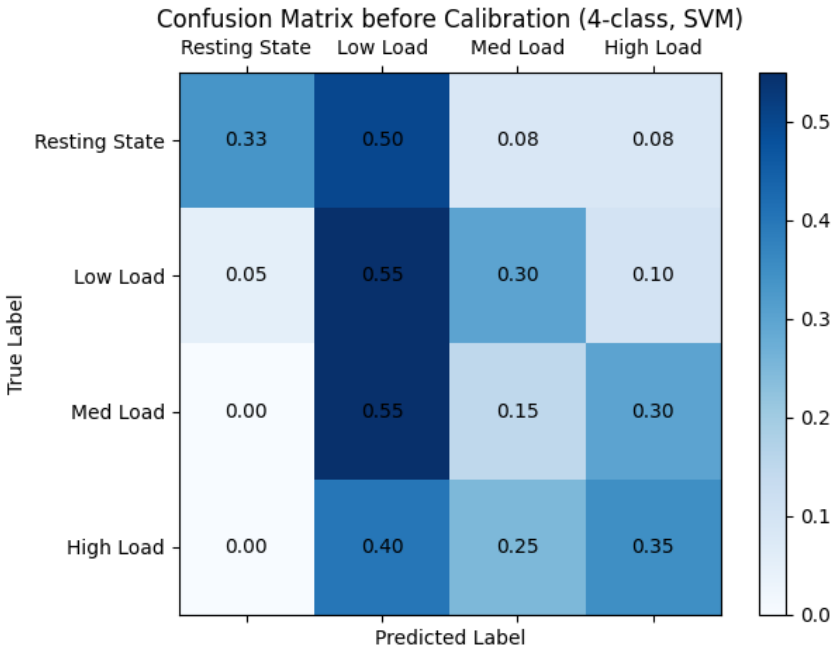


Figure 31: Confusion Matrix - Adapted Classifier 2 - Multi-class

State	Precision	Recall	F1-score
Resting	0.86	0.33	0.48
Low Load	0.28	0.55	0.37
Medium Load	0.19	0.15	0.17
High Load	0.42	0.35	0.38
Average	0.44	0.34	0.35

Table 10: Performance Metrics - Adapted Classifier 2 - Multi-class

Binary Model

As with all models tested, a binary classifier was also trained to distinguish between Resting State and High Load. Accuracy again increased for all subjects, varying from 69% to 88%, with 80% average.

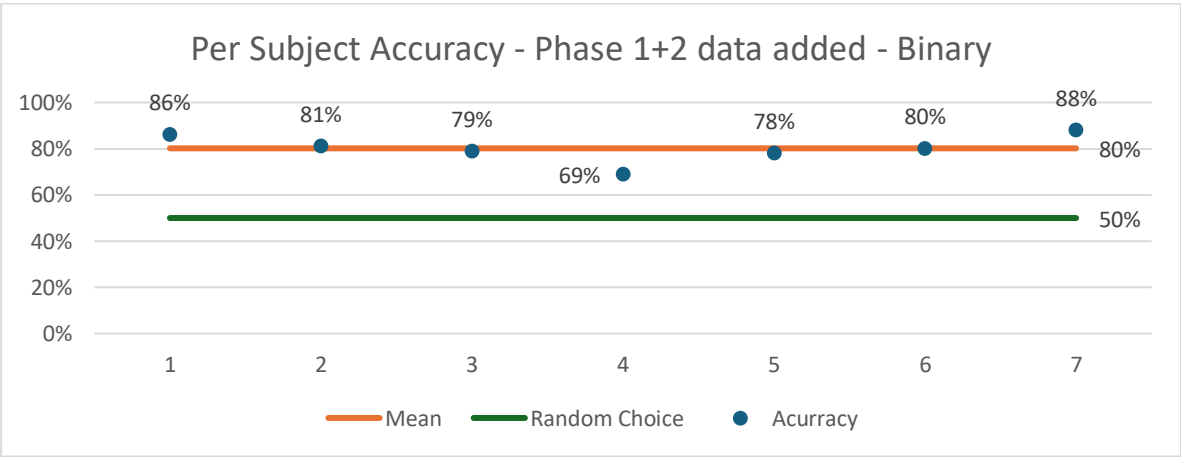


Figure 32: Per Subject Accuracy - Adapted Classifier 2 - Binary

The updated confusion matrix shows a notable improvement in class separation. In the Resting state, recall increased from 0.52 to 0.73, representing a 40% improvement, meaning the model is now significantly better at correctly identifying resting-state samples. While precision for resting slightly decreased from 0.91 to 0.84 (a 7.7% reduction), the overall F1-score improved from 0.66 to 0.78 (an 18% increase), indicating a more balanced classification. For the High Load state, precision improved from 0.66 to 0.76 (15% increase), reducing the number of false positives. Although recall decreased slightly from 0.95 to 0.86 (9.5% reduction), this suggests that the classifier is no longer overly biased toward predicting high load. The F1-score for high load improved from 0.76 to 0.81 (6.6% increase), reflecting a more robust classification performance.

Overall, the classifier's average precision increased from 0.79 to 0.80, recall from 0.73 to 0.80 (9.6% increase), and F1-score from 0.70 to 0.80 (14.3% increase). This demonstrates that the classifier is now more balanced, achieving both high precision and recall across both states. Unlike the earlier version, which had a trade-off between precision and recall, the updated model now performs well in both areas, making it more reliable for real-world cognitive load classification. Overall, the binary model successfully improved precision and recall for High Load detection, making it highly reliable for identifying cognitive effort.

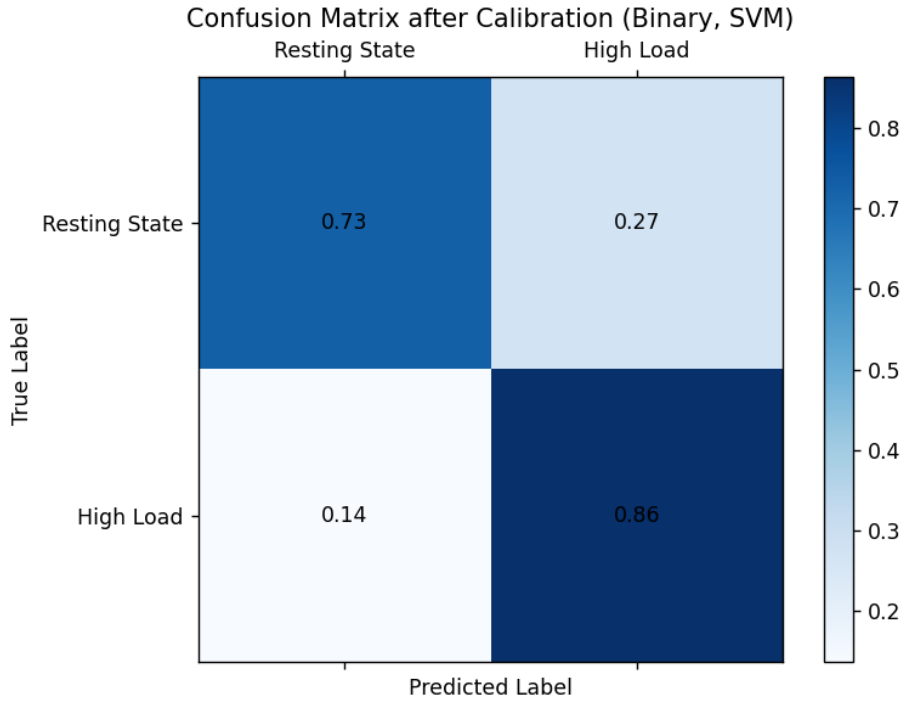


Figure 33: Confusion Matrix - Adapted Classifier 2 - Binary

State	Precision	Recall	F1-score
Resting	0.84	0.73	0.78
High Load	0.76	0.86	0.81
Average	0.80	0.80	0.80

Table 11: Perfomance Metrics - Adapted Classifier 2 - Binary

7.5.4 Training with experiment's Arithmetic Task Data

While the previous approach trained the model using N-back task data from different subjects, the subject's own N-back task data, and arithmetic task data, we hypothesized that removing

the subject's own N-back data and keeping only N-back data from other subjects and a single arithmetic session might lead to better generalization. The key reasoning behind this decision was to examine whether training on a dataset that does not include the subject's own N-back task data could reduce task-specific and subject-specific biases while still allowing the classifier to adapt to a novel task.

By including the subject's own N-back data in training, the model might overfit to individual EEG patterns and task execution styles, leading to a model that works well for this subject's specific cognitive load patterns but fails to generalize to unseen individuals or tasks. In contrast, training on N-back data from different subjects provides a more diverse representation of cognitive load patterns, preventing the classifier from relying on one individual's neural responses. Adding only one arithmetic session serves as a way to test whether the model can still adapt to a different type of cognitive load-inducing task, without depending on prior exposure to the subject's N-back task.

This approach is particularly valuable for assessing the classifier's ability to transfer knowledge across tasks, a critical factor in real-world EEG-based cognitive load estimation. If the model can successfully classify Resting and Cognitive Load states based on inter-subject N-back training data and a single arithmetic session, it suggests that the classifier is learning task-independent cognitive load representations, rather than memorizing patterns specific to a single subject or task. This would make the model more robust for practical applications, where new users might undergo different cognitive tasks that were not explicitly included in the training set.

Multiclass Model

After removing the subject's own N-back data and training the model using N-back data from different subjects along with a single session of arithmetic task data, the classifier's performance showed notable changes, increasing the accuracy from 35% to 36%. The confusion matrix and classification metrics reveal improvements in some areas while introducing new challenges. For most subjects, the accuracy increased by 3-4%, while for some remained the same, bringing the average increase to 1%, as seen in Figure 34.

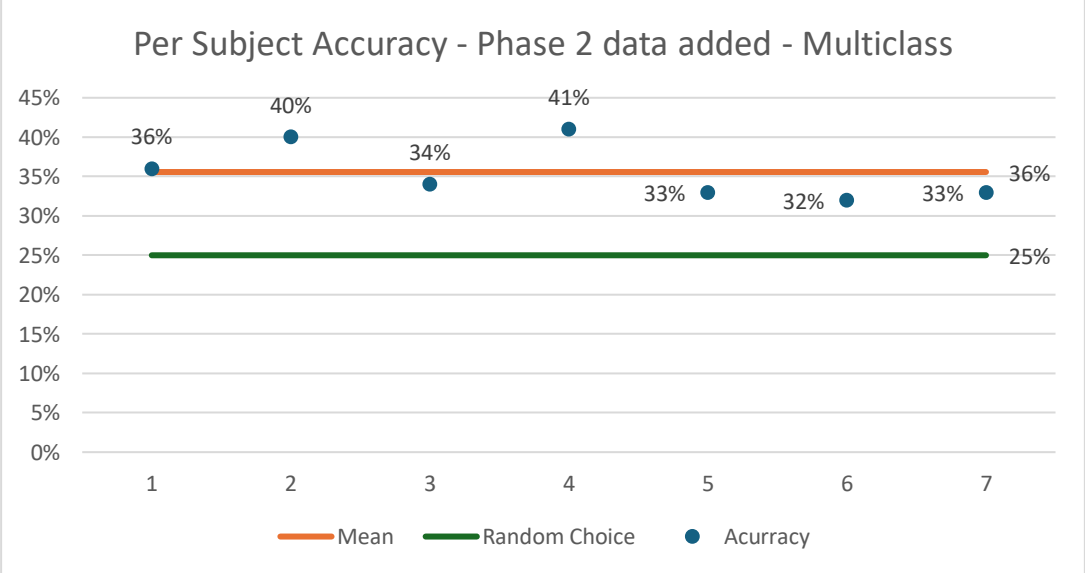


Figure 34: Per Subject Accuracy - Adapted Classifier 3 - Multi-class

Resting State classification saw a slight drop in recall (33% to 29%) but a sharp increase in precision (86% to 100%), meaning that while fewer Resting samples were correctly classified, those that were identified had no false positives. Low Load recall dropped from 55% to 40%, and precision from 28% to 26%, suggesting that Low Load is now harder to classify, likely due to increased overlap with Medium Load.

Medium Load classification improved, with recall rising from 15% to 40% and F1-score increasing from 17% to 36%, showing that the model now recognizes Medium Load better. High Load classification remained stable, with recall staying at 35%, but precision dropped from 42% to 36%, indicating slightly more misclassification.

Overall, average recall improved from 35% to 36%, and F1-score from 35% to 38%, showing a more balanced classification across cognitive states. The model now performs better in distinguishing Medium Load but slightly worse in Low Load classification, with Resting State becoming more selective but less sensitive.

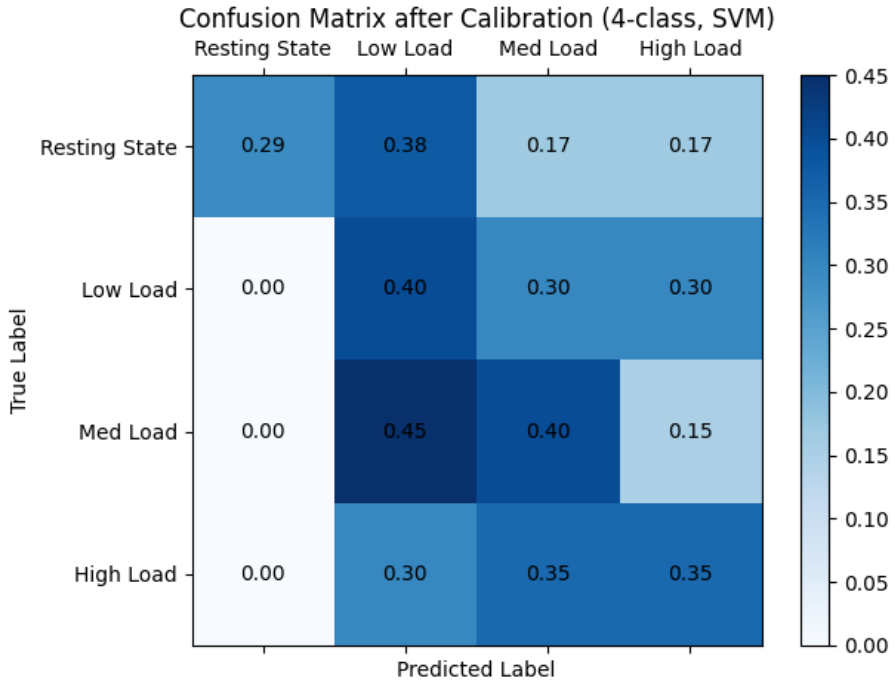


Figure 35: Confusion Matrix - Adapted Classifier 3 - Multi-class

State	Precision	Recall	F1-score
Resting	1.00	0.29	0.49
Low Load	0.26	0.40	0.32
Medium Load	0.32	0.40	0.36
High Load	0.36	0.35	0.36
Average	0.24	0.36	0.38

Table 12: Performance Metrics - Adapted Classifier 3 - Multi-class

Binary Model

A binary classification was again applied using the same data but only for Resting State and High Load State. The updated binary classifier shows a more balanced classification performance compared to the previous model. In this case, for some subjects, the accuracy dropped by 1-2% but for most it increased, bringing the average to 83%, as seen in Figure 36.

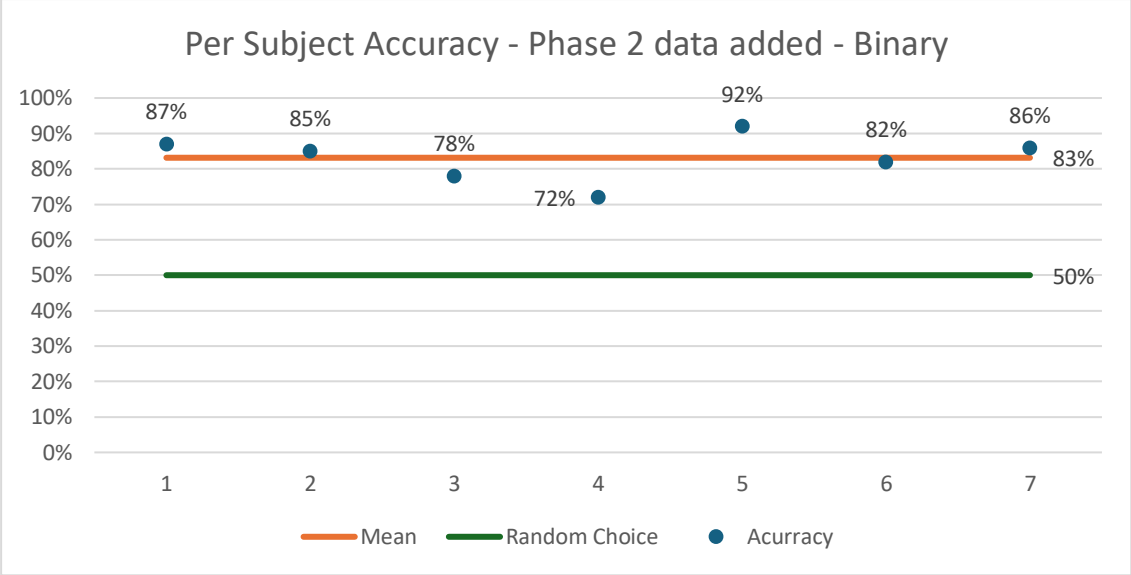


Figure 36: : Per Subject Accuracy - Adapted Classifier 3 - Binary

For the Resting state, recall increased from 0.73 to 0.89, representing a 21.9% improvement, meaning the model is now significantly better at correctly identifying resting-state samples. However, precision slightly decreased from 0.84 to 0.79 (6% reduction), indicating a small trade-off. Despite this, the overall F1-score improved from 0.78 to 0.84, reflecting a 7.7% increase, further enhancing classification balance. For the High Load state, precision improved from 0.76 to 0.87, a 14.5% increase, reducing false positives and making high-load classification more reliable. However, recall slightly decreased from 0.86 to 0.76 (11.6% reduction), suggesting that while the model is more precise, it may occasionally miss high-load instances. Despite this, the F1-score for high load remains stable, changing only slightly from 0.81 to 0.81.

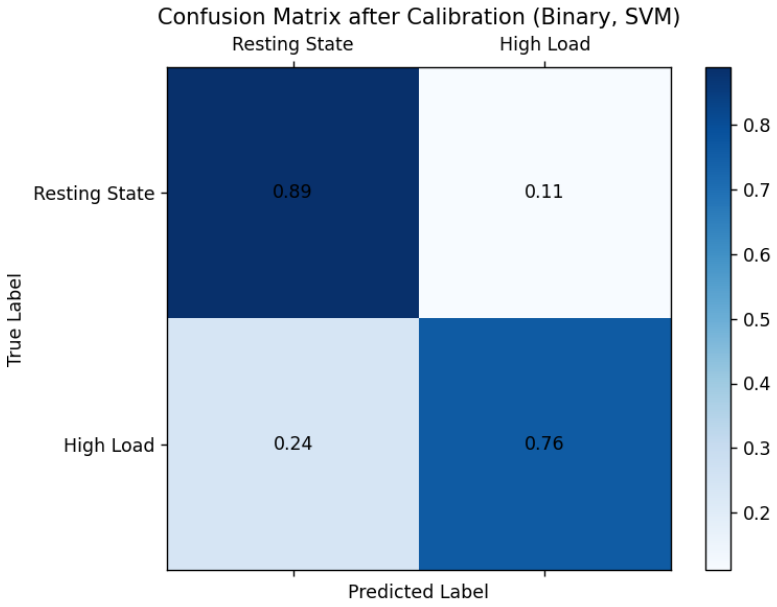


Figure 37: Confusion Matrix - Adapted Classifier 3 - Binary

Overall, the classifier's average precision increased from 0.80 to 0.83, recall remained stable at 0.83, and the F1-score improved from 0.80 to 0.83 (3.8% increase). These improvements

indicate that the third classifier continues to reduce bias while maintaining a strong balance between precision and recall, resulting in a more refined and reliable cognitive load classification system.

State	Precision	Recall	F1-score
Resting	0.79	0.89	0.84
High Load	0.87	0.76	0.81
Average	0.83	0.83	0.83

Table 13: : Perfomance Metrics - Adapted Classifier 3 - Binary

All of the final classification models showed a huge improvement compared to the initial classifier, which had accuracy equivalent to random chance (25% for 4-class and 50% for binary class). Initially, the model struggled to distinguish cognitive load states, with severe misclassification issues, particularly in High Load detection. The latest models now achieve precision, recall, and F1-scores consistently above 30% for 4-class and 80% for binary, confirming that training with additional arithmetic task data and refining the dataset significantly enhanced performance, making the classifier much more reliable for real-world cognitive load estimation.

7.6 Comparative Analysis

7.6.1 Insights into Cognitive Load Variations

The results from the four-class classification models demonstrate a consistent improvement in accuracy as additional training data was incorporated. The initial classifier, trained only on the original dataset, achieved 26% accuracy, which was only slightly better than random chance (25% expected for a four-class problem). When N-back task data from additional subjects was introduced, accuracy increased to 30%, indicating that exposure to more varied cognitive load patterns helped the model differentiate between states more effectively.

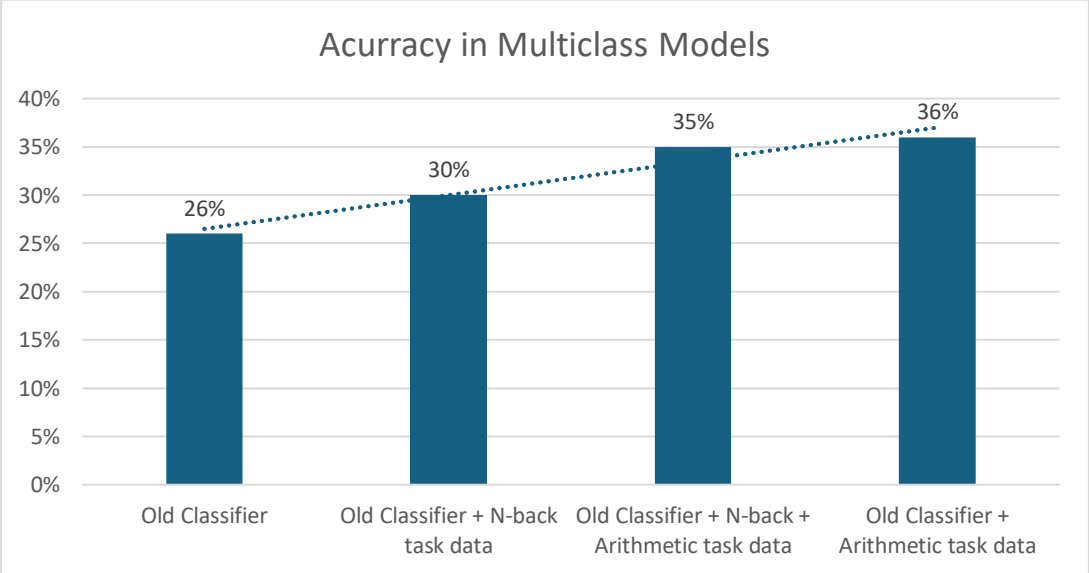


Figure 38: Accuracy across Multi-class Models

Further improvements were observed when arithmetic task data was added, leading to 35% accuracy, suggesting that including diverse task types contributed to better generalization. Interestingly, when the classifier was trained only on the original dataset and arithmetic task

data, accuracy increased further to 36%, implying that removing subject-specific N-back data did not hinder performance and may have reduced task-specific biases.

The binary classification models show a great improvement in accuracy as additional data was introduced. The initial classifier, trained only on the original dataset, had an accuracy of 50%, which is equivalent to random guessing in a two-class problem. When N-back task data from additional subjects was incorporated, accuracy increased significantly to 74%, indicating that more diverse cognitive load samples helped the model learn clearer distinctions between Resting and High Load states. The addition of arithmetic task data further boosted accuracy to 80%, showing that training with a different type of cognitive task improved the model's ability to generalize across varying cognitive load conditions. Interestingly, when the model was trained only on the original dataset and arithmetic task data (without subject-specific N-back data), accuracy increased slightly further to 83%, reinforcing the idea that N-back task data from the same subject may have introduced biases rather than improving generalization.

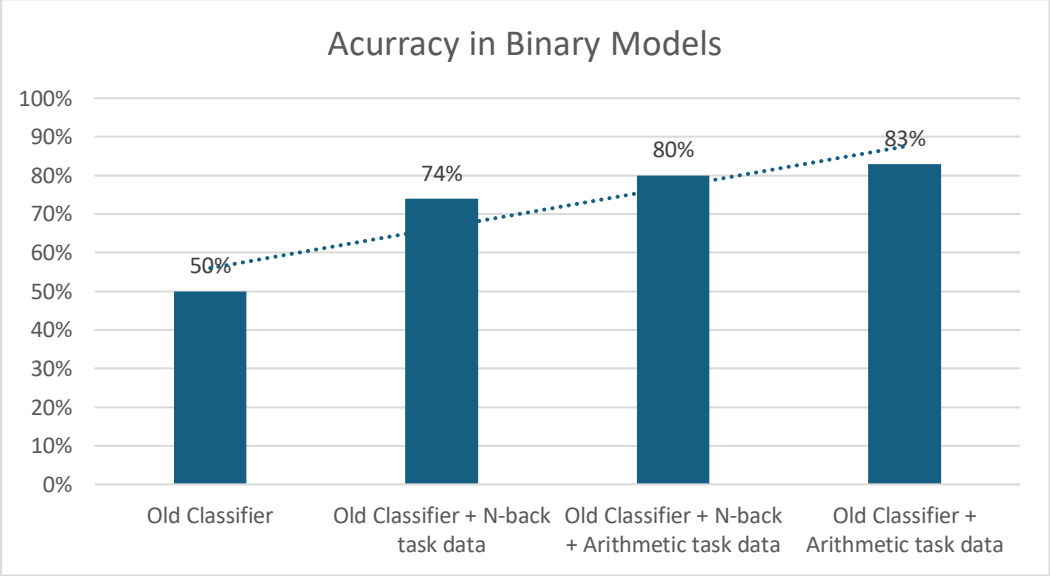


Figure 39: Accuracy across Multi-class Models

The classification results across different models revealed clear trends in how cognitive load levels were distinguished. The Resting State was generally the most precisely classified across all models, particularly in binary classification, where it achieved up to 94% recall. However, in multi-class classification, Resting State was often misclassified as Low Load, suggesting that at times, EEG activity during resting periods shared similarities with lower cognitive effort conditions. Conversely, High Load classification improved significantly when arithmetic task data was included, highlighting that this task induced more distinct EEG patterns compared to N-back tasks alone.

The binary models seemed to outperform the multi-class models, achieving up to 87% accuracy, while multi-class models reached a maximum of 36% accuracy, although when compared to the baseline of 50% and 25% each, models exceeded the baseline by 75% and 64% each. This confirms that binary classification was quite more stable, whereas multi-class classification struggled with separating adjacent load levels. In particular, Medium and Low Load were the most frequently confused, which suggests that EEG responses for these levels may not be as easily distinguishable using the current feature set. The strong performance of binary

models suggests that while EEG-based cognitive load classification is feasible, accurate differentiation among multiple cognitive load states remains challenging with limited data.

The impact of training data composition was also crucial in shaping classifier performance. Adding N-back data from multiple subjects improved generalization, demonstrating that cross-subject EEG patterns provided a broader representation of cognitive load variability. Interestingly, removing subject-specific N-back data did not harm accuracy, suggesting that the classifier did not rely on task-specific neural activity from the same individual. However, models trained only on N-back data generalized poorly to the arithmetic task, reinforcing that exposure to multiple cognitive tasks improves a model's ability to generalize across different cognitive load conditions.

7.6.2 Accuracy and Limitations of Predictions

The accuracy of classification models in training varied significantly, with binary classifiers outperforming multi-class models. Cross-subject variability and arithmetic task data improved accuracy, indicating the importance of training on multiple cognitive tasks for improved model generalization and reliability in distinguishing between Resting and High Load.

Despite these improvements, several limitations remain. The dataset size was a major constraint, as adding only 40-45 new samples caused significant accuracy shifts, indicating that the models lacked sufficient data for stable learning. This sensitivity suggests potential overfitting, where the classifier may be adapting too closely to the available samples rather than learning generalizable patterns. The confusion between Medium and Low Load further highlights that the current feature set may not be sufficient for precise differentiation between moderate cognitive states. Another challenge is task dependency, as improvements in classification were largely driven by arithmetic task data. However, it remains uncertain whether the model would generalize well to cognitive tasks beyond N-back and arithmetic.

Additionally, hardware differences in EEG recording systems could have influenced classification performance. The initial dataset was recorded using an EEG system with 64 active Ag-AgCl electrodes (ActiCap, Brain Products GmbH) and an ActiCHamp amplifier (Brain Products, GmbH), positioned according to the extended 10-20 system (Oostenveld & Praamstra, 2001). In contrast, the additional experimental data was recorded using a Biosemi system with 32 active electrodes, also following the 10-20 system. This difference in electrode count, amplifier technology, and signal processing methods may have introduced variability in the EEG features, potentially affecting classification consistency. Despite this, the classifier adapted well to the new dataset, demonstrating its ability to generalize across different EEG recording setups. However, the impact of hardware variations should be further investigated, as differences in electrode density, signal quality, and amplifier characteristics may influence feature extraction and classification accuracy, particularly in real-world applications where lower-density consumer-grade EEG devices may be used.

8. Discussion

8.1 Implications of Findings

The findings of this study provide valuable insights into the use of EEG-based machine learning techniques for cognitive load estimation. The classification results indicate that cognitive load levels can be effectively distinguished using EEG features, particularly those derived from the frontal and parental regions. The analysis of preprocessing steps, including filtering, ICA artifact removal, and channel selection, demonstrated their importance in improving classification accuracy.

A key implication of these findings is the potential application of EEG-based cognitive load monitoring in real-world scenarios. The ability to classify cognitive load states with reasonable accuracy suggests that such systems could be integrated into adaptive learning environments, human-computer interaction applications, and cognitive workload assessment tools in high-demand professions such as aviation, medicine, and military operations.

Another important implication relates to the generalizability of classifiers trained on standard cognitive tasks. The study revealed that classifiers trained on a well-established cognitive task such as the N-Back test may not generalize well to real-world conditions. This highlights the necessity of adapting models to different task environments and incorporating domain adaptation techniques to improve robustness.

8.2 Challenges Encountered

One of the primary challenges encountered in this study was the hardware malfunction on the trigger box, which led to exclusion of one participants data. Certain electrodes frequently exhibited poor contact, resulting in unreliable signals. Another significant challenge was the necessity to exclude certain subjects from analysis due to excessive noise in their EEG recordings. Subjects whose recordings had more than 40% of epochs contaminated by artifacts were excluded to ensure the integrity of the dataset. While this step was necessary, it also reduced the sample size, potentially impacting the robustness of the findings.

Moreover, EEG signals exhibit high inter-subject variability, making it difficult to develop universally applicable classifiers. The complexity of EEG data necessitates extensive preprocessing, including filtering, independent component analysis (ICA), and feature selection, to obtain meaningful patterns.

A major obstacle in EEG-based cognitive load estimation is the challenge of generalizing models across different tasks. While classifiers performed well within the experimental setup, their accuracy declined when applied to new datasets or task paradigms. This underscores the need for transfer learning techniques or subject-specific calibration to improve cross-task and cross-subject adaptability. Additionally, it was observed that classifiers exhibited instability when small amounts of new data were added, leading to significant variations in results. This suggests that the dataset used in training may not be sufficiently large to allow for robust generalization, highlighting the need for larger and more diverse datasets to improve model stability and reliability.

8.3 Comparison with Related Studies and Future Research

The findings of this study matched previous research in EEG-based cognitive load estimation and at the same time brought new challenges to light. Previous studies have found strong correlations between EEG features, especially theta and beta power spectral densities, and cognitive workload, and moreover, the frontal electrodes along with two electrodes from each other region are providing the most valuable information. We confirm these results in our study yet placing even greater emphasis on preprocessing as a way to improve classification performance.

This work differs from related work in its focus on classifier generalization. Many studies achieve high classification accuracy within a single experimental setting, but few address the issue of adapting classifiers to new tasks. The observed drop in accuracy when applying trained models to different tasks speaks for itself: future research should concentrate on domain adaptation techniques to make models more robust.

Indeed, future research efforts should involve the progress of building more robust machine learning models that generalize across different cognitive tasks and experimental setups. Domain adaptation seems a plausible direction that could mitigate the drop in performance reflected when classifiers are applied to new tasks. Approaches of transfer learning might also be used to improve adaptability to individual differences by fine-tuning pre-trained models using small amounts of subject-specific data.

Another very important aspect of future work will be to improve the real-time EEG processing capabilities. Data collection and analysis in real-time by reliable algorithms would give rise to interesting applications such as user interfaces that automatically adapt to cognitive state monitoring within workplace environments. Furthermore, such research should adapt to hardware by explorations of portable, low-cost EEG systems. By systematically analyzing how hardware differences and other real-world factors affect classification, researchers can refine guidelines for standardized data collection and develop models that adapt more gracefully to heterogeneous conditions, ultimately improving the reliability and reproducibility of EEG-based cognitive load detection. Also, minimal number of electrodes with a constant classification performance would serve the purpose of making an EEG-based estimation of cognitive load practical for widespread use.

Furthermore, future studies should aim to increase the dataset size by including a larger and more diverse pool of participants. Expanding the dataset would improve classifier robustness, reduce variability in performance across subjects, and enhance model generalization to real-world settings. The inclusion of additional participants from different demographics and cognitive backgrounds would also provide insights into how cognitive load estimation models perform across varied populations, making them more applicable to practical use cases.

9. Citation

[1] J. Sweller, “Cognitive Load During Problem Solving: Effects on Learning,” *Cogn Sci*, vol. 12, no. 2, pp. 257–285, Apr. 1988, doi: 10.1207/s15516709cog1202_4.

[2] F. G. W. C. Paas and J. J. G. Van Merriënboer, “Variability of Worked Examples and Transfer of Geometrical Problem-Solving Skills: A Cognitive-Load Approach,” *J Educ Psychol*, vol. 86, no. 1, pp. 122–133, 1994, doi: 10.1037/0022-0663.86.1.122.

[3] R. E. Mayer and R. Moreno, “MAYER AND MORENO WAYS TO REDUCE COGNITIVE LOAD Nine Ways to Reduce Cognitive Load in Multimedia Learning WHAT IS MULTIMEDIA LEARNING AND INSTRUCTION?,” 2003.

[4] A. Szulewski, D. Howes, J. J. G. Van Merriënboer, and J. Sweller, “From Theory to Practice: The Application of Cognitive Load Theory to the Practice of Medicine,” Jan. 01, 2021, *Wolters Kluwer Health*. doi: 10.1097/ACM.0000000000003524.

[5] R. Liu, L. Wang, T. A. Koszalka, and K. Wan, “Effects of immersive virtual reality classrooms on students’ academic achievement, motivation and cognitive load in science lessons,” *J Comput Assist Learn*, vol. 38, no. 5, pp. 1422–1433, Oct. 2022, doi: 10.1111/JCAL.12688.

[6] J. Sweller, “Cognitive load theory and educational technology,” *Educational Technology Research and Development*, vol. 68, no. 1, pp. 1–16, Feb. 2020, doi: 10.1007/s11423-019-09701-3.

[7] J. C. Wilson, S. Nair, S. Scielzo, and E. C. Larson, “Objective Measures of Cognitive Load Using Deep Multi-Modal Learning,” *Proc ACM Interact Mob Wearable Ubiquitous Technol*, vol. 5, no. 1, Mar. 2021, doi: 10.1145/3448111.

[8] *2019 IEEE Frontiers in Education Conference (FIE)*. Institute of Electrical and Electronics Engineers (IEEE), 2019.

[9] J. L. Sewell, L. A. Maggio, O. ten Cate, T. van Gog, J. Q. Young, and P. S. O’Sullivan, “Cognitive load theory for training health professionals in the workplace: A BEME review of studies among diverse professions: BEME Guide No. 53,” *Med Teach*, vol. 41, no. 3, pp. 256–270, Mar. 2019, doi: 10.1080/0142159X.2018.1505034.

[10] Y. Zhou, S. Huang, Z. Xu, P. Wang, X. Wu, and D. Zhang, “Cognitive Workload Recognition Using EEG Signals and Machine Learning: A Review,” Sep. 01, 2022, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/TCDS.2021.3090217.

[11] N. Hollender, C. Hofmann, M. Deneke, and B. Schmitz, “Integrating cognitive load theory and concepts of human–computer interaction,” *Comput Human Behav*, vol. 26, no. 6, pp. 1278–1288, Nov. 2010, doi: 10.1016/J.CHB.2010.05.031.

[12] M. Gjoreski *et al.*, “Cognitive Load Monitoring with Wearables-Lessons Learned from a Machine Learning Challenge,” *IEEE Access*, vol. 9, pp. 103325–103336, 2021, doi: 10.1109/ACCESS.2021.3093216.

[13] S. N. Miri Ashtiani and M. R. Daliri, “Identification of cognitive load-dependent activation patterns using working memory task-based fMRI at various levels of difficulty,”

Scientific Reports 2023 13:1, vol. 13, no. 1, pp. 1–13, Sep. 2023, doi: 10.1038/s41598-023-43837-w.

[14] I. Constant and N. Sabourdin, "The EEG signal: A window on the cortical brain activity," Jun. 2012. doi: 10.1111/j.1460-9592.2012.03883.x.

[15] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," *Advances in Psychology*, vol. 52, no. C, pp. 139–183, Jan. 1988, doi: 10.1016/S0166-4115(08)62386-9.

[16] F. Paas, "Training_strategies_for_attaining_transf," 1992.

[17] M. Krell, K. M. Xu, G. D. Rey, and F. Paas, "Editorial: Recent Approaches for Assessing Cognitive Load From a Validity Perspective," Jan. 24, 2022, *Frontiers Media S.A.* doi: 10.3389/feduc.2021.838422.

[18] Pawan and R. Dhiman, "Machine learning techniques for electroencephalogram based brain-computer interface: A systematic literature review," *Measurement: Sensors*, vol. 28, p. 100823, Aug. 2023, doi: 10.1016/J.MEASEN.2023.100823.

[19] F. Khanam, A. B. M. A. Hossain, and M. Ahmad, "Electroencephalogram-based cognitive load level classification using wavelet decomposition and support vector machine," *Brain-Computer Interfaces*, vol. 10, no. 1, pp. 1–15, Jan. 2023, doi: 10.1080/2326263X.2022.2109855.

[20] A. Saha, V. Minz, S. Bonela, S. R. Sreeja, R. Chowdhury, and D. Samanta, "Classification of EEG signals for cognitive load estimation using deep learning architectures," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11278 LNCS, pp. 59–68, 2018, doi: 10.1007/978-3-030-04021-5_6/TABLES/4.

[21] "What are the parts of the nervous system? | NICHD - Eunice Kennedy Shriver National Institute of Child Health and Human Development." Accessed: Dec. 16, 2024. [Online]. Available: https://www.nichd.nih.gov/health/topics/neuro/conditioninfo/parts#f1

[22] "Brain Basics: Know Your Brain | National Institute of Neurological Disorders and Stroke." Accessed: Dec. 16, 2024. [Online]. Available: https://www.ninds.nih.gov/health-information/public-education/brain-basics/brain-basics-know-your-brain

[23] T. Bui and J. M. Das, "Neuroanatomy, Cerebral Hemisphere," *StatPearls*, Jul. 2023, Accessed: Dec. 11, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK549789/

[24] "Lobes of the brain - Queensland Brain Institute - University of Queensland." Accessed: Dec. 16, 2024. [Online]. Available: https://qbi.uq.edu.au/brain/brain-anatomy/lobes-brain

[25] S. Jimsheleishvili and M. Dididze, "Neuroanatomy, Cerebellum," *StatPearls*, 2024, Accessed: Dec. 11, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30844194/

[26] H. Basinger and J. P. Hogg, “Neuroanatomy, Brainstem,” *StatPearls*, Jul. 2023, Accessed: Dec. 11, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK544297/

[27] M. X. Cohen, “Where Does EEG Come From and What Does It Mean?,” Apr. 01, 2017, *Elsevier Ltd*. doi: 10.1016/j.tins.2017.02.004.

[28] K. Blinowska and P. Durka, “ELECTROENCEPHALOGRAPHY (EEG).”

[29] C.-H. Im, “Basics of EEG: Generation, Acquisition, and Applications of EEG,” pp. 3–11, 2018, doi: 10.1007/978-981-13-0908-3_1.

[30] “The ten-twenty electrode system of the International Federation. The International Federation of Clinical Neurophysiology - PubMed.” Accessed: Dec. 16, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/10590970/

[31] “Fundamental of EEG Measurement.” Accessed: Dec. 16, 2024. [Online]. Available: https://www.researchgate.net/publication/228599963_Fundamental_of_EEG_Measurement

[32] L. Sun, Y. Liu, and P. J. Beadle, “Independent component analysis of EEG signals,” *Proceedings of the 2005 IEEE International Workshop on VLSI Design and Video Technology, IWVDVT 2005*, pp. 293–296, 2005, doi: 10.1109/IWVDVT.2005.1504590.

[33] P. Antonenko, F. Paas, R. Grabner, and T. van Gog, “Using Electroencephalography to Measure Cognitive Load,” Dec. 2010. doi: 10.1007/s10648-010-9130-y.

[34] P. Antonenko, F. Paas, R. Grabner, and T. van Gog, “Using Electroencephalography to Measure Cognitive Load,” *Educ Psychol Rev*, vol. 22, no. 4, pp. 425–438, Dec. 2010, doi: 10.1007/S10648-010-9130-Y/FIGURES/3.

[35] E. Donchin, “Surprise!… Surprise?,” *Psychophysiology*, vol. 18, no. 5, pp. 493–513, Sep. 1981, doi: 10.1111/J.1469-8986.1981.TB01815.X.

[36] S. M. Jaeggi, M. Buschkuehl, W. J. Perrig, and B. Meier, “The concurrent validity of the N-back task as a working memory measure,” *MEMORY*, vol. 18, no. 4, pp. 394–412, May 2010, doi: 10.1080/09658211003702171.

[37] R. Chen *et al.*, “Research on multi-dimensional N-back task induced EEG variations,” *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2015-November, pp. 5163–5166, Nov. 2015, doi: 10.1109/EMBC.2015.7319554.

[38] R. Patel, K. Gireesan, R. Baskaran, and N. V. C. Shekar, “Optimal classification of N-back task EEG data by performing effective feature reduction,” *Sadhana - Academy Proceedings in Engineering Sciences*, vol. 47, no. 4, pp. 1–12, Dec. 2022, doi: 10.1007/S12046-022-02015-W/TABLES/7.

[39] N. B. Maimon, L. Molcho, N. Intrator, and D. Lamy, “Single-channel EEG features during n-back task correlate with working memory load,” Aug. 2020, Accessed: Feb. 05, 2025. [Online]. Available: https://arxiv.org/abs/2008.04987v1

[40] C. Scharinger, A. Soutschek, T. Schubert, and P. Gerjets, “Comparison of the working memory load in N-back and working memory span tasks by means of EEG frequency

band power and P300 amplitude," *Front Hum Neurosci*, vol. 11, p. 212526, Jan. 2017, doi: 10.3389/FNHUM.2017.00006/BIBTEX.

[41] M. Liebherr *et al.*, "EEG and behavioral correlates of attentional processing while walking and navigating naturalistic environments," *bioRxiv*, p. 2021.05.27.445993, Aug. 2021, doi: 10.1101/2021.05.27.445993.

[42] M. K *et al.*, "Dynamic cognitive workload assessment for fighter pilots in simulated fighter aircraft environment using EEG," *Biomed Signal Process Control*, vol. 61, p. 102018, Aug. 2020, doi: 10.1016/J.BSPC.2020.102018.

[43] J. A. Blanco *et al.*, "Quantifying cognitive workload in simulated flight using passive, dry EEG measurements," *IEEE Trans Cogn Dev Syst*, vol. 10, no. 2, pp. 373–383, Jun. 2018, doi: 10.1109/TCDS.2016.2628702.

[44] J. M. Morales, J. F. Ruiz-Rabelo, C. Diaz-Piedra, and L. L. Di Stasi, "Detecting Mental Workload in Surgical Teams Using a Wearable Single-Channel Electroencephalographic Device," *J Surg Educ*, vol. 76, no. 4, pp. 1107–1115, Jul. 2019, doi: 10.1016/J.JSURG.2019.01.005.

[45] N. B. Maimon *et al.*, "Continuous Monitoring of Mental Load During Virtual Simulator Training for Laparoscopic Surgery Reflects Laparoscopic Dexterity: A Comparative Study Using a Novel Wireless Device," *Front Neurosci*, vol. 15, p. 694010, Jan. 2022, doi: 10.3389/FNINS.2021.694010/BIBTEX.

[46] J. Morton *et al.*, "Identifying predictive EEG features for cognitive overload detection in assembly workers in Industry 4.0," 2019.

[47] F. Schaule, J. O. Johanssen, B. Bruegge, and V. Loftness, "Employing Consumer Wearables to Detect Office Workers' Cognitive Load for Interruption Management," *Proc ACM Interact Mob Wearable Ubiquitous Technol*, vol. 2, no. 1, pp. 1–20, Mar. 2018, doi: 10.1145/3191764.

[48] R. Sun, S. Xu, Y. Han, D. Zhuang, B. Yan, and C. Sun, "Exploring impacts of thermal and lighting conditions on office workers' subjective evaluations, cognitive performance and EEG features in multi-person offices," *Journal of Building Engineering*, vol. 95, p. 110170, Oct. 2024, doi: 10.1016/J.JOBE.2024.110170.

[49] V. Ahonen, M. Leino, and T. Lipping, "Electroencephalography in Evaluating Mental Workload of Gaming," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pp. 845–848, 2021, doi: 10.1109/EMBC46164.2021.9629772.

[50] S. Das, L. Ghosh, and S. Saha, "Analyzing Gaming Effects on Cognitive Load Using Artificial Intelligent Tools," *Proceedings of CONECCT 2020 - 6th IEEE International Conference on Electronics, Computing and Communication Technologies*, Jul. 2020, doi: 10.1109/CONECCT50063.2020.9198662.

[51] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni, "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness," *Neurosci Biobehav Rev*, vol. 44, pp. 58–75, 2014, doi: 10.1016/j.neubiorev.2012.10.003.

[52] S. Ma, X. Yan, J. Billington, N. Merat, and G. Markkula, “Cognitive load during driving: EEG microstate metrics are sensitive to task difficulty and predict safety outcomes,” *Accid Anal Prev*, vol. 207, p. 107769, Nov. 2024, doi: 10.1016/J.AAP.2024.107769.

[53] D. He, B. Donmez, C. C. Liu, and K. N. Plataniotis, “High Cognitive Load Assessment in Drivers Through Wireless Electroencephalography and the Validation of a Modified N-Back Task,” *IEEE Trans Hum Mach Syst*, vol. 49, no. 4, pp. 362–371, Aug. 2019, doi: 10.1109/THMS.2019.2917194.

[54] A. Sinha, Di. Roy, R. Chaki, B. K. De, and S. K. Saha, “Readability Analysis Based on Cognitive Assessment Using Physiological Sensing,” *IEEE Sens J*, vol. 19, no. 18, pp. 8127–8135, Sep. 2019, doi: 10.1109/JSEN.2019.2917834.

[55] N. Kasabov, M. H. Fard, K. Petrova, and G. Y. Wang, “Studying Transfer of Learning using a BrainInspired Spiking Neural Network in the Context of Learning a New Programming Language,” *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering, CSDE 2021*, 2021, doi: 10.1109/CSDE53843.2021.9718472.

[56] S. Baceviciute, A. Mottelson, T. Terkildsen, and G. Makransky, “Investigating Representation of Text and Audio in Educational VR using Learning Outcomes and EEG,” *Conference on Human Factors in Computing Systems - Proceedings*, Apr. 2020, doi: 10.1145/3313831.3376872/SUPPL_FILE/PN9763AUX.ZIP.

[57] A. Diwakar, T. Kaur, C. Ralekar, and T. K. Gandhi, “Deep Learning Identifies Brain Cognitive Load Via EEG Signals,” *2020 IEEE 17th India Council International Conference, INDICON 2020*, Dec. 2020, doi: 10.1109/INDICON49873.2020.9342222.

[58] M. O. Zhuravlev, R. V. Ukolov, A. E. Runnova, and A. A. Titova, “The method of spatial-temporal analysis of various oscillational modes in the EEG brain activity,” *Proceedings of the 2020 IEEE International Conference “Quality Management, Transport and Information Security, Information Technologies”, IT and QM and IS 2020*, pp. 562–565, Sep. 2020, doi: 10.1109/ITQMIS51053.2020.9322963.

[59] Z. Y. Chin, X. Zhang, C. Wang, and K. K. Ang, “EEG-based discrimination of different cognitive workload levels from mental arithmetic,” *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2018-July, pp. 1984–1987, Oct. 2018, doi: 10.1109/EMBC.2018.8512675.

[60] M. Spüler, C. Walter, W. Rosenstiel, P. Gerjets, K. Moeller, and E. Klein, “EEG-based prediction of cognitive workload induced by arithmetic: a step towards online adaptation in numerical learning,” *ZDM - Mathematics Education*, vol. 48, no. 3, pp. 267–278, Jun. 2016, doi: 10.1007/S11858-015-0754-8.

[61] N. Kumar and J. Kumar, “Measurement of Cognitive Load in HCI Systems Using EEG Power Spectrum: An Experimental Study,” *Procedia Comput Sci*, vol. 84, pp. 70–78, Jan. 2016, doi: 10.1016/J.PROCS.2016.04.068.

[62] Q. Wang and O. Sourina, “Real-time mental arithmetic task recognition from EEG signals,” *IEEE Trans Neural Syst Rehabil Eng*, vol. 21, no. 2, pp. 225–232, 2013, doi: 10.1109/TNSRE.2012.2236576.

[63] J. Molina del Río, M. A. Guevara, M. Hernández González, R. M. Hidalgo Aguirre, and M. A. Cruz Aguilar, “EEG correlation during the solving of simple and complex logical–

mathematical problems," *Cogn Affect Behav Neurosci*, vol. 19, no. 4, pp. 1036–1046, Aug. 2019, doi: 10.3758/S13415-019-00703-5/FIGURES/3.

[64] L. Cabañero Gómez, R. Hervás, I. González, and V. Villarreal, "Studying the generalisability of cognitive load measured with EEG," *Biomed Signal Process Control*, vol. 70, p. 103032, Sep. 2021, doi: 10.1016/J.BSPC.2021.103032.

[65] Y. Ke *et al.*, "An EEG-Based mental workload estimator trained on working memory task can work well under simulated Multi-Attribute task," *Front Hum Neurosci*, vol. 8, no. SEP, p. 107808, Sep. 2014, doi: 10.3389/FNHUM.2014.00703/BIBTEX.

[66] G. N. Dimitrakopoulos *et al.*, "Task-Independent Mental Workload Classification Based Upon Common Multiband EEG Cortical Connectivity," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1940–1949, Nov. 2017, doi: 10.1109/TNSRE.2017.2701002.

[67] R. Patel, K. Gireesan, R. Baskaran, and N. V. C. Shekar, "Optimal classification of N-back task EEG data by performing effective feature reduction," *Sadhana - Academy Proceedings in Engineering Sciences*, vol. 47, no. 4, pp. 1–12, Dec. 2022, doi: 10.1007/S12046-022-02015-W/TABLES/7.

[68] S. Stober, A. Sternin, A. M. Owen, and J. A. Grahn, "Deep Feature Learning for EEG Recordings," Nov. 2015, Accessed: Feb. 07, 2025. [Online]. Available: https://arxiv.org/abs/1511.04306v4

[69] B. Yang, K. Duan, C. Fan, C. Hu, and J. Wang, "Automatic ocular artifacts removal in EEG using deep learning," *Biomed Signal Process Control*, vol. 43, pp. 148–158, May 2018, doi: 10.1016/J.BSPC.2018.02.021.

[70] K. Kyriaki, D. Koukopoulos, and C. A. Fidas, "A Comprehensive Survey of EEG Preprocessing Methods for Cognitive Load Assessment," *IEEE Access*, vol. 12, pp. 23466–23489, 2024, doi: 10.1109/ACCESS.2024.3360328.

[71] R. Mahmoud, T. Shanableh, I. P. Bodala, N. V. Thakor, and H. Al-Nashash, "Novel Classification System for Classifying Cognitive Workload Levels under Vague Visual Stimulation," *IEEE Sens J*, vol. 17, no. 21, pp. 7019–7028, Nov. 2017, doi: 10.1109/JSEN.2017.2727539.

[72] M. F. Hinss, E. S. Jahanpour, B. Somon, L. Pluchon, F. Dehais, and R. N. Roy, "COG-BCI database: A multi-session and multi-task EEG cognitive dataset for passive brain-computer interfaces", doi: 10.5281/ZENODO.7413650.

[73] M. F. Hinss, E. S. Jahanpour, B. Somon, L. Pluchon, F. Dehais, and R. N. Roy, "Open multi-session and multi-task EEG cognitive Dataset for passive brain-computer Interface Applications," *Scientific Data 2023 10:1*, vol. 10, no. 1, pp. 1–22, Feb. 2023, doi: 10.1038/s41597-022-01898-y.

[74] A. M. Brouwer, M. A. Hogervorst, J. B. F. Van Erp, T. Heffelaar, P. H. Zimmerman, and R. Oostenveld, "Estimating workload using EEG spectral power and ERPs in the n-back task," *J Neural Eng*, vol. 9, no. 4, Aug. 2012, doi: 10.1088/1741-2560/9/4/045008.

[75] M. F. Hinss *et al.*, “An EEG dataset for cross-session mental workload estimation: Passive BCI competition of the Neuroergonomics Conference 2021”, doi: 10.5281/ZENODO.4917218.

[76] A. Gramfort *et al.*, “MEG and EEG data analysis with MNE-Python,” *Front Neurosci*, vol. 7, no. 7 DEC, p. 70133, Dec. 2013, doi: 10.3389/FNINS.2013.00267/BIBTEX.

[77] “Filtering EEG Data — Neural Data Science in Python.” Accessed: Dec. 23, 2024. [Online]. Available: https://neuraldatascience.io/7-eeg/erp_filtering.html

[78] S. Tsuchimoto *et al.*, “Use of common average reference and large-Laplacian spatial-filters enhances EEG signal-to-noise ratios in intrinsic sensorimotor activity,” *J Neurosci Methods*, vol. 353, p. 109089, Apr. 2021, doi: 10.1016/J.JNEUMETH.2021.109089.

[79] N. Bigdely-Shamlo, T. Mullen, C. Kothe, K. M. Su, and K. A. Robbins, “The PREP pipeline: standardized preprocessing for large-scale EEG analysis,” *Front Neuroinform*, vol. 9, no. JUNE, pp. 1–19, Jun. 2015, doi: 10.3389/FNINF.2015.00016.

[80] A. Apicella, F. Isgrò, A. Pollastro, and R. Prevete, “On the effects of data normalization for domain adaptation on EEG data,” *Eng Appl Artif Intell*, vol. 123, p. 106205, Aug. 2023, doi: 10.1016/J.ENGAPPAI.2023.106205.

[81] A. Rajalakshmi and S. S. Sridhar, “A Study of Time Domain Features of EEG Signal Analysis”.

[82] Y. Liu *et al.*, “EEG based Mental Workload Assessment by Power Spectral Density Feature,” *2022 IEEE International Conference on Mechatronics and Automation, ICMA 2022*, pp. 450–454, 2022, doi: 10.1109/ICMA54519.2022.9856376.

[83] P. D. Welch, “The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms,” *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967, doi: 10.1109/TAU.1967.1161901.

[84] Y. Kumar, M. L. Dewal, and R. S. Anand, “Features extraction of EEG signals using approximate and sample entropy,” *2012 IEEE Students’ Conference on Electrical, Electronics and Computer Science: Innovation for Humanity, SCEECS 2012*, 2012, doi: 10.1109/SCEECS.2012.6184830.

[85] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal Of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2011, doi: 10.1613/jair.953.

[86] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “Training algorithm for optimal margin classifiers,” *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pp. 144–152, 1992, doi: 10.1145/130385.130401.

[87] “Support Vector Machines (SVM): An Intuitive Explanation | by Tasmay Pankaj Tibrewal | Low Code for Data Science | Medium.” Accessed: Jan. 04, 2025. [Online]. Available: https://medium.com/low-code-for-advanced-data-science/support-vector-machines-svm-an-intuitive-explanation-b084d6238106

[88] X. Li, X. Chen, Y. Yan, W. Wei, and Z. J. Wang, “Classification of EEG Signals Using a Multiple Kernel Learning Support Vector Machine,” *Sensors 2014, Vol. 14, Pages 12784-12802*, vol. 14, no. 7, pp. 12784–12802, Jul. 2014, doi: 10.3390/S140712784.

[89] L. Breiman, “Random forests,” *Mach Learn*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324/METRICS.

[90] J. Gómez-Ramírez, M. Ávila-Villanueva, and M. Á. Fernández-Blázquez, “Selecting the most important self-assessed features for predicting conversion to mild cognitive impairment with random forest and permutation-based methods,” *Scientific Reports 2020 10:1*, vol. 10, no. 1, pp. 1–15, Nov. 2020, doi: 10.1038/s41598-020-77296-4.

[91] G. I. Webb, “Overfitting,” *Encyclopedia of Machine Learning*, pp. 744–744, 2011, doi: 10.1007/978-0-387-30164-8_623.

[92] G. Naidu, T. Zuva, and E. M. Sibanda, “A Review of Evaluation Metrics in Machine Learning Algorithms,” *Lecture Notes in Networks and Systems*, vol. 724 LNNS, pp. 15–25, 2023, doi: 10.1007/978-3-031-35314-7_2/FIGURES/3.

[93] “Biosemi EEG ECG EMG BSPM NEURO amplifier electrodes.” Accessed: Jan. 08, 2025. [Online]. Available: https://www.biosemi.com/Products_ActiveTwo.htm

[94] “Biosemi EEG ECG EMG BSPM NEURO amplifiers systems.” Accessed: Jan. 09, 2025. [Online]. Available: https://www.biosemi.com/faq/trigger_signals.htm

[95] “MMBT-S Trigger Interface Box | NEUROSPEC AG Research Neurosciences.” Accessed: Feb. 08, 2025. [Online]. Available: https://www.neurospec.com/Products/Details/1067/mmbt-s-trigger-interface-box

[96] “ActiveTwo System Operating Guidelines”.

[97] J. Peirce *et al.*, “PsychoPy2: Experiments in behavior made easy,” *Behav Res Methods*, vol. 51, no. 1, pp. 195–203, Feb. 2019, doi: 10.3758/S13428-018-01193-Y/FIGURES/3.

[98] “ActiView LabView Acquisition.” Accessed: Jan. 09, 2025. [Online]. Available: https://www.biosemi.com/software_biosemi_acquisition.htm

[99] A. Gramfort *et al.*, “MEG and EEG data analysis with MNE-Python,” *Front Neurosci*, no. 7 DEC, 2013, doi: 10.3389/FNINS.2013.00267.

[100] L. Cabañero-Gomez, R. Hervas, I. Gonzalez, and L. Rodriguez-Benitez, “eeglib: A Python module for EEG feature extraction,” *SoftwareX*, vol. 15, p. 100745, Jul. 2021, doi: 10.1016/J.SOFTX.2021.100745.

[101] M. Arsalidou and M. J. Taylor, “Is 2+2=4? Meta-analyses of brain areas needed for numbers and calculations,” *Neuroimage*, vol. 54, no. 3, pp. 2382–2393, Feb. 2011, doi: 10.1016/J.NEUROIMAGE.2010.10.009.

[102] U. Gupta and R. Z. Zheng, “Cognitive Load in Solving Mathematics Problems: Validating the Role of Motivation and the Interaction Among Prior Knowledge, Worked Examples, and Task Difficulty,” *European Journal of STEM Education*, vol. 2020, no. 1, p. 5, 2020, doi: 10.20897/ejsteme/9252.

[103] S. G. Hart and L. E. Staveland, “Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research,” *Advances in Psychology*, vol. 52, no. C, pp. 139–183, Jan. 1988, doi: 10.1016/S0166-4115(08)62386-9.

Figures

[104] A Novel Approach to Image EEG Sleep Data for Improving Quality of Life in Patients Suffering From Brain Injuries Using DreamDiffusion - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Shows-the-different-regions-of-the-brain-by-coloration-This-includes-the-frontal_fig1_381960918 [accessed 13 Feb 2025]

[105] Shriram, Revati & Sundhararajan, Mahalingam & Daimiwal, Nivedita. (2012). EEG Based Cognitive Workload Assessment for Maximum Efficiency. IOSR Journal of Electronics and Communication Engineering (IOSR-JECE).

[106] Oscillatory electroencephalographic patterns of arithmetic problem solving in fourth graders - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Schematic-configuration-of-32-EEG-channels-arranged-in-an-extended-10-20-layout-Colors_fig2_356732959 [accessed 13 Feb 2025]

[107] Han, Y.M.Y., Chan, MC., Chan, M.M.Y. *et al.* Effects of working memory load on frontal connectivity in children with autism spectrum disorder: a fNIRS study. *Sci Rep* **12**, 1522 (2022). https://doi.org/10.1038/s41598-022-05432-3

10. Appendix

10.1 Participant Consent Form

Consent Form

Title of the Study: *Assessing Cognitive Load in an Arithmetic Task Using Brain Wave Signals*

Researcher Information:

- **Name**: Lydia Exarchou
- **Institution**: University of Patras
- **Contact Information**: lydia.exarchou@gmail.com

Purpose of the Study:

The purpose of this study is to investigate cognitive performance during tasks requiring memory and problem-solving skills, and also assess the efficacy of the brain wave classifier, calibrated using an n-back task, in identifying cognitive load levels during arithmetic tasks.

Procedure:

1. **N-Back Calibration Task**:
   You will perform a memory-based task where numbers appear on the screen. You will be required to identify numbers based on specific rules (e.g., whether the number matches the one shown earlier). This task helps calibrate the pre-existing classifier.
2. **Arithmetic Task**:
   After the calibration, you will complete a series of arithmetic problems. You will solve these problems mentally and provide your answers on screen.

- The total duration of the study is approximately 50 minutes.
- Instructions for each task will be provided before the task begins.

Voluntary Participation:

Your participation in this study is entirely voluntary. You may withdraw at any point during the experiment without any penalty or loss of benefits.

Confidentiality:

All information collected during the study will be anonymized and treated with strict confidentiality. Any data obtained will only be used for research purposes and will not include your name or identifying information. Results will be reported in a way that ensures individual anonymity.

Risks and Benefits:

- **Risks**: The task is not expected to pose any significant risks. It should be noted that the tasks involve mental effort, which may cause mild fatigue. You are free to take breaks or stop participating if you feel any discomfort.
- **Benefits**: While there are no direct personal benefits, your participation will contribute to research on cognitive processes, including memory, attention, and problem-solving.

Contact for Questions:

If you have any questions about this study, you may contact the researcher at the provided details above.

Consent:

By signing this form, you confirm that:

1. You have read and understood the information provided above.
2. You voluntarily agree to participate in this study.
3. You understand that you can withdraw at any time without consequences.
4. You understand your data will be anonymized and kept confidential.

Participant Information:

- **Name**: _____
- **Signature**: _____
- **Date**: _____

Researcher Confirmation:

- **Name**: _____
- **Signature**: _____
- **Date**: _____

10.2 Pre-Experiment and Post-Experiment Questionnaires

Pre-Experiment Questionnaire

Demographics and Background

1. Age: _____
2. Gender: _____
3. Do you have an academic background involving math/calculus? (Yes/No)
4. What is your dominant hand? (Right/Left)

Health and Cognitive State

4. Do you have normal or corrected-to-normal vision? (Yes/No)
5. Do you have a history of neurological disorders (e.g., epilepsy, migraines)? (Yes/No)
6. Are you currently experiencing any physical discomfort or mental fatigue? (Yes/No)

Lifestyle Factors

7. How many hours of sleep did you get last night? _____
8. Have you consumed caffeine in the past 4 hours? (Yes/No)
    - If yes, how much? (e.g., one cup of coffee, energy drink): _____
9. Do you currently feel alert and focused? (Yes/No)
10. Have you engaged in any mentally taxing activities today (e.g., exams, intensive work)? (Yes/No)
    - If yes, please describe: _____

Familiarity with Tasks

12. Have you previously participated in experiments involving arithmetic or n-back tasks? (Yes/No)
13. Do you feel confident using a keyboard for quick responses? (Yes/No)
14. Have you experienced cognitive fatigue during similar experiments? (Yes/No)

Consent and Understanding

15. Have you read and understood the consent form? (Yes/No)
16. Do you agree to participate and understand that you can withdraw at any time? (Yes/No)
17. Do you have any questions about the experiment procedure? (Open-ended)

Post-Experiment Questionnaire

For each question, rate on a scale of 1 to 10:
1 = Very Low and 10 = Very High

1. Mental Demand:

   How mentally demanding were the arithmetic tasks?

   (1) ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ (10)

2. Physical Demand:

   How physically demanding was the task (e.g., typing responses)?

   (1) ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ (10)

3. Temporal Demand:

   How hurried or rushed did you feel during the tasks?

   (1) ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ (10)

4. Performance:

   How successful do you feel you were in completing the tasks?

   (1 = Poor, 10 = Excellent)

   (1) ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ (10)

5. Effort:

   How hard did you have to work to accomplish your level of performance?

   (1) ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ (10)

6. Frustration:

   How insecure, discouraged, irritated, or stressed did you feel during the tasks?

   (1) ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ (10)

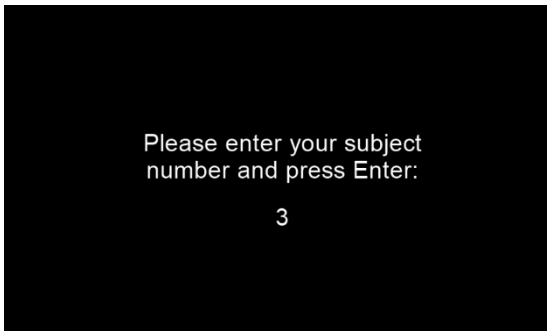
7. Did you find it difficult to stay focused throughout the experiment? (Yes/No)

   If yes, during which phase? _____

8. Do you feel the arithmetic tasks reflected real-world problem-solving situations? (Yes/No)
   If no, why not? _____

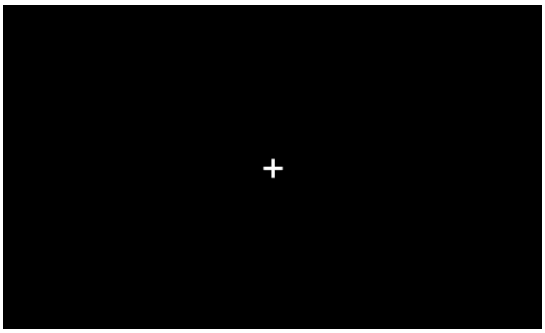
10.3 Task Interface

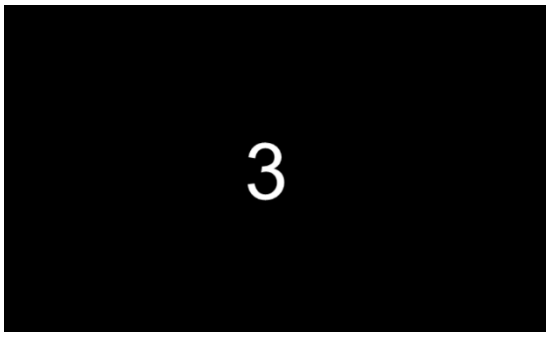
10.4.1 N-Back Task















10.3.2 Arithmetic Task

