POLITECNICO DI TORINO

MASTER's Degree in BIOMEDICAL ENGINEERING

MASTER's Degree Thesis

Signal Processing and Machine Learning Approaches for frontal EEG-Based Mental Workload Assessment: From Protocol Design to Classification

Supervisors

Prof. DANILO DEMARCHI

Ph.D. St. MARCO POGLIANO

Candidate

PALUSHI LORENZO

Academic year 2024-2025

Summary

Mental workload has become an important research topic, as increasing demands on multitasking and decision-making can impair performance and compromise safety in domains such as aviation, surgery, and driving. Understanding when and how workload arises enables the design of adaptive systems that monitor cognitive state and automate specific actions, improving safety and reducing fatigue.

To investigate cognitive variations, physiological measures such as ECG, respiration, electrodermal activity, and eye tracking are often used. In this work, frontal EEG was selected as the main modality, since brain electrical activity directly reflects cognitive processing. Eight frontal electrodes were used to minimize intrusiveness and simplify sensor placement. This thesis aimed to design an experimental protocol capable of eliciting distinct workload levels and to extract four feature groups (temporal, spectral, coherence, and complexity) from the processed EEG, which were then used to train machine learning classifiers for automatic workload recognition.

Nine task difficulty levels were defined within the Multi-Attribute Task Battery II (MATB-II) by manipulating event frequency. Eighteen participants completed a 5-minute rest phase followed by four MATB-II sessions combined with a secondary arithmetic task. Each session included five 2-minute task windows and a 20-second self-evaluation using the Bedford workload scale. EEG was recorded from eight frontal electrodes (F10, AF8, AF4, FP2, FP1, AF3, AF7, F9) with a g.HIAMP amplifier (g.tec). Two subjects were excluded due to corrupted data, yielding a final dataset of 16 participants.

EEG pre-processing included resampling from 1200 Hz to 512 Hz, a 0.5–80 Hz band-pass filter and 50 Hz notch filter. Artifacts such as blinks and abrupt movements were removed using envelope subtraction and amplitude thresholding: samples exceeding three standard deviations, or 0.5 second windows around artifacts, were discarded based on signal quality.

Each cleaned signal was segmented into 2, 3, and 4 second windows with 50% overlap. From each segment, 4 temporal, 15 spectral, 7 complexity, and 168 coherence features were extracted. Four statistical indices (mean, variance, skewness, kurtosis) were computed for each feature, except coherence, where only

mean and variance were used.

Bedford ratings, collected on a 10-point subjective workload scale, were grouped into three levels (low, medium, high) and used as class labels for the machine learning models. A Friedman test ($p<0.05$) confirmed significant differences in subjective workload across the easy, medium, and hard conditions, validating the effectiveness of the experimental manipulation. A classification analysis was performed on normalized features to identify the optimal combination of feature type and window size. Each classifier was optimized using five-fold cross-validation, with 80% of the data used for training and 20% reserved for testing.

In binary classification, both coherence features and complexity features yielded the highest performance, with an F1 score of 92.46% and an accuracy of 98.51% on the test set. These results, consistent across multiple classifier–feature selector combinations, highlight signal complexity and inter-channel connectivity as key workload indicators. No consistent trend emerged for window length. In multi-class classification, performance predictably decreased due to class imbalance and subjective ratings used as class labels.

Acknowledgements

I would like to express my gratitude to prof. Danilo Demarchi for allowing me to work on such a complex yet beautiful research topic. My sincere thanks goes also to Ph.D. student Marco Pogliano for accompanying and supervising me during these months; I truly appreciated your dedication.

Finally, I would like to thank my family, without whom I could not have accomplished all of this. Your sacrifices, support, and encouragement have been fundamental throughout my journey.

To my family

Table of Contents

List of Tables

List of Figures

Acronyms

MWL
Mental WorkLoad

CNS
Central Nervous System

PNS
Peripheral Nervous System

Nav
Sodium voltage-gated channel

Kv
Potassium voltage-gated channel

EEG
Electroencephalography

ECoG
Electrocorticography

VEOG
Vertical Electrooculogram

HEOG
Horizontal Electrooculogram

REOG
Radial Electrooculogram

CLT

Cognitive Load Theory

RSME

Rating Scale Mental Effort

NASA-TLX

NASA Task Load Index

ECG

Electrocardiography

HR

Heart Rate

HRV

Heart Rate Variability

EDA

Electrodermal Activity

fNIRS

functional Near-Infrared Spectroscopy

PPG

Photoplethysmography

EMG

Electromyography

MATB-II

Multi-Attribute Task Battery II

SYSMON

System Monitoring task

TRCK

Tracking task

COMM

Communication task

RESMAN

Resource Management task

FIR

Finite Impulse Response

IIR

Infinite Impulse Response

PSD

Power Spectral Density

RMS

Root Mean Square

LR

Logistic Regression

SVC

Support Vector Machine

KNN

K-Nearest Neighbors

LDA

Linear Discriminant Analysis

DT

Decision Tree

RF

Random Forest

BRF

Balanced Random Forest

MLP

Multi-Layer Perceptron

PCA

Principal Component Analysis

LDADR

Linear Discriminant Analysis for Dimensionality Reduction

SMOTE

Synthetic Minority Oversampling Technique

KW

Kruskal Wallis

GNB

Gaussian Naive Bayes

MNB

Multinomial Naive Bayes

BNB

Bernoulli Naive Bayes

Chapter 1

Introduction

The rapid evolution of modern technology has profoundly changed the way the human brain manages multiple concurrent tasks. The increasing interaction between humans and complex technological systems has raised the cognitive demands required in many activities, stimulating research on how mental workload (MWL) and stress manifest and affect performance. Understanding these phenomena is essential for developing technologies that can assist operators and enhance safety and efficiency.

Although mental workload and stress are two concepts highly correlated, they are not interchangeable concepts. Mental workload is the demand placed on an operator's mental resources used for attention, perception, reasonable decision-making and action. In contrast, stress can be described as a discrepancy between the external demands placed on the subject and his ability to cope with them [1].

In many safety-critical fields, such as aviation, surgery and industrial operations, maintaining an appropriate level of performance is essential. In fact, when workload becomes too high (overload) the error rate increases, leading to critical situations. For this reason, understanding how workload manifests and how can be controlled is a central topic in human safety and ergonomics.

Developing a clear understanding of mental workload and how it can be measured is crucial for designing adaptive human–machine systems capable of modifying their behavior in response to the operator's state. In aviation, for example, adaptive automation has been explored as a means to track a pilot's workload and provide timely assistance, helping to avoid performance degradation during highly demanding phases of flight [2].

In this context, research on mental workload supports the development of safer and more efficient technological systems.

Current approaches to assess cognitive load can be broadly classified into three main categories. The first category comprises subjective rating scales, in which operators are asked to report their perceived workload on predefined scales. When

properly designed, these instruments can achieve good sensitivity and reliability; however, their accuracy may be affected by the respondent's self-awareness, motivation, and honesty, making them susceptible to different forms of bias.

The second one includes performance-based methods, in which workload is evaluated based on parameters such as error rate, reaction times and other factors related to task completion. These measures allow to capture the relation between increased workload and performance. However, this type of measure may not catch accurate features that describe cognitive load behavior.

The last category includes physiological measures, in which mental workload is evaluated through the monitoring of some vital parameters. Techniques such as electrocardiography (ECG), electrodermal activity (EDA) or electroencephalography (EEG) provide continuous and objective information about the operator's state, allowing real-time assessment.

In this thesis, a novel multimodal approach combining subjective and physiological measures was adopted. Specifically, frontal EEG activity was selected as the main source of workload-related information, given that variations in electrical brain activity are directly linked to changes in cognitive state. The exclusive use of frontal and prefrontal electrodes was motivated by the intent to design a configuration suitable for practical applications in operational environments, minimizing discomfort and interference with the subject's hair.

In addition, the Bedford Workload Scale was employed to record subjective scores from the participants. The self-evaluations were used as ground truth, in order to validate the measures conducted using EEG data.

A novel experimental protocol was designed to elicit efficiently different MWL states. The Multi-Attribute Task Battery II (MATB-II), in combination with an arithmetical secondary task, served as workload inductor in the study. MATB-II was chosen for its multitasking structure, comparable to the one faced by aircraft pilots. By changing the event rate and duration of MATB-II, different MWL conditions were generated, allowing their measurement using EEG data.

EEG recordings collected during the experiment were processed using Visual Studio Code and MATLAB. A customized signal-processing pipeline was developed to clean the raw data and extract relevant features from both the time and frequency domains. These features were then used to train several machine learning models, with the aim of evaluating whether different workload states could be accurately distinguished based solely on EEG activity.

Finally, the analysis aimed to determine whether mental workload could be distinguished not only among different task intensities but also from the resting state, thereby establishing a foundation for future developments in real-time workload monitoring.

The development of a strong experimental protocol and the evidence of correlation between changes in EEG features with workload changes represent an

important starting point for adaptive human–machine interaction systems.

The following sections first introduce the relevant background, including an overview of brain anatomy, the fundamentals of EEG, and the theoretical framework of mental workload. The subsequent chapters describe the experimental setup, data acquisition procedures, and signal-processing pipeline, as well as the classification methods employed in this work. Finally, the results are presented and discussed in relation to existing literature, highlighting the main findings and their implications.

Chapter 2

Background

Understanding the mechanisms underlying human cognition and behavior requires a solid grasp of how the nervous system operates and how brain activity can be objectively measured. This chapter provides an overview of the anatomical and physiological foundations of the nervous system, focusing on the structures and processes responsible for information transmission and integration. Particular attention is devoted to electroencephalography (EEG), one of the most widely used and accessible techniques for capturing the brain's electrical dynamics with high temporal resolution.

Together, the concept of mental workload (MWL) is described, as a center topic in human performance, attention and decision-making. MWL is the cognitive demand placed on operator's mental resources used for task completion. Correlating MWL and physiological measures, such as EEG enables an objective assessment of cognitive states, allowing to develop accurate brain-machine interfaces.

The following sections described shortly the structure and functions of the nervous system, the main characteristics and principles of EEG, and finally introduces the MWL problem and the methodologies used to assess it.

2.1 The nervous system

The nervous system is the major controlling, regulatory, and communicating system in the body. It is the center of all mental activity, including thought, learning, and memory. The nervous system, in conjunction with the endocrine system, plays a key role in regulating and maintaining homeostasis. Through its receptors, the nervous system keeps us in touch with our environment, both external and internal [3]. From the anatomical point of view, the nervous system is divided into the central nervous system (CNS) and peripheral nervous system (PNS), as illustrated in Figure 2.1. The CNS includes the brain and spinal cord, while the

PNS consists of all nerves, ganglia, and sensory receptors outside the CNS. The CNS's responsibilities include receiving, processing, and responding to sensory information [4].

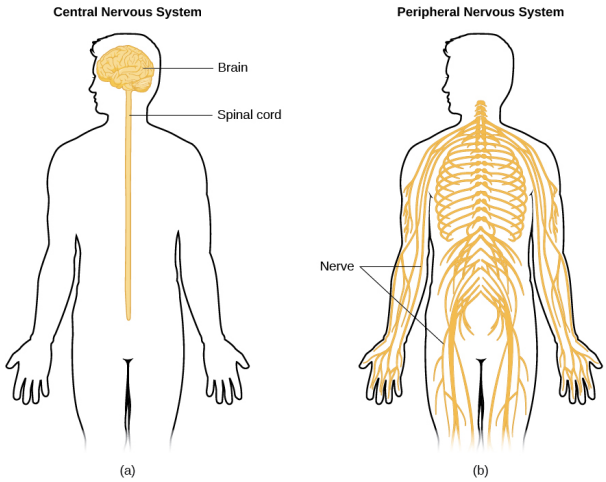


Figure 2.1: CNS and PNS composition.

2.1.1 Brain

The brain is the primary organ of the CNS and represents the most complex structure in the human body, containing billions of information-processing cells called neurons. It is entirely enclosed and protected within the skull. The brain is responsible for a wide range of functions, including learning, memory, perception, and the initiation of voluntary movements. Anatomically, the brain consists on three components: cerebrum, cerebellum and brainstem [5], as showed in Figure 2.2.

The cerebrum is the largest part of the brain and it is divided into the left and right hemispheres. Both hemispheres are composed of an outer layer of gray matter called cerebral cortex and an inner subcortical white matter. The cerebrum is involved in processing sensory inputs, controlling movement, language, emotions, learning and reasoning.

The cerebellum is a bilaterally symmetrical structure, with a cortex situated on the outside and nuclei on the inside. The cerebellum's primary function is to modulate motor coordination, posture, and balance [6].

The brainstem is the most caudal portion of the brain, and contains ten of the twelve pairs of cranial nerves, which are peripheral nerves that propagate directly from the brain and not from the spinal cord. In the brainstem there is also the

reticular formation, a network that plays a key role in regulating the sleep-wake cycle, cortical arousal, and states of consciousness [6].

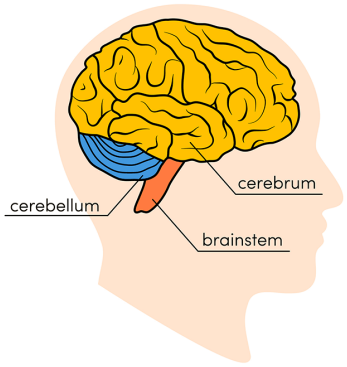


Figure 2.2: Cerebrum, cerebellum and brainstem location.

2.1.2 Cerebral cortex

The cerebral cortex is the outer layer of the cerebrum and represents the most evolutionarily advanced region of the brain, as well as the last to develop. It is composed by folds and convolutions, with the ridges between the convolutions called gyri and the valleys between the gyri called sulci [5]. This structural organization increases the surface area of the cortex, allowing a greater volume of gray matter to be accommodated within the limited space of the skull.

The cerebral cortex enables the perception of the surrounding environment, the formulation of thoughts, the expression of emotions, and the recollection of past events. It is also the region from which all motor commands for voluntary movements originate [6]. To perform these tasks, each hemisphere is divided into four regions called lobes (Figure 2.3):

- Frontal lobe;
- Parietal lobe;
- Occipital lobe;
- Temporal lobe.

Within each lobe, the cerebral cortex is further divided into regions that support specific functions. A clear example of this organization is the topographic arrangement of the primary motor and somatosensory cortices. Their somatotopic maps,

known as the motor and sensory homunculi, illustrate the correspondence between different body parts and distinct cortical areas. In these representations, the size assigned to each body region reflects the amount of cortical resources dedicated to it, as shown in Figure 2.4.

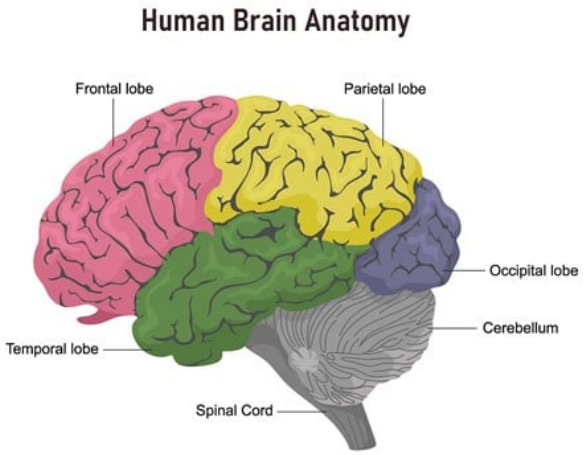


Figure 2.3: Lateral view of brain lobes

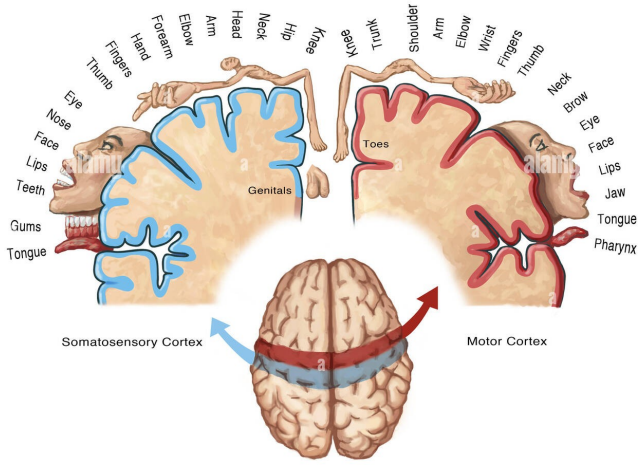


Figure 2.4: Motor and sensory homunculi from the transversal section of the primary motor and sensory cortex.

2.1.3 Neuron

Neurons (also called nerve cells) are the fundamental units of the brain and nervous system, the cells responsible for receiving sensory input from the external world, for sending motor commands to our muscles, and for transforming and relaying the electrical signals at every step in between [7]. Alongside neurons, the brain also contains glial cells, which represent the second major class of cells in the nervous system. These cells do not conduct electrical impulses but play crucial supportive roles, including maintaining homeostasis, forming myelin, and providing structural and metabolic support to neurons. A typical neuron comprises four morphologically distinct regions: the cell body (soma), dendrites, the axon, and the presynaptic terminals. Each of these regions has a distinct role in the generation and transmission of neural signals.

The cell body acts as the metabolic center of the neuron. It contains the nucleus, which holds the cell's genetic material, as well as the endoplasmic reticulum, where proteins essential for neuronal function are synthesized. Extending from the soma are the dendrites and, in most cases, a single axon. The dendrites receive incoming signals, while the axon is a long, tubular projection that carries electrical impulses to downstream neurons.

The axon originates from the cell body and carries action potentials, the fundamental electrical signals of the nervous system, which are initiated at a specialized region called the axon initial segment. These signals travel without loss of strength or distortion at speeds between 1 and 100 meters per second due to the all-or-none nature of action potentials, which are regenerated along the axon. To increase conduction speed, many axons are wrapped in a myelin sheath, a lipid-rich insulating layer generated by the Schwann cells. This sheath is periodically interrupted by nodes of Ranvier, which are essential for the regeneration of the action potential, enabling rapid and efficient signal propagation through a mechanism known as saltatory conduction [8]. A schematic overview is illustrated in Figure 2.5.

2.1.4 Synapses

Synapses are specialized junctions that mediate information transfer between neurons, typically from the axon terminal of a presynaptic neuron to the dendrite of a postsynaptic neuron. However, synapses can also occur between axon and soma, between two axons, or even between two dendrites. Synapses are broadly classified into electrical and chemical types [9].

Electrical synapses (Figure 2.6(a)) are direct cytoplasmic connections formed by protein channels between adjacent neurons, allowing for almost instantaneous transmission of electrical signals with minimal delay. Though relatively rare, they are prevalent in certain brain regions such as the thalamus and support synchronous activity. These synapses can also undergo forms of plasticity.

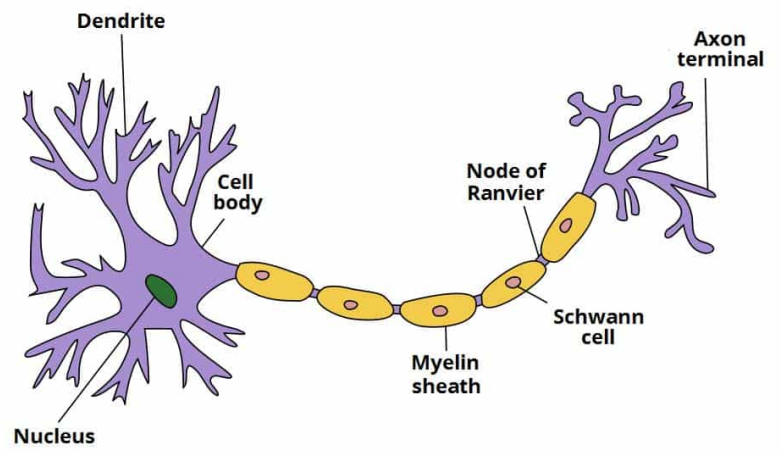


Figure 2.5: Anatomy of the neuron.

Chemical synapses (Figure 2.6(b)), the more common type, consist of a presynaptic terminal, a synaptic cleft, and a postsynaptic membrane. Neurotransmitters stored in vesicles within the presynaptic terminal are released in response to action potentials, triggered by calcium ion influx. These neurotransmitters then diffuse across the cleft and bind to specific receptors on the postsynaptic membrane, eliciting either fast (ionotropic) or slow (metabotropic) responses.

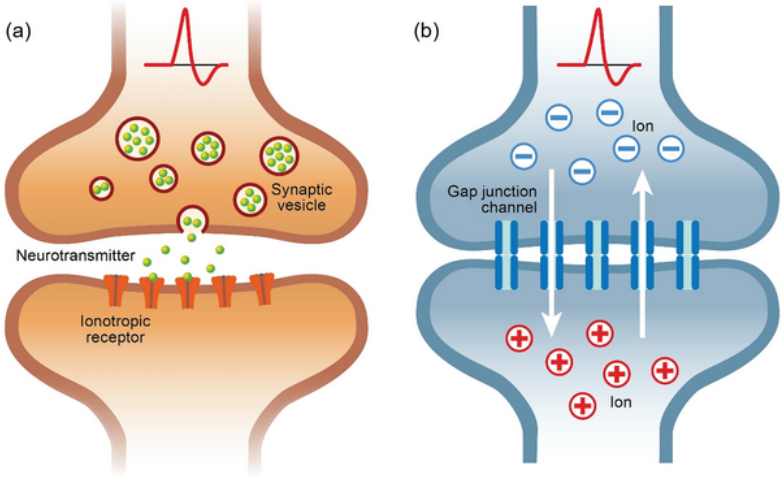


Figure 2.6: Basic structure and operation of chemical (a) and electrical synapses (b) [10].

2.1.5 Action potential

An action potential is a rapid sequence of changes in the voltage across a membrane. The membrane voltage, or potential, is determined at any time by the relative ratio of ions, extracellular to intracellular, and the permeability of each ion. In neurons, the rapid rise in potential, depolarization, is an all-or-nothing event that is initiated by the opening of sodium ion channels within the plasma membrane. The subsequent return to resting potential, repolarization, is mediated by the opening of potassium ion channels [11].

A neuronal action potential can be described in three consecutive phases: depolarization, repolarization, and hyperpolarization. Depolarization begins when the membrane potential reaches the threshold level, triggering the opening of voltage-gated sodium channels (Nav). The resulting influx of sodium ions further depolarizes the membrane and creates a positive feedback effect that drives the opening of additional Nav channels. In mature neurons, this phase lasts roughly 1 ms, after which the sodium channels enter an inactivated state and are temporarily unable to conduct ions.

Repolarization follows with the activation of voltage-gated potassium channels (Kv). Although they are sensitive to a threshold similar to Nav channels, Kv channels exhibit slower kinetics. Consequently, their opening occurs approximately when Nav channels are already inactivated. The outward flow of potassium ions drives the membrane potential back toward resting values. Due to their delayed closure, Kv channels remain open slightly longer than required, producing a transient hyperpolarization, in which the membrane potential falls below its resting level. While the channel remains open when the cell is above the threshold voltage, the channel is said to be inactivated because it does not allow ion movement. Therefore, following each action potential the cell has an absolute refractory period in which Nav are inactivated and cannot be recruited to induce another action potential. The process is illustrated in Figure 2.7

The propagation of the action potential depends on whether the axon is myelinated or unmyelinated. In myelinated axons, the insulating myelin sheath restricts ion flow, concentrating depolarization events at the nodes of Ranvier. This mechanism, known as saltatory conduction, allows the signal to jump from node to node, thereby increasing conduction velocity by more than an order of magnitude compared to unmyelinated axons. In contrast, in unmyelinated fibers, the depolarization must spread continuously to adjacent membrane regions, generating a slower wave of excitation.

Action potentials typically originate at the axon hillock, where excitatory inputs are integrated and the threshold for activation is most easily reached. In sensory neurons, however, initiation often takes place at the distal terminals of the axon [11].

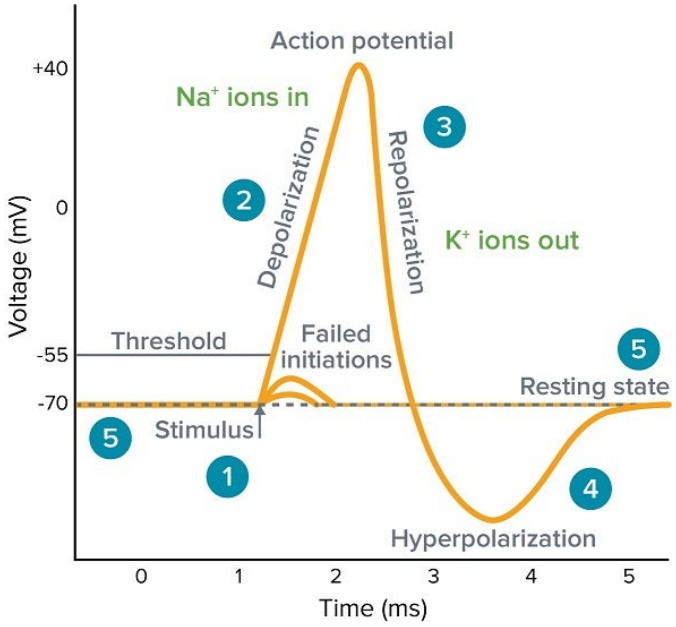


Figure 2.7: Schematic representation of the action potential,

2.2 Electroencephalography

Electroencephalography (EEG) is a record of the electric signal generated by the cooperative action of brain cells, or more precisely, the time course of extracellular field potentials generated by their synchronous action. The term electroencephalography originates from the Greek words *enkephalos* (brain) and *graphein* (to write) [12].

EEG signals can be recorded using electrodes placed either on the scalp or directly on the cortical surface. In the latter case, the technique is referred to as electrocorticography (ECoG). When electric fields are measured intracortically, they are known as local field potentials (LFPs). EEG activity recorded in the absence of external stimuli is termed spontaneous EEG, whereas EEG responses elicited by external or internal stimuli are referred to as event-related potentials (ERPs). EEG has multiple applications, including the diagnosis of epileptic seizures, monitoring of anesthesia during surgery, assessment of brain lesions, and investigation of sleep disorders. It is also employed in the evaluation of neurological and psychiatric conditions such as Alzheimer's disease, Parkinson's disease, schizophrenia, and depression. Furthermore, EEG is used to monitor brain activity in patients admitted to intensive care units and in the application of therapeutic interventions such as deep brain stimulation.

Richard Caton (1842–1926), an English scientist, is credited with discovering

the electrical properties of the brain, by recording electrical activity from the brains of animals using a sensitive galvanometer, noting fluctuations in activity during sleep and absence of activity following death. Hans Berger (1873–1941), a German psychiatrist, recorded the first human EEGs in 1924 [13].

2.2.1 EEG characteristics

The EEG signal exhibits notable characteristics that can be observed in the time domain:

- **Amplitude**: variable between 10 and 500 μV, and can be distinguished in low (<30 μV), medium (30-70 μV) and high (>70 μV) amplitude.
- **Morphology**: represents the way in which a repetitive signal with a dominant frequency manifests and can be classified as either polymorphic or monomorphic. A polymorphic signal consists of a sequence of potentials within the same frequency band but with irregular periodicity and often varying amplitudes. In contrast, a monomorphic signal consists of a sequence of potentials with perfectly regular periodicity and, frequently, identical amplitudes.
- **Topography**: defines the cerebral areas where an electrical event takes place and can be identified with reference to the distinction between lobes (frontal, parietal, occipital or temporal) and hemispheres (left or right).
- **Symmetry/Asymmentry**: signals are considered symmetric if they manifest in both hemispheres with the same frequency, amplitude and duration, even if they occur at different times. Signals are considered asymmetric if they appear only in one hemisphere or, when bilateral, exhibit different characteristics.
- **Synchrony/Asynchrony**: signals are considered synchronous if they appear simultaneously in both hemispheres, and asynchronous if they appear in both hemispheres at different times.

2.2.2 EEG rhythms

EEG signals exhibit distinct components in the frequency domain that have been associated with different physiological or cognitive states. While they are primarily used in sleep analysis, they also hold great potential for applications in other areas, such as mental workload assessment and stress classification. These components, also called EEG rhythms, are shown in Table 2.1. Figure 2.8, instead, depict the primary five waveforms of the different bands.

Rhythm	Frequency (Hz)	Amplitude (μV)	Characteristics
delta (δ)	0.5-4	20-200	Pathological conditions, deep sleep
theta (θ)	4-8	5-100	Falling asleep, light sleep
alpha (α)	8-12	10-200	Mental relax
beta (β)	12-30	1-20	Attention, concentration
gamma (γ)	>30	1-20	Learning process, high concentration and memory tasks

Table 2.1: EEG rhythms and their associated status.

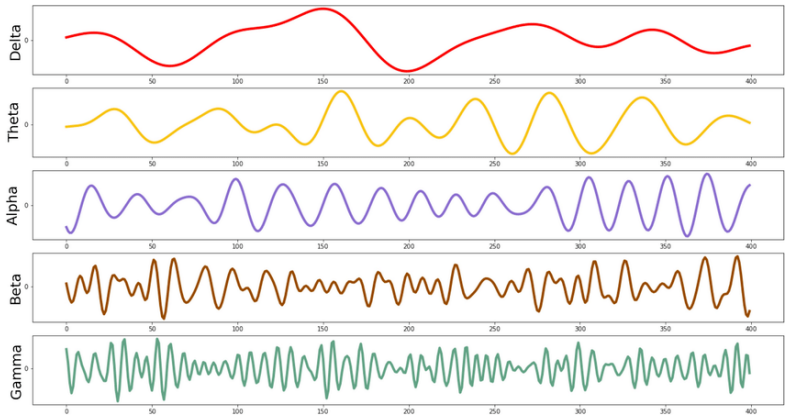


Figure 2.8: EEG signal decomposed into its frequency bands. From the top: δ (0.5-4 Hz), θ (4-8 Hz), α (8-12 Hz), β (12-30 Hz), γ (>30 Hz) [14].

2.2.3 EEG recording standards and the International 10-20 System

EEG can be recorded invasively, for example through methods such as ECoG, or more commonly in a non-invasive manner using electrodes placed on the scalp. In this work, the latter approach is considered, as it enables the acquisition of brain activity without surgical intervention while maintaining adequate spatial and temporal resolution.

In scalp EEG, electrode–skin impedance plays a crucial role in defining the electrical properties of the recording interface. Lower impedance values generally lead to higher signal quality, as they enhance the transmission of the brain's electrical potentials to the recording system. Moreover, reduced electrode–skin

impedance minimizes power line interference and increases the robustness of EEG recordings against motion-related artifacts, including those generated by cable movements. There are two types of electrodes that can be used to record an EEG signal, as illustrated in Figure 2.9:

- **Wet/Gel electrodes**: exhibit the lowest electrode–skin impedance (typically below 5 kΩ). The most common type is the silver/silver chloride (Ag/AgCl) electrode, which uses an electrolyte gel to ensure good conductivity. Although considered the clinical gold standard, wet electrodes have notable drawbacks: the gel can dry out over time, adjacent electrodes may form conductive bridges, and both scalp preparation and gel application can be uncomfortable or inconvenient for the subject [15].
- **Dry electrodes**: have been introduced as an alternative to wet electrodes for long-term EEG recordings, as they avoid scalp preparation and gel application. They can be divided into contact, noncontact, and insulating types. Contact dry electrodes, which rest directly on the scalp, are the most common due to their lower impedance and better performance compared to other dry designs. However, the absence of conductive gel results in higher electrode–skin impedance (up to several MΩ at 50/60 Hz), making the signal more sensitive to noise, interference, and motion artifacts. Proper shielding is therefore required. Passive dry electrodes, connected to the amplifier through unshielded cables, often provide lower signal quality, so active dry electrodes (that features on-site amplification to reduce environmental noise) are generally preferred [15].

Up to 1947 there was a lack of uniformity in the positioning, numbering system and montages of electrodes on the scalp. In 1949, the first standardized system was presented at the 2nd International Congress of IFSECN in Paris, and published by Jasper in 1958. It is still universally used and known as the International 10-20 System (SI 10-20) [16].

The traditional 10–20 electrode placement system defines the positions of 19 EEG electrodes on the scalp, along with two additional electrodes located on the earlobes (A1/A2), in relation to specific anatomical landmarks. The system is based on distances corresponding to 10% or 20% of the total length between these landmarks (Fig. 3). The naming convention of each electrode derivation consists of two parts: the first indicates the row of the array from the front of the head (Fp, F, C, P, O, T), while the second specifies the hemisphere, with even numbers assigned to the left side, odd numbers to the right, and the central line denoted by 'z' (or '0') [12], as can be observed in Figure 2.10.

Progress in topographic representation of EEG recordings brought demand for a larger amount of derivations. Electrode sites halfway between those defined by the standard 10–20 system were introduced in the extended 10–20 system, as shown in Figure 2.11.

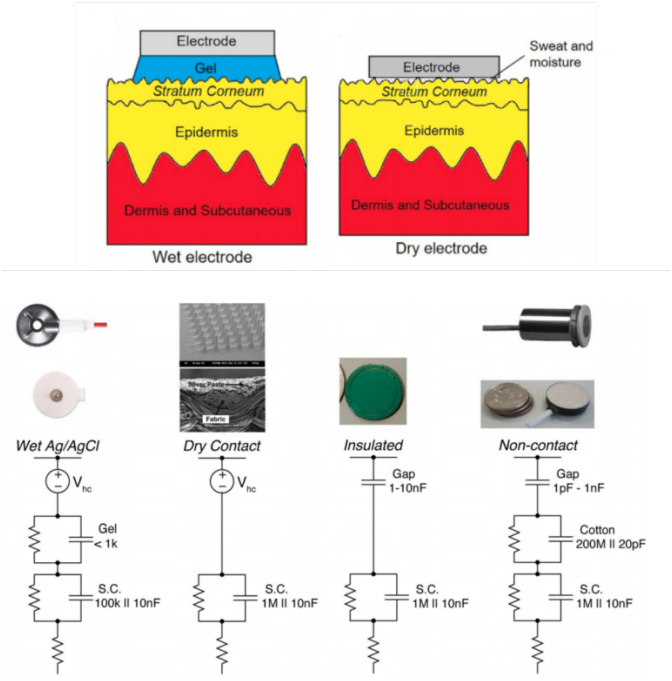
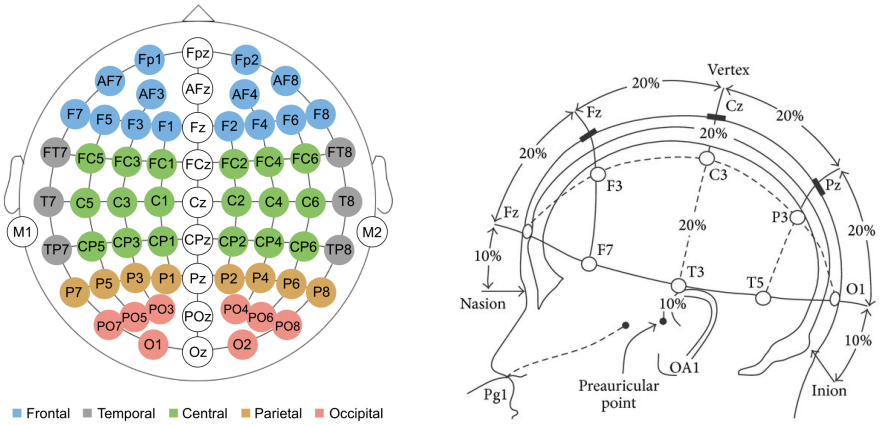


Figure 2.9: Top: comparison of wet (on the left) and dry (on the right) electrodes. Bottom: electrical model of the various electrode-skin interface [15]

.



(a) Frontal view of the International 10-20 System

(b) Lateral view of the International 10-20 System

Figure 2.10: Frontal and lateral views of the International 10-20 System.

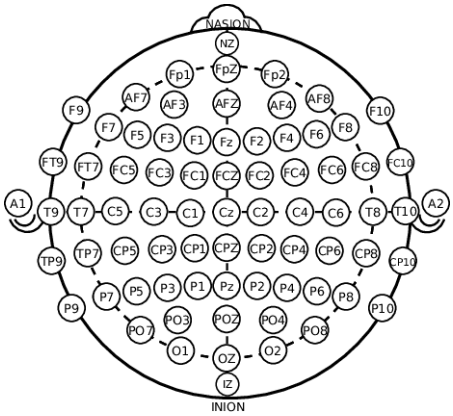


Figure 2.11: Frontal view of the extended 10-20 system.

2.2.4 Unipolar and bipolar configurations

Once the electrodes are positioned on the scalp, there are two possible approaches to perform the signal acquisition. The first one is monopolar (referential) acquisition, in which each electrode records the potential difference with respect to a common reference. The reference electrode is ideally electrically neutral and can be placed on the earlobe, mastoid, chin, or neck, although no universally accepted standard exists for its placement [12]. This configuration, showed in Figure 2.12(a) provides information on the absolute potential at different scalp sites and is particularly sensitive to widespread brain activity. However, it is also more susceptible to noise and artifacts if the reference is not truly inactive.

The second approach is bipolar acquisition (Figure 2.12(b)), where the signal is measured as the potential difference between two adjacent active electrodes. This configuration enhances the detection of local activity and helps reduce common noise, since artifacts shared by both electrodes tend to cancel out. A limitation of bipolar recordings is that they may obscure broader brain dynamics, as they primarily capture relative differences between neighboring sites rather than absolute activity patterns.

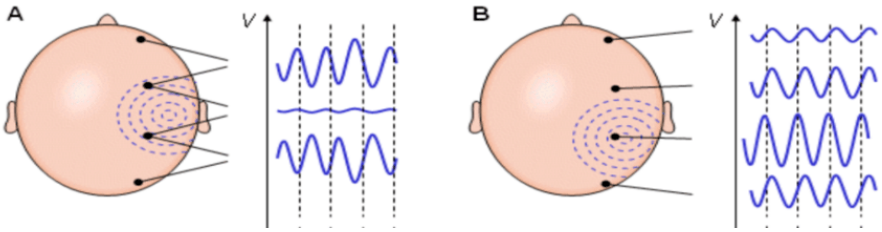


Figure 2.12: (a): bipolar configuration for EEG recordings. (b): monopolar configuration for EEG recordings [17].

2.2.5 EEG artifacts

Although the electroencephalograph is designed to record brain activity, it also captures unwanted signals originating from non-cerebral sources, known as artifacts. These artifacts act as noise that contaminate the detection of genuine neural activity, making it more difficult to extract the signal of interest. The main sources of EEG artifacts, as reported in [18, 19] are:

- **Ocular artifacts**: Electrical activity from eye movements and blinks, mainly detected by frontal electrodes; often monitored with vertical (VEOG), horizontal (HEOG) and radial (REOG) reference channels. Produces strong interference contaminating EEG, especially large-amplitude artifacts from blinks; depends on electrode proximity and eye movement direction and is a physiological source of interference.

- **Muscle artifacts**: Electrical activity on the scalp caused by the contraction of muscles (e.g., during swallowing, talking, walking). Involves variable shapes and amplitudes depending on muscle type and contraction level, overlapping with EEG activity and difficult to correct, due to its variability. It is a physiological source of noise.

- **Cardiac artifacts**:Electrical activity generated by the heart that can be detected on the scalp. Its amplitude depends on both electrode placement and the subject's body characteristics. These artifacts include electrocardiographic signals, which appear as regular waveforms corresponding to the heartbeat, and pulse artifacts, produced when electrodes are positioned over arteries. While these artifacts may sometimes resemble epileptiform activity, they are generally easier to identify and correct using a reference electrocardiographic waveform. Pulse artifacts, instead, are more difficult to remove due to their similarity to EEG signals, although they typically affect only a single electrode.

- **Motion artifacts**: Caused by subject movement altering electrode position or electrode–skin contact, which changes electrical coupling and conduction volume. Induce signal distortions and potential changes at recording sites. It is a physiological source of noise.

- **Power line artifacts**: Electromagnetic interference originating from the coupling with the electrical power grid (typically 50 or 60 Hz). It produces rhythmic noise overlapping with brain oscillations. This contamination is often shared across electrodes and can be difficult to remove without proper shielding, grounding, and filtering. It represents an environmental source of noise.

- **Electromagnetic interference (EMI)**: External electromagnetic fields generated by nearby electronic devices or medical equipment. These fields can induce broadband or narrowband noise that masks neural activity and distorts spectral analyses. The severity depends on the device type, its proximity, and the level of shielding. This is also an environmental noise source.

- **Electrode–skin impedance**: Quantifies the resistance to alternating current flow at the interface between the recording electrode and the scalp. Fluctuations in this impedance are primarily influenced by insufficient skin preparation, drying of the conductive gel and individual skin characteristics. Produces unstable impedance, manifested as baseline drifts and reduction in signal reliability. It is an experimental source of noise

- **Skin and other artifacts**: Includes minor artifacts such as perspiration artifacts (slow baseline shifts), sympathetic skin response (slow autonomic waves), and other interferences from tongue movements, dental restorations, breathing, and electrodermal activity. These are all physiological sources of noise.

Figures 2.13, 2.14, 2.15 2.16 and 2.17 represents examples of common EEG artifacts.

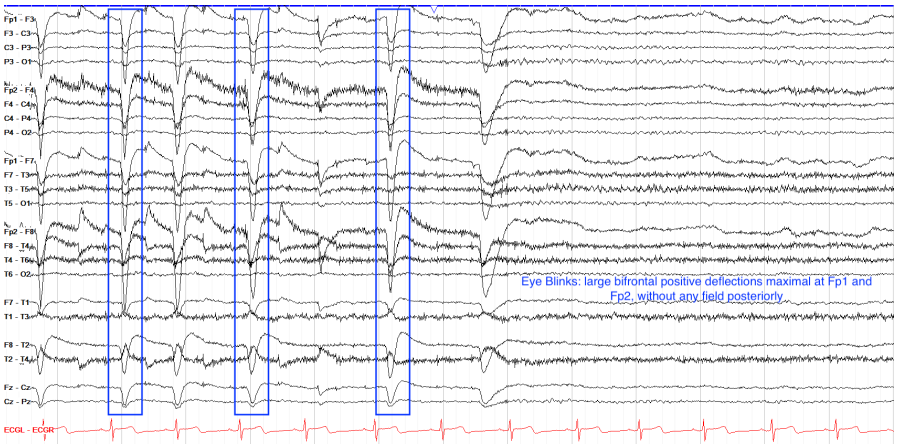


Figure 2.13: Example of eye blinks artifact [20].

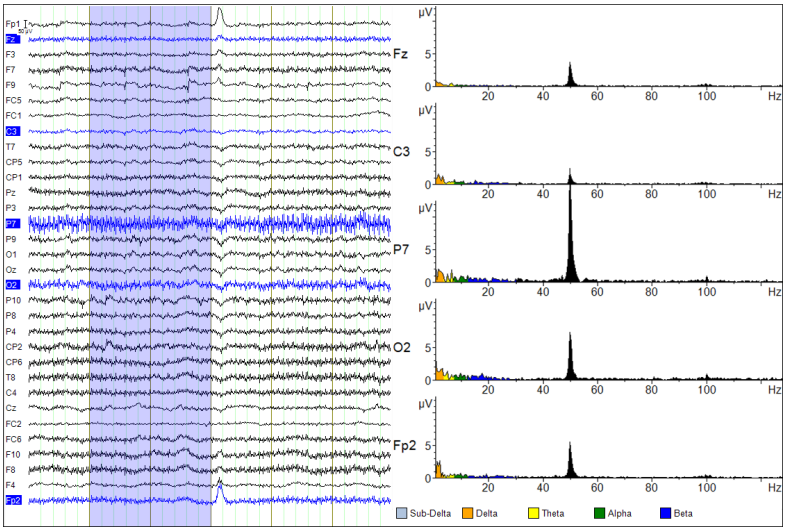


Figure 2.14: Example of power line artifact. Different channels are affected to a different degree, but the noise is present in all channels [21].

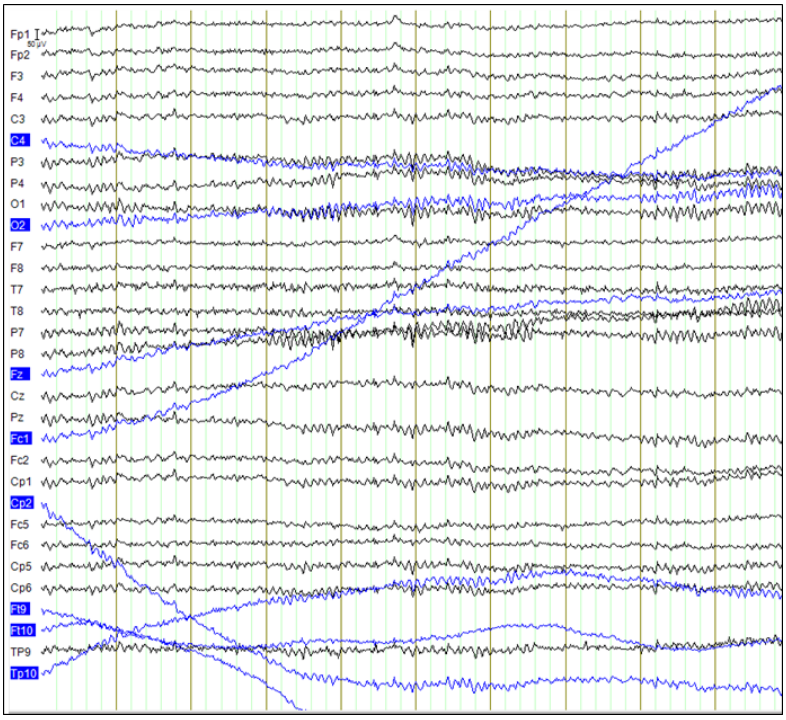


Figure 2.15: Example of the effect of electrode-skin impedance on the EEG signal, highlighted in blue [21].

Figure 2.16: Example of lateral eye movement artifact [20].

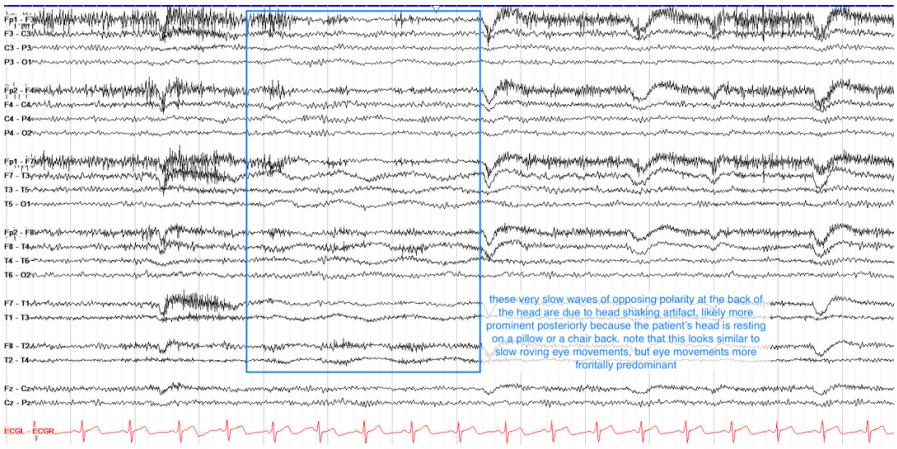


Figure 2.17: Example of head movement artifact and its effect on the EEG signal [20].

2.3 Mental Workload

2.3.1 Definition and characteristics

Mental workload (MWL) is a subjective measure that reflects the total cognitive, perceptual and physical demands placed on an individual during task performance. It encompasses factors such as task difficulty, time pressure, mental and physical effort, stress, fatigue and activity type (e.g., skill-based, rule-based or knowledge-based behaviors). It represents the perceived cost or load experienced by a person in accomplishing a task, influenced by the interaction between task requirements and the individual's skills, perceptions and circumstances [22], as illustrated in

Figure 2.18.

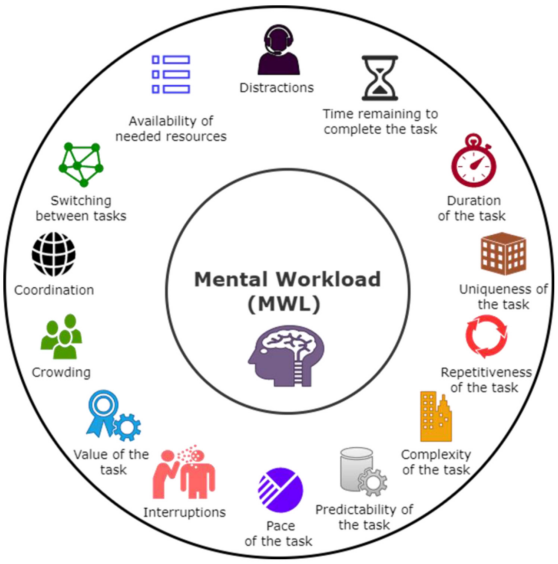


Figure 2.18: Example of factors that influence MWL.

Evaluation of MWL is based on the understanding of human cognitive processes and methodologies and the combination of situational demands with the cognitive efforts derived from the memory structure. Implementing the information processing model is the basis of the Cognitive Load Theory (CLT) [23]. CLT posits that human working memory has a limited capacity; when instructional tasks demand excessive cognitive resources, learning becomes less efficient [24]. Instructional design should therefore aim to manage and optimize this finite capacity to prevent overload. CLT distinguishes among three types of cognitive load ([23]):

- **Intrinsic Cognitive Load**: Arises from the inherent complexity of the material and the learner's existing knowledge base. It reflects the number of interacting elements a learner must process at once to build or adapt schemas [25, 26].

- **Extraneous Cognitive Load**: Results from suboptimal instructional design that forces learners to expend working memory resources on non-essential processing [27]. Because intrinsic and extraneous loads are additive, excessive extraneous load leaves fewer resources for meaningful learning [24].

- **Germane Cognitive Load**: Refers to the cognitive effort dedicated to schema construction, abstraction and elaboration (processes that directly contribute to meaningful learning) [25].

2.3.2 Relationship with performance

One of the main reasons to study MWL is its direct relationship with operator performance. This relationship is bidirectional: while reduced performance may indicate elevated MWL, performance failure itself can also increase subjective perceptions of workload [28]. The primary goal is to identify when workload becomes suboptimal, thereby increasing the likelihood of errors or incidents. Suboptimal workload may manifest as overload or underload [29].

Overload occurs when task demands exceed the operator's processing capacity, often impairing selective attention and resulting in inefficient information sampling. Conversely, underload arises when insufficient stimulation reduces attentional engagement, leading to lapses and performance decline. Both overload and underload are recognized as equally detrimental, each associated with degraded performance and heightened error risk [28, 30]. This dynamic is often represented by a U-shaped relationship between performance and MWL, as shown in Figure 2.19.

Operators may attempt to compensate for suboptimal workload by allocating additional cognitive resources. This strategy can temporarily help maintain performance, but it typically comes at the cost of increased mental strain. While such compensatory effort can counteract the effects of underload, it may also amplify the detrimental consequences of overload.

Given these dynamics, accurately measuring MWL is essential for understanding how it relates to performance. Reliable assessment makes it possible to identify workload thresholds that are associated with elevated error rates and a higher risk of accidents.

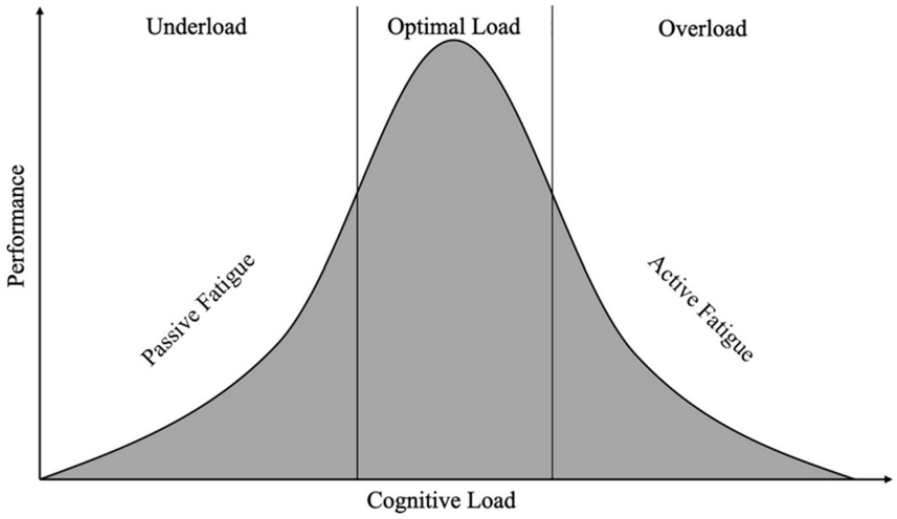


Figure 2.19: The relationship between mental workload and performance.

2.4 Measurements

The measurement of MWL is crucial from both scientific and applied perspectives. From a research standpoint, quantifying MWL enables the prediction of operator and system responses, supports the optimization of human-machine interactions and helps identify sources of error in order to improve performance across various domains, including medicine. From a human factors perspective, understanding and regulating MWL is equally critical for safeguarding well-being and mitigating the consequences of of excessive cognitive demands, such as stress and fatigue [31]. There are mainly three methods used to measure mental workload: self assessment or subjective rating scales, performance methods and physiological measures [32].

2.4.1 Subjective rating scales

Subjective rating scales are commonly employed to capture individuals' self-assessments of the workload they experience. These scales are specifically designed to reflect personal perceptions of task demands, exerted effort, fatigue or stress, thereby emphasizing the psychological dimension of workload. When properly constructed, such scales can achieve both sensitivity and reliability. Their widespread is largely attributed to their simplicity, cost-effectiveness and flexibility. Nevertheless, their validity is influenced by the respondent's level of self-awareness and honesty, and they remain vulnerable to various forms of bias. Finally, respondents need to know well the scale proposed. In fact, in many subjective scales, each level of the scale is shortly described, helping the subject by giving him all possible tools for the self-assessment duty. The main rating scales are the NASA Task Load Index (NASA-TLX), BedFord Workload Scale and the Rating Scale Mental Effort (RSME), that will be shortly described below.

NASA-TLX

The NASA-TLX (Figure 2.20), developed by Hart and Staveland in 1988, is the most widely applied subjective workload assessment tool, especially in aviation and air traffic control. The instrument measures workload across six dimensions, as can be seen in Figure 2.20 . Each dimension is rated on a 20 point bipolar scale, producing scores from 0 to 100. The combined ratings are assumed to represent an individual's overall workload. Originally developed through expensive laboratory research, the NASA-TLX has been validated in a wide range of operational and experimental contexts, including multitasking environments such as simulated and real flight, as well as air combat scenarios [33].

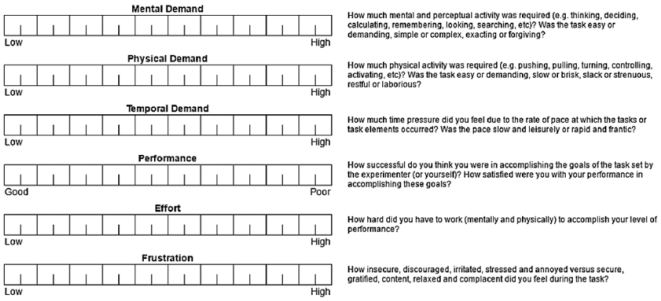


Figure 2.20: NASA-TLX questionnaire with a short description of each dimension.

BedFord Workload Scale

The BedFord scale (Figure 2.21) is a uni-dimensional workload assessment tool developed to evaluate an operator's spare mental capacity during task performance. Unlike multidimensional measures (e.g. NASA-TLX), it employs a hierarchical decision-tree structure to guide the operator through a ten-point rating scale, where each point is associated with a descriptive workload level. The process begins with the operator judging whether the task was possible to complete, whether the workload was tolerable, and whether performance was satisfactory without requiring workload reduction [34]. Based on these judgments, a workload score ranging from 'insignificant' (1) to 'task abandoned' (10) is assigned. This approach provides a simple yet structured means of assessing subjective workload [35],

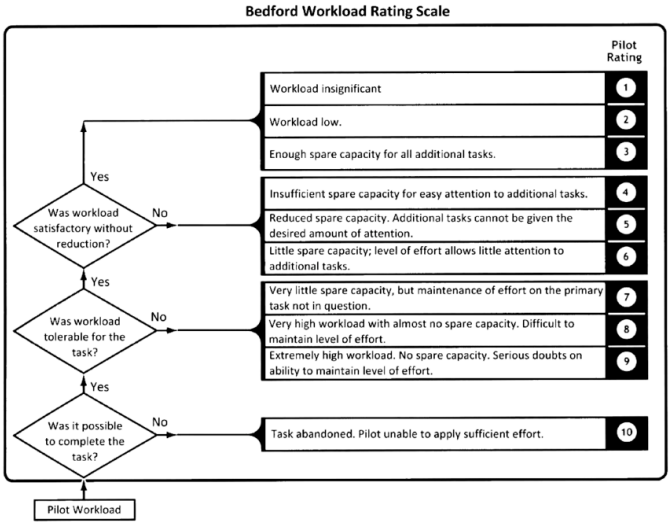


Figure 2.21: BedFord rating scale.

Rating Scale Mental Effort (RSME)

RSME (Figure 2.22) is a unidimensional instrument designed to assess perceived mental effort. It consists of a line marked with nine anchor points, each labeled to represent increasing levels of workload. The operator is able to see only the line, as the labels are hidden to prevent any bias on the rating. Conceptually, the RSME is comparable to the effort subscale of the NASA-TLX, as it relies on operators' ability and willingness to self-report their mental state. The scale is simple, cost-effective and requires no specialized equipment, making it suitable for workplace applications where quick responses are needed without disrupting ongoing tasks. Given the limitations observed in other workload indices, the RSME has been investigated as a practical tool for evaluating mental workload, including in healthcare contexts such as nursing [36].

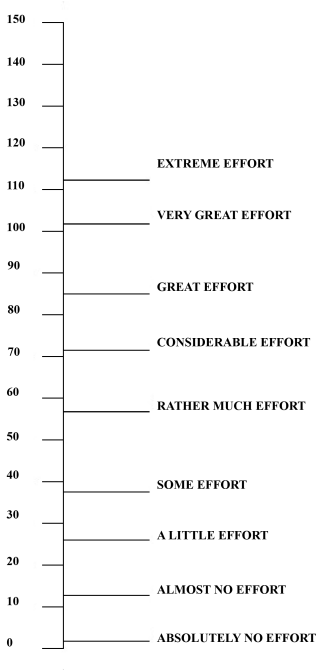


Figure 2.22: RSME rating scale.

2.4.2 Performance methods

Performance measures are employed to evaluate how effectively operators execute both primary and secondary tasks within workload assessments. Primary task measures focus on the accuracy and efficiency with which the main task is carried out, thereby reflecting the direct impact of workload on task performance and achievement. Secondary task measures, in contrast, performance on an additional

task that is deliberately introduced to impose further demand or reveal residual operator capacity. Together, these measures provide valuable diagnostic insights into how workload influences task execution, offering a more objective means of quantifying workload effects and enabling inferences about remaining capacity and the distribution of cognitive resources [32].

In practice, performance measures are typically carried out by evaluating factors like percentage of error in a specific task, response delay and other task-related indicators. However, it has been observed a lack of consistency between performance ratings and subjective ratings of cognitive load, difficulty, and effort when the two methods are used in combination.

2.4.3 Physiological measures

Physiological measures evaluate workload by monitoring physiological and neural responses associate with cognitive and emotional demands. Common techniques are, as illustrated in Figure 2.23:

- **Electrocardiography (ECG)**: records the electrical activity of the heart and provides valuable information about autonomic nervous system responses. Derived measures such as Heart Rate (HR) and Heart Rate Variability (HRV) are closely associated with variations in mental workload. In general, higher HR and reduced HRV are indicative of increased cognitive demand and mental effort [37].

- **Respiration**: monitors breathing activity, typically through measures such as respiration rate and amplitude. Variations in these parameters are associated with changes in cognitive and emotional states. An increase in respiration rate is often observed under higher mental workload, reflecting heightened metabolic and autonomic activation, although this relationship can vary depending on the nature of the task and the degree of physical involvement [38].

- **Electrodermal activity (EDA)**: reflects the electrical conductance of the skin, which varies with the activity of sweat glands controlled by the autonomic nervous system. In presence of negative emotional states, EDA levels typically increase and exhibit greater fluctuations, making it a reliable physiological indicator of emotional arousal [39].

- **functional Near-Infrared Spectroscopy (fNIRS)**: measures variations in the concentration of oxygenated and deoxygenated hemoglobin in the brain, thereby providing an indirect indicator of neural activity. Under conditions of stress or increased cognitive load, brain oxygenation levels undergo significant changes that can be detected through fNIRS signal[40].

- **EEG**: this technique is widely employed to assess variation in cognitive states during tasks. The most common approach is power spectral analysis, which decomposes EEG signals into frequency bands and quantifies synchronous neural oscillations. Changes in spectral power across specific frequency bands, observed during different tasks or over time, provide valuable insights into the neural mechanisms underlying cognitive processes. Overall, there are numerous EEG features used in the mental workload assessment field [23, 41, 42].
- **Eye tracking**; eye movements, such as saccades, fixations and pupil size variations are related to mental workload. Saccades are rapid shifts of gaze between fixation points, while fixations represent periods during which the gaze remains stable. An increase in saccadic activity, together with fluctuations in pupil size, is often associated with heightened cognitive load [43].
- **Other techniques**: there are a lot of minor biological indicators used to assess cognitive load, such as photoplethysmography (PPG), that measures changes of blood volume in the tissue and electromyography (EMG), that measures electrical activity of muscles.

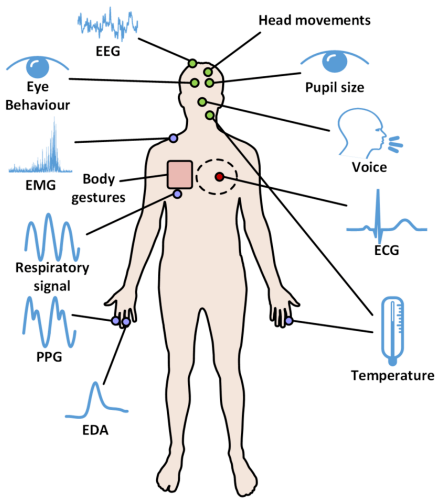


Figure 2.23: Physiological signals related to MWL.

These measures are particularly valuable as they provide objective, continuous, and real-time data that can capture dimensions of workload not easily detected through performance metrics. Moreover, they are capable of revealing subtle or unconscious changes in an operator's state that may occur prior to observable declines in task performance. As such, physiological measures serve as complementary approach to subjective rating scales and performance-based assessments in

the comprehensive evaluation of mental workload. For this study, EEG measures in combination with self assessment measures will be implemented to evaluate the MWL of a pool of subjects.

2.5 MWL Tests

In the literature, a wide range of tests has been developed to induce subjects into conditions of heightened attention, concentration and stress, thereby enabling the assessment and study of mental workload. These tests typically combine auditory and visual stimuli with tasks that require memory, multitasking, or simultaneous processing..

2.5.1 Arithmetical Tests

Mental arithmetical operations are commonly perceived as stressful and represent one of the simplest methods to induce cognitive load in experimental subjects. Performing a large number of operations over time requires sustained attention and active engagement of working memory. The level of workload can be modulated by the type of operation and the number of digits involved.

For instance, in the study of Al-Shargie et al. (2016) [44], three levels of difficulty were defined. The first level involved three one-digit integers combined with addition and subtraction. The second level introduced multiplication, while the third level required four integers ranging from 0 to 99 and included division as an additional operator.

The main advantages of these tasks are their simplicity and the minimal training required for participants. However, their one-dimensional structure does not accurately reflect real-world conditions. They can become monotonous, potentially reducing attention over time, and they fail to capture scenarios that involve multitasking or complex decision-making.

Although arithmetic tasks alone do not fully exploit their potential, they can be highly effective when integrated as secondary task with other cognitive workload tests.

2.5.2 N-back Test

N-back tasks are among the most commonly used paradigms for assessing and manipulating working memory capacity. In a typical n-back task, participants are presented with a continuous sequence of stimuli and are required to indicate whenever the current stimulus matches the one presented n steps earlier in the sequence. The main independent variable is the load factor n, which systematically

adjusts the difficulty of the task and the demands on working memory. Figure 2.24 illustrates an example of N-Back test.

A wide variety of stimulus types have been employed in n-back tasks, including letters, numbers, emotional words, faces, shapes, pictures and auditory tones. Although multiple cognitive processes contribute to task performance, the outcomes are largely independent of the specific stimuli used. Task difficulty can also be influenced by the introduction of lure stimuli, which are near matches appearing at positions adjacent to the correct n-back location [45]. Further increases in cognitive load can be achieved through dual n-back tasks, in which participants perform two n-back tasks simultaneously using different stimulus streams [46]. These variants are considered effective for taxing multiple components of working memory.

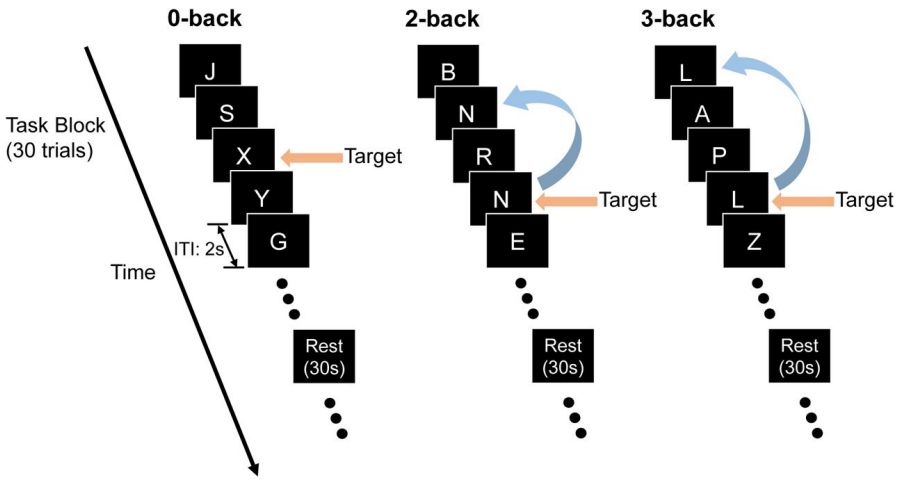


Figure 2.24: Example of 0-back, 2-back and 3-back task implemented in a test.

The flexibility, reliability and ability to systematically manipulate working memory load make n-back particularly suitable for integration with other cognitive measures such as EEG [47].

2.5.3 Multi-Attribute Task Battery-II (MATB-II)

The Multi-Attribute Task Battery II (MATB-II) is a computer-based task designed to evaluate operator performance and workload. MATB provides a benchmark set of tasks and analogous to activities that aircraft crew-members perform in flight, with freedom to use by non-pilot subjects [48].

The MATB was originally developed by Comstock and Arnegard. In 2011, an updated version called MATB-II was released, which integrated the original tasks into a single-computer setup. removing the need for an additional machine to support auditory monitoring [49].

MATB-II mainly consist on four types of tasks, as can be seen in Figure 2.25, [50]:

- **SYSTEM MONITORING (SYSMON)**: this task is divided into two subtasks. In the first, participants are presented with four sliders bars (F1-F4). Their task is to reset each slider whenever its indicator reaches either end of the bar, accomplished by clicking anywhere along the slider. The second subtask involves two lights: F5, which alternates between grey and green to indicate off/on states, and F6, which alternates between grey and red for on/off states. In this case, participants are required to click on the respective box when F5 becomes gray (default state: on) or when F6 becomes red (default state: off).

- **TRACKING (TRCK)**: the participant is required to keep the crosshair inside the blue square at the center by using a joystick. The tracking task has two modes: *automatic*, in which the crosshair remains inside the square without participant input, and and *manual*, in which the participant must actively adjust it.

- **COMMUNICATION (COMM)**: in this task two categories of auditory cues are present: relevant and irrelevant. Only the relevant cues, identified by the call sign "NASA504", require a response, while all other call signs must be disregarded. Each relevant message specifies a communication channel (NAV1, NAV2, COM1 or COM2) along with a six-digit frequency. The task of the participant is to select the appropriate channel associated with the message and subsequently input the corresponding frequency into the designed field.

- **RESOURCE MANAGEMENT (RESMAN)**: the goal of this subtask is to maintain tanks A and B at their designed levels, indicated by blue markers on the sides of the tanks. Since these tanks continuously deplete during the experiment, they must be refilled using tanks C, D, E and F. Fuel circulation between tanks is managed through pumps labeled 1-8, with arrows ('>') indicating the direction of flow. Tanks C and D have limited storage capacity, whereas tanks E and F provide unlimited supply. Each pump can be toggled between two operational state: *inactive* (idle, shown in white) and *active* (pumping, shown in green) by clicking on it. Pumps may also fail during the task, appearing in red to indicate malfunction. After a certain period, broken pumps are automatically restored to working condition.

In addiction, a fifth panel is present on the interface, under the label 'SCHEDULING'. In this section participants can see incoming COMM and TRCK events as a function of vertical timeline, with nearer events signaled at the top of the timeline. Figure 2.25 shows the interface of MATB-II during its execution.

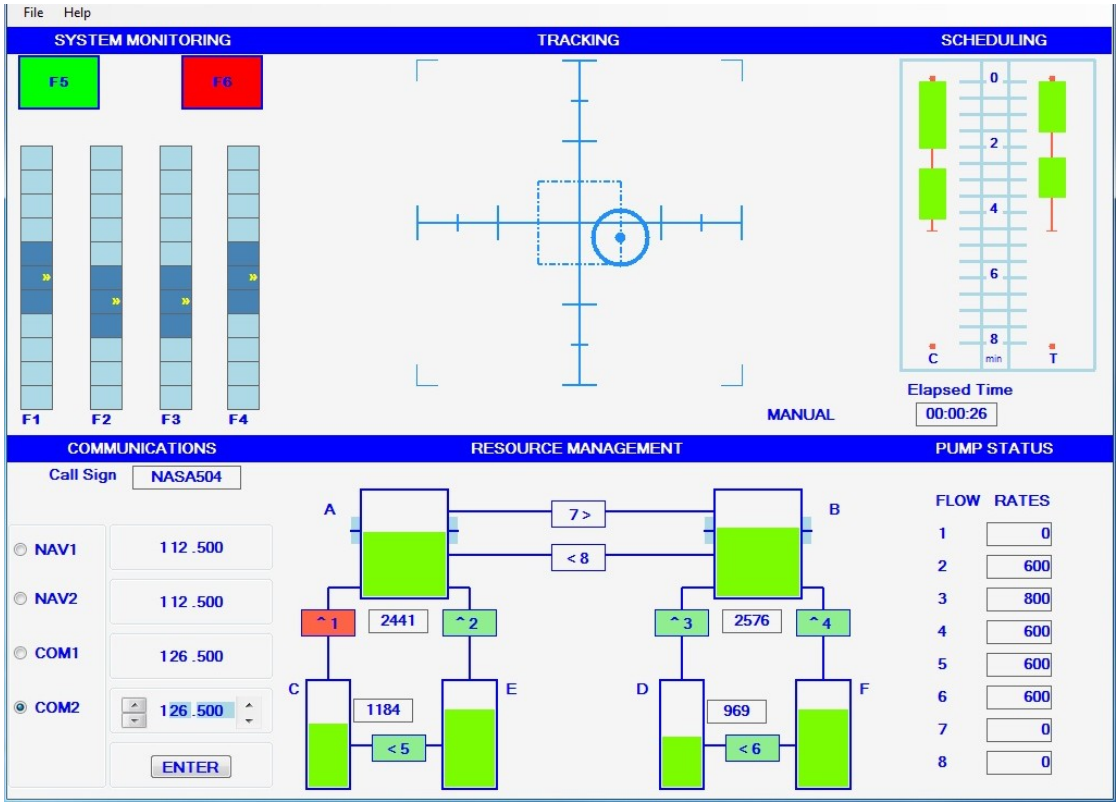


Figure 2.25: Interface of the MATB-II. Top left: SYSMON task. Top center: TRCK task. Top right: SCHEDULING panel. Bottom left: COMM task. Bottom center: RESMAN task.

The complexity of MATB-II can be adjusted by changing task event rate or task mode duration (e.g, *manual* mode in TRCK). For instance, in the SYSMON task, the event rate of green and red light can be changed, as the pump failure frequency in the RESMAN task. For the TRCK task, the duration and number of transitions between *automatic* and *manual*. Lastly, for the COMM task, the proportion of relevant or irrelevant messages can be adjusted.

All of these variables enable the induction of different levels of mental workload in participants, making it possible to use this task in combination with EEG recordings to investigate associated cognitive changes.

Chapter 3

Methods and Materials

In this chapter, the configuration of the MATB-II task and the experimental protocol defined for EEG data collection using the g.HIAMP system are described. Subsequently, the signal processing methods and the extracted features are presented. Finally, a comparative analysis of different feature selection methods and machine learning algorithms is discussed.

EEG signals were recorded using the g.Recorder software. The data processing and all analysis were carried out in Python language using Visual Studio Code. Graphics and signal visualization were produced in MATLAB.

3.1 MATB-II levels

To induce different mental workload states using the MATB-II test, the number of events for each subtask can be modulated, as described in the literature [50, 51].

A total of nine difficulty levels were defined, and grouped into three categories: low (levels 1-3), medium (levels 4-6) and high (levels 7-9) difficulty.

For the SYSMON task, F5-F6 lights systems and F1-F4 sliding bars system event rate increased linearly across levels, with an increase of one event per minute across all levels.

For the TRCK task, the difficulty was changes by extending the duration of *manual* mode. In addition, two other parameters were manipulated: the update frequency of the crosshair position and the joystick response. A higher frequency results in a more sensitive tracker, increasing the attention of the user to keep the crosshair in position. A greater attention was also imposed by a lower joystick response, causing the cursor to react more slowly to the participant's input.

In the COMM task, both the total number of calls and the number of relevant calls increase with each level. All relevant calls are identified by the call sign 'NASA504', ensuring consistency in the recognition of the correct response.

Finally, in the RESMAN task, the adjustable parameters are the number of pump failures and the failure duration. The number of failures increases at a rate of 0.5 failures per minute (failures weren't necessarily contemporary). The failure duration was kept constant for all the levels, at 20 seconds for the first three difficulty and at 30 seconds for the remaining levels.

In addition, a pump failure probability distribution was defined. Pumps 2, 4, 5, and 6 are considered the most critical since they are connected to the infinite fuel tanks; therefore, a higher failure probability was assigned to them. Conversely, pumps 7 and 8 are the least critical and were assigned a lower failure probability. This probability distribution remained the same across all difficulty levels and is reported in Table 3.1.

Pump	P1	P2	P3	P4	P5	P6	P7	P8
Failure probability	0.11	0.17	0.11	0.17	0.17	0.17	0.05	0.05

Table 3.1: Failure probability distribution of RESMAN pumps.

In addition, for the SYSMON, TRCK, and COMM tasks, a timeout parameter was introduced to prevent two consecutive events of the same task from occurring too close to one another.

For the SYSMON task, a timeout of 5 seconds was assigned to the F5 and F6 light systems, while a timeout of 6 seconds was applied to the F1–F4 sliding bars.

For the TRCK task, the timeout was determined according to the following equation:

$$T_{\text{TRCK}} = D_{\text{tracking}} + 5$$

where D_{tracking} is the duration of the *manual* mode in the defined difficulty level, and 5 seconds is a security margin.

Finally, for the COMM task, a timeout of 20 seconds was defined, as this task is the most time-consuming among all MATB-II .

These timeout values were also used to define the latest possible time instant at which the last event could start before the end of the test. For the RESMAN task, the timeout was set equal to the pump failure duration, and it was used solely to determine the maximum allowable start time for the last event.

Table 3.2 summarize the events defined for each of the nine levels on a 2 minute test duration.

Level	SYSMON			TRCK				COMM		RESMAN	
	F5	F6	F1-F4	Man	Upd	Resp	Dur (s)	Act	Rel	Fail	Dur (s)
1	2	2	2	0	M	H	30	2	0	0	20
2	4	4	4	1	M	H	20	2	1	1	20
3	6	6	6	1	M	H	30	3	1	2	20
4	8	8	8	2	M	M	20	3	2	3	30
5	10	10	10	2	H	M	30	4	2	4	30
6	12	12	12	1	H	M	70	4	3	5	30
7	14	14	14	1	H	L	90	5	4	6	30
8	16	16	16	1	H	L	105	6	5	7	30
9	18	18	18	1	H	L	119	6	6	8	30

Table 3.2: Event configuration for each difficulty level (Levels 1-9) for a 2-minute test. L,M,H stand for 'low', 'medium' and 'high' respectively; 'Man', 'Upd', 'Resp', 'Dur (s)' stand for 'Manual', 'Update', 'Response' and 'Duration (seconds); 'Act', 'Rel' stand for 'Activations' and 'Relevant'; 'Fail', 'Dur (s)' stand for 'Failures' and 'Duration (seconds)'.

3.2 Task configuration

For the experiment, different MATB-II difficulty levels were arranged in sequence to create a single test session. First, the overall structure of the test was defined: each sequence of a given difficulty lasted 2 minutes, with a total of four sequences per test. A 20-second pause was introduced between sequences, allowing participants to perform a self-assessment of MWL. In this way, a continuous test lasting 11 minutes and 40 seconds was generated. Participants were unaware of the difficulty

levels at any point during the experiment, to prevent subjective influences on task perception.

Four different level sequences were designed to induce varying mental workload states within a single session:

- **Sequence 1** (Figure 3.1a) : The test begins with a medium difficulty level, providing participants with an initial reference for task complexity. Difficulty then increases linearly, from level 2 up to level 8.

- **Sequence 2** (Figure 3.1b) : The test follows an inverted U-shape, in which difficulty raises to a peak before gradually decreasing.

- **Sequence 3** (Figure 3.1c) : The opposite pattern of Sequence 2 is applied: difficulty decreases from level 8 to level 2, and then increases again up to level 9.

- **Sequence 4** (Figure 3.1d) : The test follows a sinusoidal pattern, progressively increasing and decreasing until returning to a medium level.

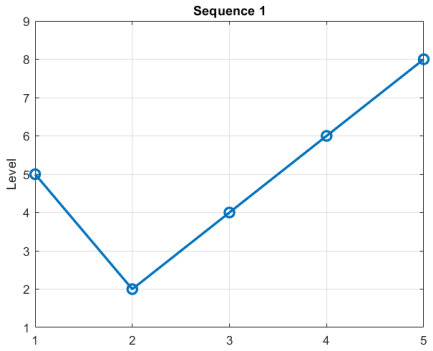
It is important to note that sudden changes between high and low difficulty were avoided to maintain coherence in MWL progression, reflecting the gradual changes in cognitive load that occur in real-life situations.

3.3 Secondary task

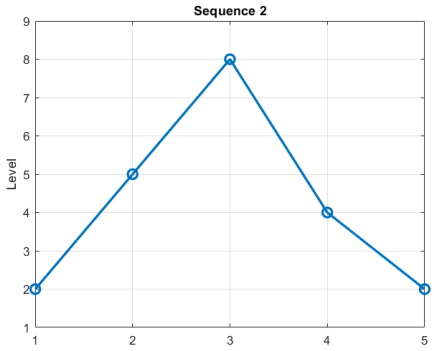
In addition to the MATB-II, a secondary task was included in the experiment to further increase the multitasking demands of the test. The secondary task was performed simultaneously with the MATB-II, including during the breaks between sequences. It consisted of simple arithmetic operations, involving addition and multiplication of single-digit numbers, with four possible answers, only one of which was correct. The interface presented to participants is shown in Figure 3.2.

Each operation lasted 10 seconds. If the participant responded before the end of the allotted time, the next question was presented immediately. Conversely, if no answer was provided within the first two-thirds of the time window, an auditory cue was played to regain the participant's attention.

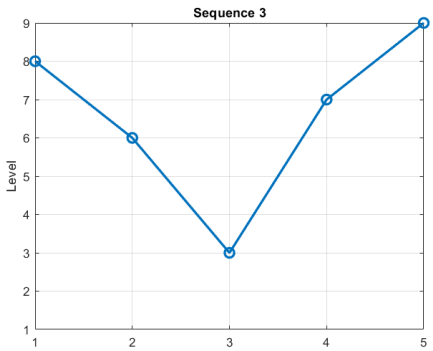
The questions were generated randomly, and the difficulty of the secondary task was kept constant, since its primary purpose was to divert the participant's attention from the main task, the MATB-II.



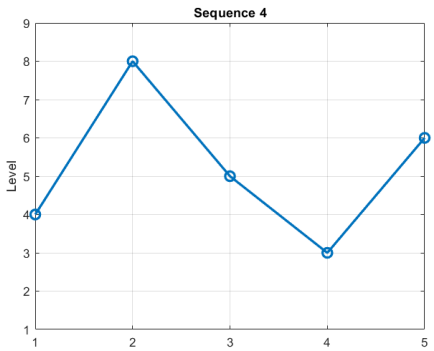
(a) Sequence 1: 5, 2, 4, 6, 8.



(b) Sequence 2: 2, 5, 8, 4, 2.



(c) Sequence 3: 8, 6, 3, 7, 9.



(d) Sequence 4: 4, 8, 5, 3, 6.

Figure 3.1: Sequences of MATB-II generated.

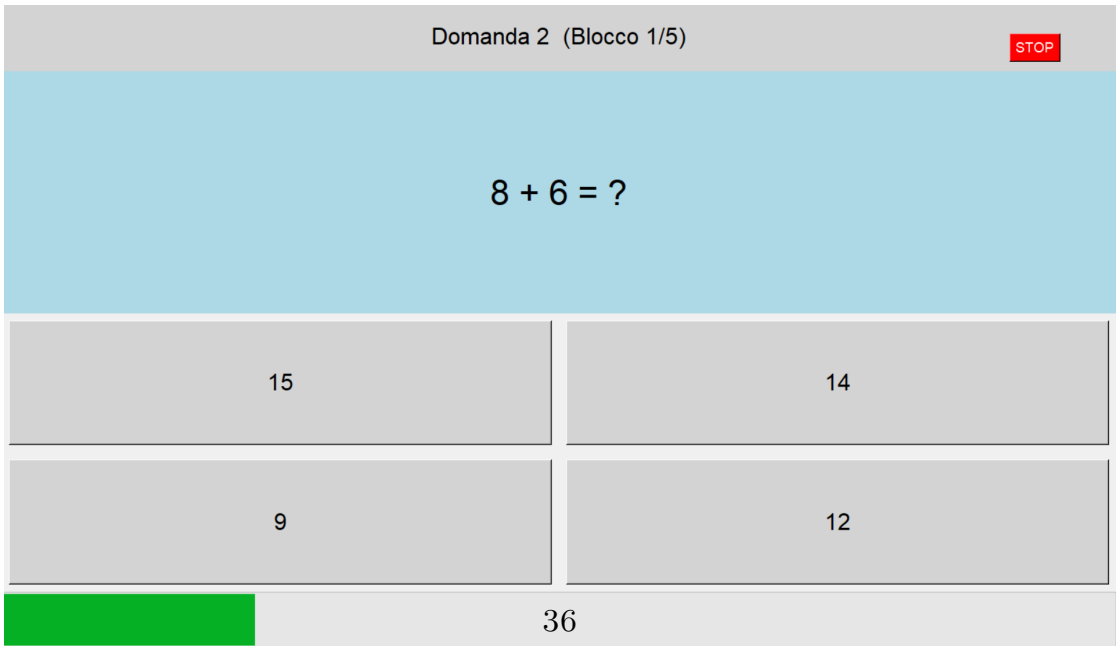


Figure 3.2: Interface of the secondary task.

3.4 Setup

For this study, the g.HIamp Channel Amplifier (Figure 3.3) by g.tec was used to record EEG data from the frontal lobe of the participants. The acquisition system includes a power supply cable, a central hub with four input ports for connection cables leading to a 64-channel passive/active connector box (Figure 3.4), and a USB cable that connects the hub to an external computer for data recording. Additional input ports are available for other compatible devices.

Figure 3.3: g.HIAMP recorder central hub [52].

Figure 3.4: Frontal view of the gHIAMP connector box. The yellow input is dedicated to the ground connection.

The central hub was connected to a computer via the USB interface, enabling real-time signal visualization and recording through the gRecorder software. This

software allowed the creation of a standardized recording model used across all sessions, where parameters such as sampling frequency (set at 1200 Hz) and the mapping between connector box inputs and electrode positions were defined and replicated for every experiment.

The electrodes used for the recordings were Kendall Arbo ECG H124SG (Figure 3.5), selected for their versatility in physiological measurements and their minimally invasive design. These electrodes incorporate a highly conductive solid gel and strong adhesive properties, ensuring stable signal acquisition throughout the experiment. The flexible foam backing provides comfort and facilitates proper placement, while the silver/silver chloride (Ag/AgCl) sensor and liquid-sealed backing support high-quality and reliable signal conduction [53].

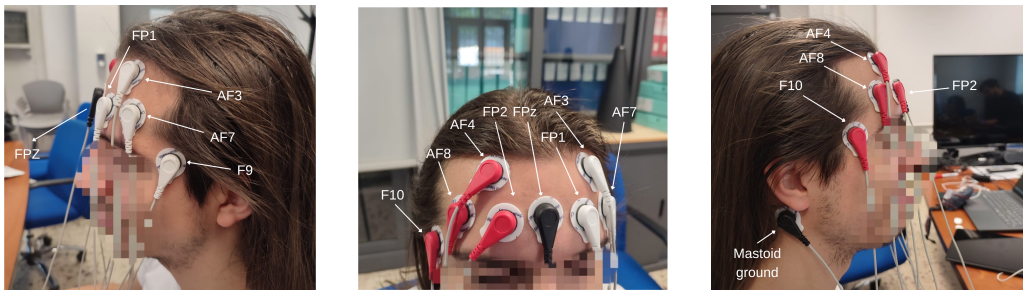
Figure 3.5: Frontal view of the Kendall ECG electrode. [53].

The electrode placement used followed the instructions of the International 10-20 system, and included only frontal and prefrontal regions of the scalp. This configuration was employed to reduce the system invasiveness and avoid interference with subjects' hair. In addition, the frontal lobe is the main source of high-level information elaboration, decision-making and memory functions, making it appropriate for the problem considered.

A total of eight channels (F10, AF8, AF4, FP2, FP1, AF3, AF7, and F9 locations) were employed in a monopolar configuration, with the reference placed at 'Fpz', centrally between hemispheres.

An additional electrode, used as the ground reference, was positioned on the right mastoid, just below the ear. This site was selected because it is considered electrically neutral, which helps ensure cleaner and more stable signal recordings.

The full positioning system is represented in Figure 3.6.



(a) Left-side view of the electrode setup.

(b) Frontal view of the electrode setup.

(c) Right-side view of the electrode setup.

Figure 3.6: Complete view of the electrode setup.

The MATB-II software was executed on a computer connected via USB and HDMI cables to a touchscreen monitor. This configuration provided participants with greater flexibility in their task execution strategies. To perform the TRCK task, an Airbus joystick (Figure 3.7) was connected to the computer through a USB interface.

The secondary task was implemented on a tablet, positioned on the opposite side of the joystick to ensure ergonomic balance and ease of interaction during the experiment. Figure 3.8 shows the complete experimental setup.

A separated computer was connected to the g.HIAMP to collect the data in real time.

Figure 3.7: AirBus joystick used for the TRCK task.

Finally, the Biosignalsplux kit Figure 3.9) was also employed in this study. The kit includes a central hub, a trigger device, a USB cable for connection to an external computer, and several physiological sensors such as ECG, EDA, and fNIRS. However, the only component used in this experiment was the external trigger, a handheld switch that changes its state when pressed. The trigger was used to mark the start and end of each experimental procedure, ensuring accurate synchronization between the physiological recordings and the task events.

Figure 3.8: Experimental setup of the test.

(a) Biosignalsplux central hub. **(b)** Handheld switch.

Figure 3.9: Biosignalsplux elements used in the study [54].

3.5 Experimental Protocol

In the experimental protocol, all four sequences defined in chapter 3.2 were included. In particular, each sequence was presented in the same order across all participants. This choice was done because the level of difficulty already changed within the same sequences; in this way all subjects performed the test under the same experimental conditions.

A total of 18 volunteers took part in the study (mean age of 24.39 $\pm$ 2.62 years), with a gender distribution of 14 males (78%) and 4 females (22%). Due to corrupted data, two participants were excluded from the analysis, while one subject recording was incomplete because of technical issues, resulting in the loss of approximately three quarters of the EEG data.

Before the experimental session, each participant completed a training phase designed to familiarize them with the MATB-II environment and its tasks. The training was conducted at least one day prior to the main experiment to ensure adequate preparation. It lasted approximately 30 minutes and consisted of three 10-minute MATB-II runs, all set to a medium difficulty level. At the beginning of the session, participants were seated in front of the touchscreen monitor and asked to indicate their dominant hand, which was then used to operate the joystick. Each MATB-II subtask was explained in detail beforehand. Importantly, participants were not informed about the presence of a secondary task, so as to increase the workload and induce a mild stress response during the actual experiment.

At the end of the training, the Bedford Workload Scale (Chapter 2.4.1) was introduced and explained, as it was the instrument used for subjective workload self-assessment throughout the experiment.

On the test day, participants were instructed to avoid intense physical activity, alcohol consumption, and caffeine intake in the hours preceding the session, in order to minimize potential interference with EEG data quality.

Before the experimental session began, the procedure was clearly explained, including the execution of the secondary task. Subsequently, participants were required to sign an informed consent form, which formally authorized their involvement in the study and permitted the utilization of limited personal data, specifically their age and sex. To uphold privacy and data security standards, all other potentially identifiable personal data were anonymized.

The experimental setup was then arranged: the joystick was positioned near the dominant hand, and the tablet used for the secondary task was placed next to the monitor on the opposite side. Participants sat in front of the touchscreen monitor, and the EEG cap was prepared using the g.GAMMAcap from g.tec (Figure 3.10). Electrode positions were first marked on the scalp using a pen, after which the cap was removed to apply the electrodes and then repositioned to ensure proper alignment. Two headphones were placed in the participant's ears, one for the

COMM communications within the MATB-II, and the other for the audio cues of the secondary task.

Once the g.HIAMP was connected and the g.Recorder software was ready, a 5-minute resting state was recorded. Participants were instructed to remain relaxed and silent during this period.

Following the rest phase, the first sequence (Figure 3.1a) was presented. After the initial two minutes of MATB-II activity, a black screen appeared on the monitor, accompanied by an audible countdown emitted on the secondary task tablet. Simultaneously, the secondary task paused and the Bedford Workload Scale appeared on the tablet, allowing participants 20 seconds to select the rating that best reflected their perceived workload. The MATB-II task then resumed automatically. This sequence of task–self-evaluation cycles was repeated until the end of the first sequence.

At the conclusion of each sequence, a 3-minute resting period was introduced to allow physiological parameters to return to baseline levels. The entire block, including task execution, self-assessment, and resting phase, was repeated for the remaining three sequences, presented in the same order across all participants (sequence 2, sequence 3, and sequence 4; see Figure 3.1). At the beginning and end of each phase, including the resting phase as well as the start and end of each task sequence, the trigger was activated to mark those events, resulting in a total of ten triggers per experiment.

The complete experimental procedure is illustrated in Figure 3.11, and lasted approximately one hour and 15 minutes, including 46 minutes and 40 seconds of test performance, 9 minute of rest between tasks, 5 minute of initial rest and the time dedicated to apply and remove the electrodes from the subjects.

Figure 3.10: g.GAMMAcap used to position the EEG electrodes [55].

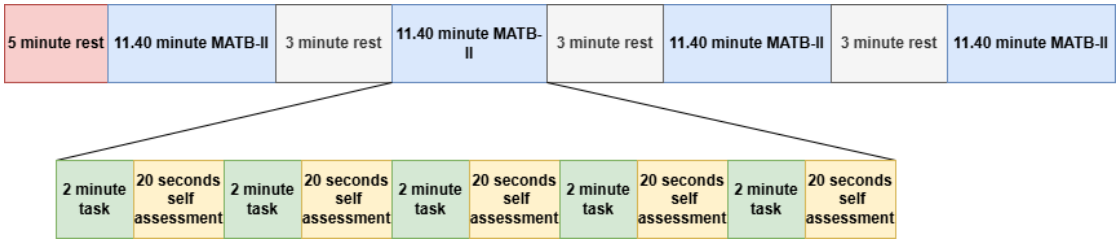


Figure 3.11: Complete experimental procedure.

The Bedford rating scale, presented to participants through the interface shown in Figure 3.12, was selected because, unlike the more commonly used NASA-TLX, it is more suitable for near real-time assessments. This feature allows for a higher temporal resolution in the acquisition of subjective workload data.

Each level of the Bedford scale is accompanied by a short description, which helps participants better distinguish between different cognitive states. The scale is structured according to a sequence of three discriminative questions:

- Was it possible to complete the task?
- Was the workload tolerable for the task?
- Was the workload satisfactory without reduction?

Based on the participants' answers, four levels of workload can be identified. The first category, corresponding to the inability to complete the task, was excluded, as the experiment was designed to be always feasible. The remaining categories allowed for the classification of subjective mental workload (MWL) into low, medium, and high levels.

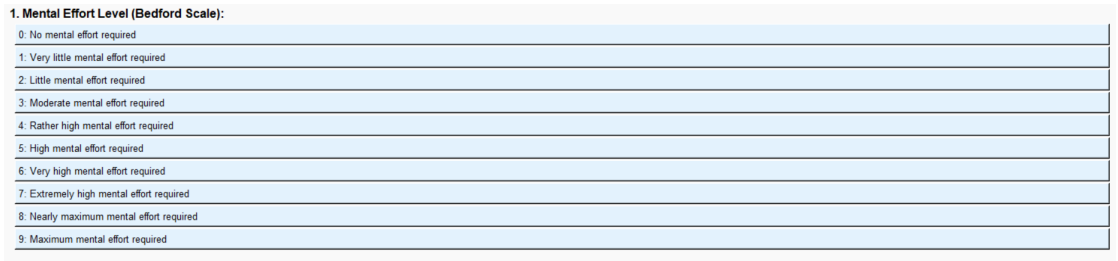


Figure 3.12: Interface of the Bedford Workload Scale.

Chapter 4

Processing

4.1 Data extraction

The EEG signals recorded with the gRecorder software were stored in files with the ".hdf5" format. Each file contained the complete dataset from a single participant's experimental session. The file name included information about the date and time of recording (hour and minute), allowing temporal identification of each dataset.

Alongside the EEG data, the Biosignalplux system generated a ".txt" format file containing the time instants corresponding to the external trigger activations. The file included the recording date, hour, minute and second in its name, allowing temporal alignment with the EEG data.

All subject-specific folders were first collected into a single main directory. Visual Studio Code, combined with the Python programming language, was used for all data processing operations, while MATLAB was employed primarily for signal visualization.

A dedicated Python script was developed to extract EEG signals from both the resting phase and the active workload phases, and to assign them to the corresponding subject.

Inside an iterative loop, the script analyzed the subfolder of each subject, performing the following steps:

- **EEG data extraction**: The script first searched for a ".hdf5" file. If found, it extracted the recording start time (hour, minute, and second) and passed the file path, together with the predefined channel labels, to a dedicated function. This function retrieved the sampling frequency and generated a DataFrame in which each column corresponded to one EEG channel.

- **Trigger data extraction**: The script then searched for the ".txt" file containing the trigger information. A dedicated function identified the column

corresponding to the trigger signal (labeled "DI"), which contained binary values (0 and 1). The function detected all 0-to-1 transitions, corresponding to trigger activations, sorted them in chronological order, and extracted the sampling frequency of the Biosignalplux device.

- **Subjective workload data extraction**: Next, the script searched for the ".csv" files containing the Bedford Workload Scale evaluations. Each participant had four files, corresponding to the four task sequences. The sequence number was embedded in the file name, allowing the script to correctly associate each subjective rating with the appropriate sequence window. Since the order of sequences was the same for all participants, the corresponding difficulty levels were automatically assigned. The function returned a DataFrame containing the sequence number, the Bedford Workload Scale rating, and the associated difficulty level.

- **Signal segmentation and data alignment**: A final function was responsible for extracting the EEG signal segments corresponding to the resting and active task phases, and for associating the subjective ratings with the relevant segments. This was achieved by aligning the start time of the ".hdf5" EEG recording with the start time of the trigger ".txt" file, thereby synchronizing the trigger samples with the correct EEG samples. The resting phase was extracted from the interval between the first and second triggers. Subsequently, the start and end samples of each task sequence were determined. Knowing the predefined duration of each phase (2 minutes of task execution and 20 seconds of evaluation), the corresponding EEG segments were separated accordingly.

  All processed information, including EEG segments, trigger events, subjective ratings, and sequence difficulty, was stored in a dictionary, structured as illustrated in Table 4.1, with all elements organized chronologically.

Field	**Description**
test_id	Identifies the test sequence.
window_id	Identifies the active window within the sequence.
segment	Sequential number that identifies the EEG matrix.
difficulty_level	MATB-II difficulty level.
bedford_rating	Subjective workload score.

Table 4.1: Structure of all informations contained in the dictionary.

A higher-level dictionary was then defined within the main loop, in which the previously described dictionary was associated with the corresponding subject ID.

The resulting data structure was saved in the '.h5' format, allowing the preservation of the nested dictionary hierarchy and facilitating efficient data access in subsequent analyses.

4.2 Signal pre-processing

Once the signals of interest were extracted, a cleaning phase was performed. The signal pre-processing pipeline, illustrated in Figure 4.1, outlines the sequence of operations applied to prepare the data for subsequent analysis.

The main script was responsible for extracting a structured dictionary containing all subjects' recordings. Each dataset was then processed within a loop that instantiated the EEG class, where both the pre-processing routines and the feature extraction procedures were implemented.

At last, the processed signals were stored in a MATLAB file, enabling further visual inspection of the pre-processing outcomes.

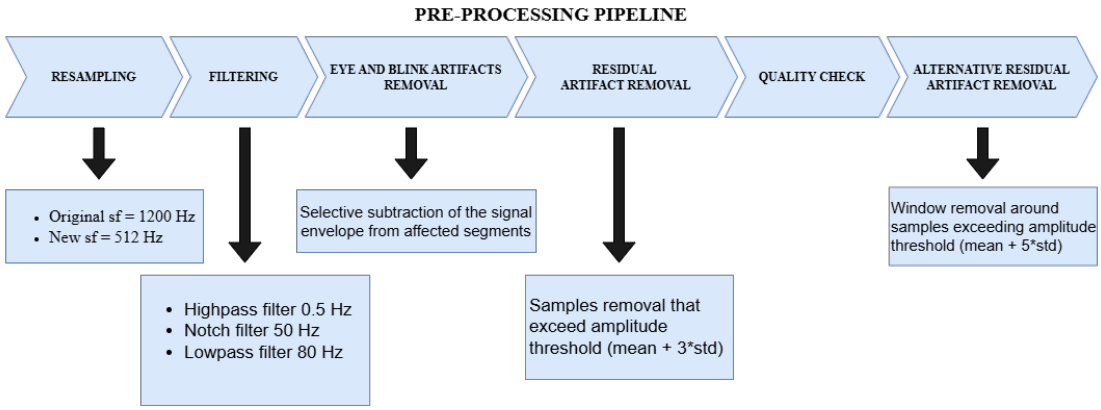


Figure 4.1: Signal pre-processing pipeline.

4.2.1 Resampling

The first step of the pipeline involved reducing the sampling frequency, as the initial rate of 1200 Hz exceeded the requirements for EEG analysis. Moreover, such a high frequency would have considerably increased the computational cost. Therefore, the sampling frequency was reduced to 512 Hz.

The EEG signals were resampled using a polyphase finite impulse response (FIR) resampling method, implemented through the 'resample_poly' function from the SciPy library. Specifically, the algorithm computes an exact rational ratio between the target and original sampling frequencies, allowing an efficient and

accurate conversion from the original rate (1200 Hz) to the new one (512 Hz). Each EEG channel was resampled independently to preserve spatial independence among electrodes. The polyphase approach performs an implicit upsampling, low-pass filtering, and subsequent downsampling, ensuring that frequency components above the new Nyquist limit are effectively attenuated.

This function is particularly suitable for EEG data, as it provides a linear-phase and and anti-aliasing filter that preserves the morphology of the signal and avoids spectral distortion. Its computational efficiency allows multi-channel and long-duration recordings.

4.2.2 Filtering

The filtering of the EEG signals was performed to reject frequency components not related to cerebral activity, aiming at an overall improvement of the signal-to-noise ratio. Pre-processing involved a chain of three zero-phase digital filters in sequence: high-pass, notch, and low-pass. Filtering was performed using the *filtfilt* function, which applies forward and reverse filtering to achieve zero-phase distortion and thus preserve the temporal morphology of EEG waveforms.

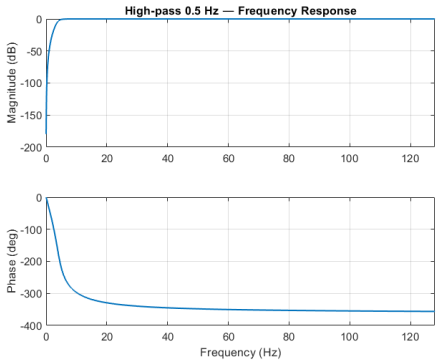
First, the high-pass filter with a cut-off frequency of 0.5 Hz (Figure 4.2a) was used to suppress low-frequency components such as slow drifts, baseline wander, and electrode polarization artifacts, since these components typically arise from sweat, subtle head movements, or instrumental instability and obscure relevant neural oscillations. The filter was designed as a fourth-order Butterworth Infinite Impulse Response (IIR) filter, providing steep roll-off but still maintaining stability in the range of interest for EEG frequencies.

After, an order-six Butterworth low-pass filter at 80 Hz cutoff frequency removes high-frequency noise including muscular activity (EMG) and electronics interference (Figure 4.2b). With this last step, only the spectral content of the main rhythms has been retained for subsequent analyses.

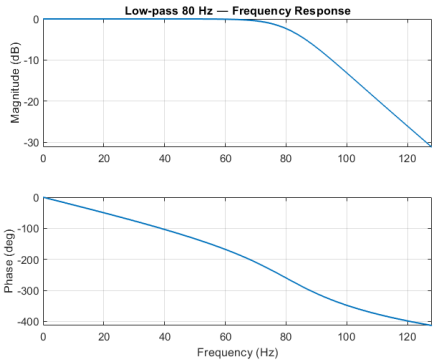
Finally, a notch filter centered at 50 Hz (Figure 4.3) was applied to suppress power line interference, among the most frequent sources of contamination in EEG recordings. The narrow stopband of this biquadratic IIR filter effectively attenuates the fundamental mains component without significantly affecting adjacent neural activity, especially the beta band representing activity within a frequency range between 12 to 30 Hz, which lies close to the notch region.

The combination of these filters allowed the extraction of a clean and physiologically relevant EEG signal, suitable for both time and frequency domain feature computation.

Figures 4.4 and 4.5 illustrates an example of the effect of the filtering process on an EEG signal.



(a) High-pass filter (0.5 Hz) mask.



(b) Low-pass filter (80 Hz) mask.

Figure 4.2: High-pass and low-pass filters masks.

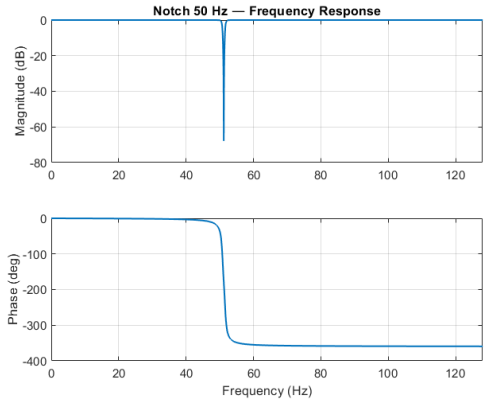


Figure 4.3: Notch filter (50 Hz) mask.

4.2.3 Artifact removal

The removal of transient artifacts such as eye blinks and head movements was performed through an adaptive correction procedure based on amplitude thresholding and envelope subtraction. For each EEG channel, the algorithm first identified potential artifacts by computing a dynamic threshold defined as the mean signal amplitude plus three times its standard deviation. Samples exceeding this threshold were marked as contaminated. To ensure full removal of the artifact influence, the detected regions were expanded by a temporal margin of approximately one second on both sides, accounting for the gradual onset and recovery typically observed in ocular and motion-related disturbances.

An envelope of the original signal was then estimated using a short-term moving average filter (window length of 0.1 s), which provided a smooth representation

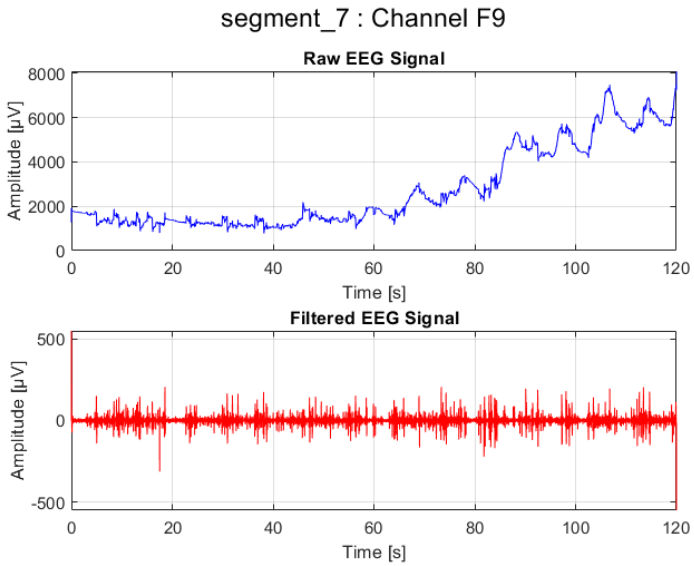
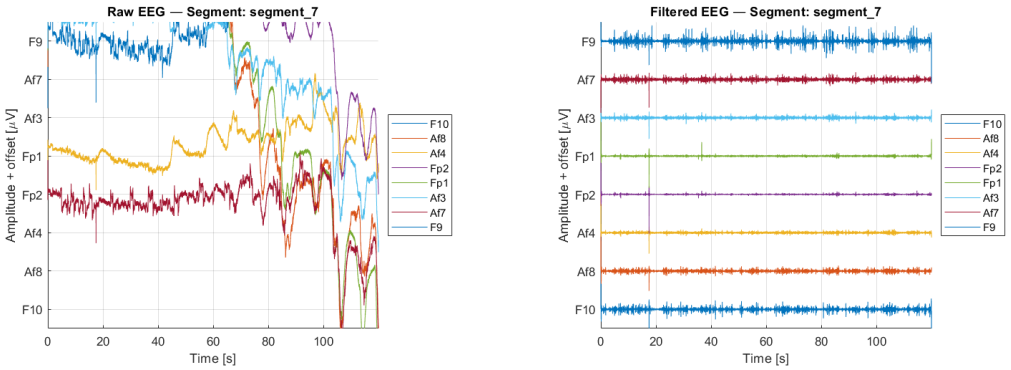


Figure 4.4: Effect of filtering on a single channel during a task phase. On the top: raw signal. On the bottom: filtered signal.



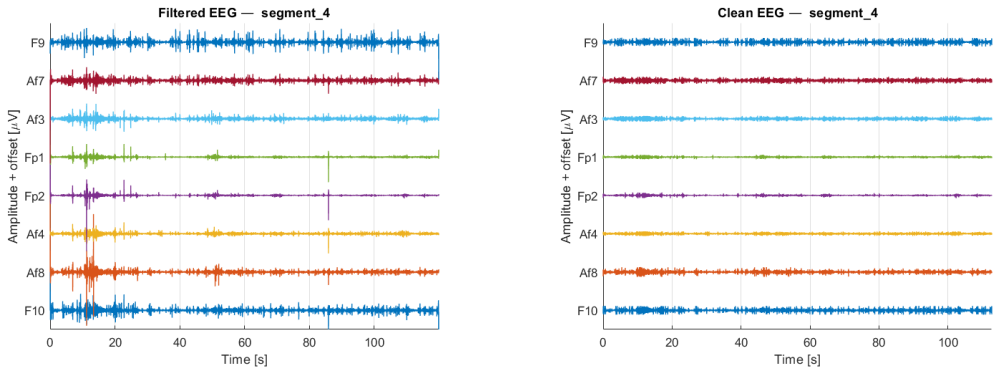
(a) Raw EEG channels. **(b)** Filtered EEG channels.

Figure 4.5: Comparison between raw and filtered EEG channel matrices.

of low-frequency drifts associated with slow amplitude fluctuations. Within the artifact regions, this envelope was subtracted from the raw signal, effectively attenuating large deflections while preserving the underlying neural oscillations. The resulting output exhibits reduced contamination from high-amplitude, non-neural transients, improving the quality and interpretability of the EEG signal for subsequent analysis.

An additional step was applied to remove residual artifacts that simultaneously affected multiple EEG channels. This procedure aimed to eliminate short, high-amplitude noise that may persist, such as abrupt movements or muscular contractions producing broadband deflections across electrodes.

For each channel, residual samples deviating more than a specified threshold (set as three standard deviations from the mean) were identified as potential outliers. A global mask was then computed by combining the individual channel masks through a logical intersection, ensuring that only samples consistently clean across all channels were retained. In this way, any time point exhibiting an abnormal amplitude in one or more electrodes was discarded from the dataset. The resulting cleaned signal therefore preserves only temporally and spatially consistent neural activity, improving the reliability of subsequent analyses such as spectral estimation or functional connectivity assessment. Results can be observed in Figure 4.6



(a) EEG channels after filtering.

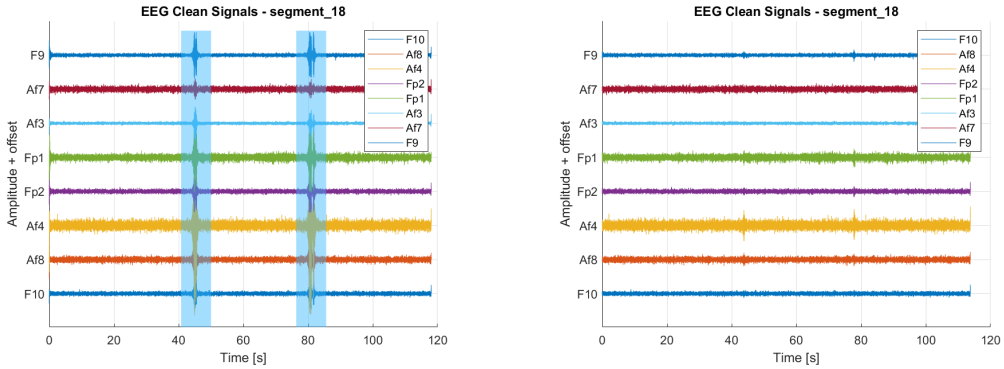
(b) EEG channels after artifact removal.

Figure 4.6: Comparison between EEG channel matrices before and after artifact removal.

After a first qualitative processing inspection, signals that still exhibited high noise underwent an alternative artifact removal step, replacing the standard one defined above. This approach was designed to remove brief and high-amplitude noise that occurred sporadically throughout recordings. This passage was applied to a small number of subjects, most likely due to experimental factors such as muscular artifacts or changes in skin-electrode impedance.

For each EEG signal, a threshold of mean amplitude plus five times the standard deviation was computed. Samples of any channel that exceeded this threshold were marked, and a window of ± 500 ms around the sample was defined, and a mask corresponding to that window was saved. By merging masks across all channels, the algorithm discarded windows that were contaminated in all channels, ensuring temporal consistency across electrodes.

This approach proved particularly effective for mitigating abrupt, broadband distortions caused by muscular or mechanical artifacts, while preserving the physiological EEG components. The resulting signal thus maintained the integrity of genuine neural oscillations and ensured improved data quality for subsequent quantitative analyses. Figure 4.7 illustrates the effect of the cited approach.



(a) Noisy EEG segment after the original processing step.

(b) Noisy EEG segment after the alternative processing step.

Figure 4.7: Comparison between EEG channel matrices with the original and alternative artifact removal. The blue window indicates the removed samples in the alternative processing step.

Finally, any signal segments that remained corrupted after pre-processing were discarded. In one participant, the channel 'AF7' appeared to have been improperly positioned, resulting in persistent artifacts throughout the entire recording. For this reason, the participant was excluded from the subsequent analyses to ensure consistency and reliability of the dataset.

4.2.4 Feature Extraction

The final part of the signal processing was the extraction of relevant features for further analysis among differences of these in different groups of mental workload levels.

The *EEG* class contained, together with the signal pre-processing, the feature extraction pipeline. The preprocessed EEG data were first segmented into overlapping temporal windows of 2, 3, and 4 seconds with a 50% overlap, ensuring an adequate trade-off between temporal resolution and statistical stability. Different temporal supports were tested in this work, as the right size for signal processing is an open debate in literature.

Each channel was segmented into smaller epochs through this procedure, allowing feature computation within locally stationary signal portions. For each window and each channel, a set of features was computed to capture distinct aspects of the signal dynamics:

- **Temporal features** (Table 4.2)
- **Complexity features** (Table 4.3)
- **Spectral features** (Table 4.4)
- **Intra-channel connectivity features** (Table 4.5)

Although Hjorth parameters are derived from linear computations, they were included in the set of non-linear and entropy-based features. This choice is motivated by their ability to describe the temporal dynamics and structural complexity of EEG waveforms rather than simple amplitude variations. In particular, Hjorth Mobility and Complexity provide indirect information about the dominant frequency content and the degree of waveform irregularity, making them conceptually aligned with other non-linear descriptors such as the Hurst exponent and fractal dimension.

For frequency domain analysis, the power spectral density (PSD) of each segment was estimated using the Welch method combined with a Hamming window. The spectrum was then normalized for all EEG bands.

Feature	**Description**
Mean amplitude	Average voltage of the EEG segment, indicating the general signal offset and baseline level.
Variance	Statistical dispersion of the signal around its mean, representing the total power and variability.
Root Mean Square	Effective amplitude of the signal, combining both magnitude and duration of fluctuations.
Normalized absolute energy	Mean of the absolute signal amplitude over time, providing a normalized measure of global activity.

Table 4.2: Amplitude-based temporal feature.

Feature	Description
Permutation entropy	Quantifies the degree of local order and randomness based on the relative arrangement of neighboring samples.
Spectral entropy	Measures the flatness of the power spectrum; higher values indicate more uniform spectral content.
Hjorth Activity	Represents the overall signal variance, proportional to signal power.
Hjorth Mobility	Indicates the mean frequency of the signal, obtained from the standard deviation ratio of the derivative to the signal itself.
Hjorth Complexity	Describes the degree of waveform irregularity and deviation from a pure sine wave.
Hurst exponent	Estimates the long-term temporal correlation or self-similarity of the signal (persistent vs. anti-persistent behavior).
Higuchi fractal dimension	Characterizes the fractal complexity of the time series, capturing the irregularity and roughness of the waveform.

Table 4.3: Complexity features describing signal irregularity and complexity.

Feature	Description
Relative band power: *delta*	Normalized PSD over the δ band (0.5–4 Hz), reflecting slow-wave activity.
Relative band power: *theta*	Normalized PSD over the θ band (4–8 Hz), often associated with drowsiness and memory processes.
Relative band power: *alpha*	Normalized PSD over the α band (8–12 Hz), linked to relaxed wakefulness and cortical idling.
Relative band power: *beta*	Normalized PSD over the β band (12–30 Hz), related to active processing and sensorimotor rhythms.
Relative band power: *gamma*	Normalized PSD over the γ band (30–80 Hz), indexing higher-frequency activity.
Inter-band power ratios (all pairs)	Ratios between relative band powers, highlighting spectral redistribution sensitive to cognitive load.

Table 4.4: Frequency-domain features.

Feature	Description
Broadband coherence (pairwise)	Mean magnitude-squared coherence averaged over all frequencies for each channel pair (i, j); indexes overall synchrony.
Band-limited coherence: *delta*, *theta*, *alpha*, *beta*, *gamma*	Mean coherence within each canonical EEG band for each pair (i, j); captures frequency-specific functional coupling.

Table 4.5: Connectivity features based on magnitude-squared coherence between channel pairs.

For each extracted feature excluding coherence four statistical indices were computed across all overlapping segments: mean, variance, kurtosis, and skewness. The mean represents the average value of a feature over time, providing an estimate of its central tendency, while the variance quantifies its dispersion and temporal variability. Kurtosis measures the "tailedness" of the distribution, indicating how frequently extreme values occur compared to a normal distribution (higher kurtosis reflects the presence of occasional large deviations). Skewness, on the other hand, quantifies the asymmetry of the feature distribution: positive skewness denotes a longer right tail (dominance of higher feature values), whereas negative skewness indicates a longer left tail. Together, these indices summarize both the average behavior and the statistical distribution shape of each feature over time, allowing a more complete characterization of the EEG dynamics beyond simple point estimates.

Given that the features were computed for each of the eight EEG channels, a total of 128 amplitude-based features, 224 complexity features, and 480 spectral features were extracted for each signal corresponding to an experimental phase (rest and workload).

For the coherence features, only the mean and the variance were extracted. Since the dataset already included 168 unique channel pairs, adding more statistics would have unnecessarily increased redundancy. With these two measures, the coherence feature set reached a total of 336 values.

The complete dataset was saved in a file named *Results.csv*. The dataset contained, along with the features, all the information related to the signal segments (subject ID, segment name, test ID, window ID, Bedford rating and difficulty level).

Chapter 5

Classification

The final stage of the processing pipeline involved the classification of mental workload using machine learning algorithms, aimed at evaluating differences both between the resting and workload conditions, and among different workload levels. Specifically, this section focused on two approaches: a binary classification distinguishing rest from workload, and a multiclass classification comprising four classes (rest, low, medium, and high workload).

In the binary case, class labels were defined based on segment naming. Since the 5-minute resting phase was always performed before the experimental task, the first segment of each subject, named 'segment_0', corresponded to the rest condition. Logically, all the other segments belonged to the workload class. Accordingly, a new column named binary was added to the dataset, where label 0 indicated rest and label 1 denoted workload.

For the multiclass analysis, four class labels were defined: rest (0), low (1), medium (2), and high workload (3). As in the binary setup, the rest condition was assigned to all segments labeled as 'segment_0'. The remaining workload levels were extracted by users self-evaluations using the Bedford scale.

The Bedford scale distinguish naturally between four levels of workload, although the highest one (level 10, test cannot be performed) was not considered in the study. Ratings ≤ 3 were classified as low workload, from 4 to 6 as medium workload, and ratings ≥ 7 as high workload. Based on these groups, a new multiclass label was added to the dataset.

The complete dataset was divided according to the four feature categories and the three window sizes, resulting in a total of twelve smaller datasets. In addition, three further datasets were created by grouping data only by window size, in order to allow comparisons with the complete model. Each dataset contained 318 examples, corresponding to the signal segments from which the features were extracted.

Subsequently, all datasets were normalized using the min-max method, expressed

by the following equation:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

where x is the original feature value, $x_{\min}$ and $x_{\max}$ represent the minimum and maximum values of the feature, respectively, and x' is the normalized feature. Through this transformation, all feature values are rescaled within the range 0 and 1.

To prevent data leakage (that is, the unintended transfer of information from the test set into the training process, which could artificially inflate performance) normalization was performed within each subject individually. In this way, the minimum and maximum values were computed separately for each participant, ensuring that no information from other subjects or from the test data influenced the training phase. A dedicated script, called *Normalization*, implemented the logic described above.

After the data preparation, a dedicated script, called *Generic_subdivision_train_test_sets*, performed the division of each dataset into train and test set. The script automated a subject-wise, stratified train/test split for all datasets in a folder, ensuring that the same subjects consistently fall into train or test across files. After collecting candidate datasets, it checked for the presence of the subject ID column and the chosen label column, then on the first file invokes the updated utility function *df_split_stratified_grouped_categorical_output* to search over many random seeds and select the split that best matches the class distribution between train and test (minimizing an RMSE between expected and observed test counts, with all rows from each subject kept entirely in one set). It saved the resulting subject lists and key metadata (such as seed and class distribution) to a pickle file, and then generated the corresponding train and test .CSV files for each dataset, ensuring that subjects remain consistent across all splits.

Tables 5.1 and 5.2 resume the examples distribution across training and test sets, respectively for binary and multiclass distribution.

Set	**Numerosity**		
	Examples	**Label 0**	**Label 1**
Train	251	11	240
Test	67	4	63

Table 5.1: Distribution of examples in the train and test sets for the binary case.

Set	Sample size				
	Examples	**Label 0**	**Label 1**	**Label 2**	**Label 3**
Train	249	11	61	131	46
Test	69	4	16	35	14

Table 5.2: Distribution of examples in the train and test sets for the multiclass case.

Following dataset partitioning, a set of feature selection methods, machine learning algorithms and dimensionality reduction methods was defined, allowing the evaluation of a large number of combinations in order to identify the configuration that yielded the best performance.

5.1 Feature Selection methods

Feature selection was performed in a dedicated script, *FeatureSelector*, which evaluates multiple univariate and model-based selection strategies. For each method, the script defines the subset of features considered informative for the classification problem. The function takes in input a feature matrix and a label vector, applies the configured selectors and stores the features in a dictionary. Two global hyperparameters control the feature selection:

- *alpha* (default 0.05): significance threshold for statistical-based selectors.
- threshold_variance (default 0.95): cumulative importance threshold for score-based selectors; features are ranked by their score and retained until the normalized cumulative score reaches this value.

In statistical-based selectors, features with p $\leq$ alpha were retained. For score-based methods, the smallest set of feature that produced the best normalized scores reaching the threshold defined were retained.

The feature selection methods implemented in the study, are shortly described:

- **None**: Baseline that performs no selection; returns all input features unchanged. Useful as a reference to compare against other methods.
- **ANOVA_F**: Univariate Analysis of Variance F-test computed independently for each feature against the class labels. Features with p-value $\leq \alpha$ are retained. Suitable for approximately normally distributed features with class-conditional mean differences.

- **CHI2**: Computes the chi-squared statistic to assess the dependence between each (non-negative) feature and the class labels. Since the test requires non-negative inputs, the feature matrix is transformed using $X = |X|$. Features with p-value $\leq \alpha$ are retained.
- **KruskalWallis (KW)**: This method applies the non-parametric Kruskal–Wallis H-test to evaluate each feature across the different classes. Because it does not assume normality and compares distributions based on their medians, it is well suited for data that are not Gaussian. Features with p-value $\leq \alpha$ are retained.
- **MUTUAL INFORMATION**: Estimates the mutual information between each feature and the class labels, allowing the detection of potentially non-linear relationships. Features are ranked according to their mutual information scores, and the smallest subset whose normalized cumulative score reaches *threshold_variance* is selected.
- **RF importance**: Trains a Random Forest classifier and uses its impurity-based feature importances as relevance scores. Features are sorted by importance, and the minimal subset required to reach *threshold_variance* of the total importance is kept.
- **LR coef**: Fits an L2-regularized Logistic Regression model and evaluates feature relevance through the mean absolute coefficient value across classes. Features are ranked accordingly, and the smallest subset reaching the *threshold_variance* criterion is selected.

5.2 Machine Learning Algorithms

The classification analysis was carried out using a set of machine learning algorithms. Due to the relatively small dataset size, deep learning approaches were deemed inappropriate. For each algorithm, a limited grid of hyperparameter values was specified to systematically evaluate multiple configurations and determine the combination achieving the optimal performance.

Minor differences were present between binary and multiclass problems, both in choice of machine learning and hyperparameters tuning. The following sections provide a brief description of each classifier implemented.

Logistic Regression (LR)

Logistic Regression is a linear probabilistic model that estimates the probability of class membership using the logistic function. It assumes a linear relationship

between predictors and the log-odds of the response variable, and applies regularization to prevent overfitting. Table 5.3 reports the selected hyperparameters and their corresponding values.

Hyperparameter	**Description**	**Binary values**	**Multiclass values**
penalty	Type of regularization	[l1, l2]	l2
C	Inverse regularization strength	[0.001, 0.01, 0.1, 0.5, 1, 5, 10, 100]	[0.01, 0.1, 1]
solver	Optimization algorithm	[liblinear, saga]	saga
class_weight	Balance of class weights	[None, balanced]	balanced

Table 5.3: Logistic Regression hyperparameters and value ranges.

Support Vector Machine (SVC)

The Support Vector Machine constructs a decision boundary that maximizes the margin between different classes. By employing kernel functions, it can efficiently handle both linear and non-linear classification problems. Table 5.4 reports the selected hyperparameters and their corresponding values.

Hyperparameter	**Description**	**Binary values**	**Multiclass values**
C	Regularization parameter	[0.01, 0.1, 1, 10, 100]	[0.01, 0.1, 1]
kernel	Kernel type	[linear, rbf, poly]	[linear, rbf]
gamma	Kernel coefficient	[scale, auto, 0.01, 0.1, 1]	[scale, 0.01]
degree	Degree (poly kernel)	[2, 3, 4]	\
class_weight	Balance of class weights	[None, balanced]	balanced

Table 5.4: SVC hyperparameters and value ranges.

K-Nearest Neighbors (KNN)

K-Nearest Neighbors is a non-parametric algorithm that assigns a class label to a new observation based on the majority class among its k closest neighbors, according to a chosen distance metric. Table 5.5 reports the selected hyperparameters and their corresponding values.

Hyperparameter	**Description**	**Binary values**	**Multiclass values**
n_neighbors	Number of neighbors k	range(1, 30, 2)	[5, 11]
weights	Neighbor weighting	[uniform, distance]	distance
p	Distance norm (1=Manhattan, 2=Euclidean)	[1, 2]	[1, 2]

Table 5.5: KNN hyperparameters and value ranges.

Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis is a statistical method that seeks linear combinations of features that best separate multiple classes. It assumes normally distributed data and equal covariance matrices across classes. Table 5.6 reports the selected hyperparameters and their corresponding values.

Hyperparameter	Description	Binary values	Multiclass values
solver	Solver for discriminant functions	[svd, lsqr, eigen]	[svd, lsqr]
shrinkage	Covariance shrinkage	[None, auto, 0.1, 0.5, 0.9]	[None, auto]

Table 5.6: LDA hyperparameters and value ranges.

Naïve Bayes Classifiers

Naïve Bayes classifiers are probabilistic models based on Bayes' theorem, built under the simplifying assumption that features are conditionally independent given the class label. In this study, three variants were implemented:

- **GaussianNB (GNB)**: assumes that continuous features follow a Gaussian distribution.
- **MultinomialNB (MNB)**: intended for features representing counts or event frequencies.
- **BernoulliNB (BNB)**: designed for features that take binary values.

Tables 5.7, 5.8 and 5.9 describes the hyperparameters selection and the values defined respectively for the Gaussian, Multinomial and Bernoulli NB.

Hyperparameter	Description	Binary values	Multiclass values
var_smoothing	Variance stabilization constant	[1e-12, 1e-11, 1e-10, 1e-9, 1e-8, 1e-7, 1e-6]	\

Table 5.7: Gaussian Naïve Bayes hyperparameters and value ranges.

Hyperparameter	Description	Binary values	Multiclass values
alpha	Additive smoothing	[0.01, 0.1, 0.5, 1.0, 2.0]	\
fit_prior	Learn class prior probabilities	[True, False]	\

Table 5.8: Multinomial Naïve Bayes hyperparameters and value ranges.

Hyperparameter	Description	Binary values	Multiclass values
alpha	Additive smoothing	[0.01, 0.1, 0.5, 1.0, 2.0]	[0.1, 0.5, 1.0]
binarize	Threshold for binarization	[0.0, 0.5, 1.0]	auto
fit_prior	Learn class prior probabilities	[True, False]	True

Table 5.9: Bernoulli Naïve Bayes hyperparameters and value ranges.

Decision Tree Classifier (DT)

The Decision Tree algorithm divides the feature space into increasingly homogeneous regions by performing recursive binary splits based on impurity measures such as the Gini index or entropy. Table 5.10 summarizes the selected hyperparameters and their corresponding values.

Hyperparameter	Description	Binary values	Multiclass values
max_depth	Maximum tree depth	[None, 3, 5, 10, 20, 50]	[3, 5, 8]
min_samples_split	Min. samples to split a node	[2, 5, 10, 20]	[5, 10, 20]
min_samples_leaf	Min. samples at a leaf	[1, 2, 4, 10]	[5, 10, 20]
criterion	Split quality function	[gini, entropy, log_loss]	[gini, entropy]
class_weight	Balance of class weights	[None, balanced]	\

Table 5.10: Decision Tree hyperparameters and value ranges.

Random Forest Classifier (RF)

Random Forest is an ensemble of decision trees trained on random subsets of data and features. It combines their predictions through majority voting, improving robustness and reducing overfitting. Table 5.11 reports the selected hyperparameters and their corresponding values.

Hyperparameter	Description	Binary values	Multiclass values
n_estimators	Number of trees	[100, 200, 300]	200
max_depth	Maximum tree depth	[None, 10, 20, 50]	[5, 10]
min_samples_split	Min. samples to split a node	[2, 5, 10]	[5, 10]
min_samples_leaf	Min. samples at a leaf	[1, 2, 4]	[5, 10]
criterion	Split quality function	[gini, entropy]	\
class_weight	Balance of class weights	[None, balanced]	balanced
max_features	Features per split	\	sqrt

Table 5.11: Random Forest hyperparameters and value ranges.

Balanced Random Forest (BRF)

Balanced Random Forest extends the standard Random Forest approach by addressing class imbalance through the use of balanced bootstrap samples, drawn so that each class is equally represented during the construction of each tree. Table 5.12 lists the selected hyperparameters along with their corresponding values.

Hyperparameter	Description	Binary values	Multiclass values
n_estimators	Number of trees	\	300
max_depth	Maximum tree depth	\	[5, 10]
max_features	Features per split	\	sqrt
min_samples_leaf	Min. samples at a leaf	\	[5, 10]

Table 5.12: Balanced Random Forest hyperparameters and value ranges.

Easy Ensemble Classifier

The Easy Ensemble method trains multiple base learners on different balanced subsets of the data, and aggregates their outputs to improve performance on imbalanced classification problems. The classifier was included as an imbalance-aware ensemble model. In this implementation, no hyperparameter tuning was performed, since the classifier is known to provide stable performance with its default configuration. The model used ten sub-ensembles (`n_estimators = 10`) and parallel computation (`n_jobs = -1`), ensuring a balanced representation of classes within each subset. This algorithms was implemented only in the multiclass case.

Multi-Layer Perceptron (MLP)

The Multi-Layer Perceptron is a feedforward artificial neural network composed of one or more hidden layers equipped with non-linear activation functions. This architecture allows the model to learn complex and non-linear relationships between the input features and the target labels. Table 5.13 presents the selected hyperparameters and their corresponding values.

Hyperparameter	Description	Binary values	Multiclass values
hidden_layer_sizes	Neurons per hidden layer	[(50,), (100,), (100,50), (150,100)]	[(50,), (100,)]
activation	Non-linear activation	[relu, tanh]	[relu, tanh]
solver	Optimization algorithm	adam	\
alpha	L2 regularization strength	[1e-5, 1e-4, 1e-3, 1e-2]	[0.001, 0.01, 0.1]
learning_rate	Learning rate schedule	[constant, adaptive]	\

Table 5.13: MLP hyperparameters and value ranges.

5.3 Dimensionality Reduction Methods

Dimensionality reduction techniques are employed to transform high-dimensional feature spaces into lower-dimensional representations while preserving as much relevant information as possible. These methods help mitigate the effects of the so-called 'curse of dimensionality', reduce computational complexity, and enhance model interpretability by removing redundant or noisy features. Moreover, dimensionality reduction can improve the generalization ability of machine learning algorithms, especially when the number of samples is limited compared to the number of features, as in this thesis work.

In this study, two classical linear approaches were implemented:

- **Principal Component Analysis (PCA)**: an unsupervised linear transformation method that projects the data onto a new coordinate system defined by orthogonal directions, called principal components, which maximize the variance of the projected data. PCA is widely used to capture the most informative variance structure in the data and to eliminate redundancy among correlated features.

- **Linear Discriminant Analysis for Dimensionality Reduction**: a supervised technique that identifies linear combinations of features that best discriminate between classes. Unlike PCA, which focuses solely on capturing variance, LDA maximizes class separability by optimizing the ratio between-class and within-class variance. When applied as a pre-processing step, LDA_DR yields a compact and discriminative representation of the data.

5.4 Classification pipeline

The Python script *generic_training* is responsible for executing both the training and testing phases of the machine learning model for a given dataset. First, the script automatically identifies the training and testing partition files associated with the dataset, which are expected to be stored in the same directory as the original data file. Then, the class label used for the classification is selected by choosing the appropriate column in the dataset (either 'class_binary' or 'class_multiclass', depending on the type of classification being performed).

A data cleaning step is subsequently applied to prepare the dataset for model learning. Specifically, metadata columns (*subject*, *segment_name*, *test_id*, *window_id*, *bedford_rating*, *difficulty_level*, *class_binary*, *class_multiclass*) are removed, as they do not carry discriminative spectral or physiological information relevant to the classification task. Following this, the script performs a NaN (Not a

Number) check to verify data consistency. In the present study, no missing values were detected in any of the datasets, and therefore no sample removal was required.

The feature selection is performed by invoking the *FeatureSelection* function, which returns the optimal subset of features to be utilized in the subsequent classification stage. All necessary parameters, including the selected feature subset, the desired classification type (binary or multiclass), the k-value of the cross validation, are fed into the separate module called *ClassificationPipeline*. Finally, the execution of the *run* method within this module performs the entire classification routine, providing the calculation and reporting the corresponding performance metrics.

Figure 5.1 illustrates the classification pipeline used for this work. For both binary and multiclass cases, the passages were the same.

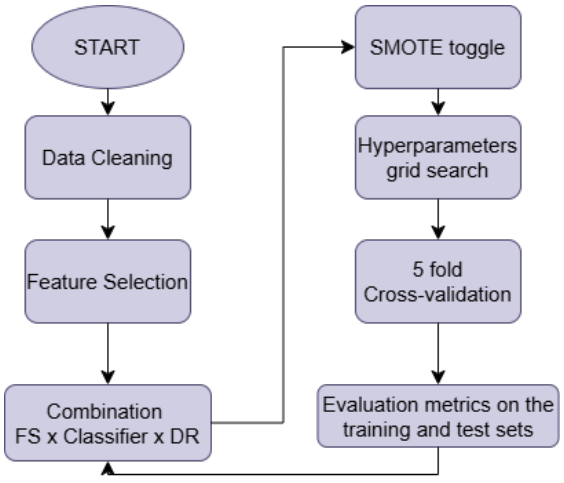


Figure 5.1: Classification pipeline implemented. FS = Feature Selection. DR = Dimensionality Reduction method.

After the data cleaning and feature selection, in *ClassificationPipeline* all valid combinations of feature selection methods, dimensionality reduction techniques and classifiers are generated. Combination involving tree-based classifiers (DT, RF, BRF and Easy Ensemble) and dimensionality reduction were excluded. Similarly, KNN and LDA_DR combination was omitted due to instability issues.

For each valid combination, the columns of the dataset corresponding to the selected features were extracted. If dimensionality reduction was active, it was applied and finally the classifier was implemented.

To address the class imbalance present in both the binary and multiclass settings, the Synthetic Minority Oversampling Technique (SMOTE) was integrated as an optional resampling step within the classification pipeline. SMOTE generates

synthetic samples for the minority class by interpolating between existing observations and their nearest neighbors, allowing the class distribution to be equalized without simply duplicating data. In this implementation, SMOTE was included as a parameter in the grid search (the "resample" option), with two possible configurations: applying SMOTE or disabling it. This design allows the cross-validation procedure to automatically determine whether class balancing has a positive or negative impact on model performance. Importantly, SMOTE was applied inside the pipeline during cross-validation, ensuring that synthetic samples were generated only within the training folds and preventing data leakage.

Following the SMOTE configuration, hyperparameter optimization was performed using a grid search strategy. This method systematically evaluates all feasible combinations of classifier hyperparameters under both resampling conditions, allowing cross-validation to identify the configuration that yields the most reliable performance.

Cross-validation was employed to obtain a reliable estimate of the generalization performance of each model. The training data were partitioned into k folds (with $k = 5$), and for each iteration, the model was trained on $k - 1$ folds and validated on the remaining one. This process was repeated until every fold had been used for validation, and the average performance across folds was taken as the final cross-validation score. This approach reduces the variance associated with a single train-test split, ensures that all samples are used for both training and validation, and provides a more robust criterion for model selection during hyperparameter optimization. Given that the information about subject division was available, a GroupKFold strategy was applied. This approach ensures that all samples belonging to the same subject appear entirely within one fold (either in the training set or in the validation set, but never in both). In other words, no subject's data is ever used simultaneously for training and validation. This constraint prevents data leakage across folds, which would otherwise occur if the model were allowed to learn subject-specific characteristics rather than generalizable features related to mental workload. Group-based validation therefore provides a stricter and more realistic measure of how the model would perform when applied to new, unseen participants.

The probabilistic classifiers (LR, SVC, and MLP) were wrapped inside a custom *PriorAdjustedClassifier* function. This wrapper adjusts the predicted probabilities according to predefined class priors, ensuring that all classes are treated as equally likely during inference. In practice, the base model first estimates the posterior probabilities, which are then rescaled by dividing each class probability by its prior and subsequently renormalizing the distribution. This correction reduces the tendency of models to favor more frequent classes and leads to fairer decision boundaries in imbalanced multiclass scenarios. The adjusted classifier was integrated directly into the grid-search pipeline, allowing its performance to be

evaluated alongside the other models during cross-validation.

The metrics implemented for general considerations and hyperparameter selection were accuracy and F1-score for the binary problem, and accuracy and macro F1-score for the multiclass problem. Each configuration inside the 5-fold cross validation was assessed using one metric at a time, and the combination that achieved the highest mean performance across all 5 folds was selected as the optimal configuration.

Once the best model was identified, the dedicated metrics were evaluated both on the training and test set, along with the generation of the confusion matrices. Results were saved in a ".pkl" format file for further evaluations.

Chapter 6

Results

In this section, the principal results obtained from the experimental analysis are presented. First, a statistical examination of the relationship between the assigned levels of task difficulty and the subjective difficulty ratings is conducted, highlighting the degree of coherence and the main trends observed. Subsequently, the performance results of the classification models are reported, considering both the binary and the multiclass configurations. These findings allow an assessment of the effectiveness of the proposed approach in identifying perceived difficulty levels and enable a comparison between different modeling strategies.

6.1 Difficulty levels analysis

To evaluate the validity of the experimental protocol and the proposed difficulty segmentation, a statistical analysis was conducted comparing the predefined difficulty levels with the participants' self-assessment using the Bedford workload scale.

Since both the experimental difficulty levels and the Bedford ratings range from 1 to 9 (the rating '10' of the Bedford scale was excluded from the analysis, as it was not assigned by any participant), the scores were grouped into the same three categories:

- **Easy**: scores from 1 to 3;
- **Medium**: scores from 4 to 6;
- **Hard**: scores from 7 to 9.

To assess the consistency between the experimentally assigned difficulty levels and the participants' perceived workload, the Bedford ratings were analysed using a non-parametric approach. Since each participant provided self-evaluations for

all three difficulty conditions (Easy, Medium, and Hard), the data exhibited a repeated-measures structure. Moreover, Bedford ratings are ordinal in nature and do not necessarily satisfy assumptions of normality. For these reasons, the Friedman test was adopted as the primary inferential method. This test represents the non-parametric equivalent of a repeated-measures ANOVA and evaluates whether systematic differences exist across three or more related conditions, while appropriately accounting for within-subject variability.

The Friedman test was used on the subject-wise mean Bedford ratings for each difficulty level. The analysis revealed a significant effect of task difficulty on perceived workload ($F = 32.00, ; p = 1.125 \times 10^{-7}$). This result shows that the distributions of subjective ratings were different across the Easy, Medium, and Hard conditions. With such a small p-value, it is very unlikely that these differences occurred by chance. Looking at the median ratings, there was a clear increase in reported workload as task demands rose. Participants indicated a higher workload with increasing difficulty. These findings confirm that the difficulty manipulation created distinct subjective workload levels, supporting the study's internal validity.

To better understand the relationship between difficulty and perceived workload, we calculated a Spearman rank-order correlation. Spearman's ρ assesses the strength of a monotonic association using ranked values, making it suitable for ordinal subjective ratings. The analysis showed a positive and statistically significant correlation ($\rho = 0.462, ; p = 2.074 \times 10^{-17}$). This means that higher difficulty levels were consistently associated with higher Bedford scores. This trend supports the results of the Friedman test, indicating that participants perceived greater mental effort in more challenging conditions.

Overall, these results show that manipulating task difficulty was successful. Increases in objective task demands matched with clear and significant increases in perceived workload. This connection between difficulty and subjective experience enhances the internal validity of the study and lays a strong foundation for the following neurophysiological and machine-learning analyses.

In addition to the inferential analyses, a boxplot representation was used to visually examine the distribution of the Bedford ratings across the three difficulty levels (Easy, Medium, Hard). The plot displays the median, interquartile range, and variability of the subjective workload scores within each condition. As shown in Figure 6.1, the median Bedford ratings increased progressively from the Easy to the Hard condition, and the overall distribution shifted upward with task difficulty. Although some overlap between groups was present, reflecting individual variability in perceived workload, the general trend clearly indicated higher subjective effort in more demanding tasks. This graphical evidence is consistent with the results of the Friedman and Spearman analyses, providing additional support for the effectiveness of the difficulty manipulation in eliciting distinct levels of perceived mental workload.

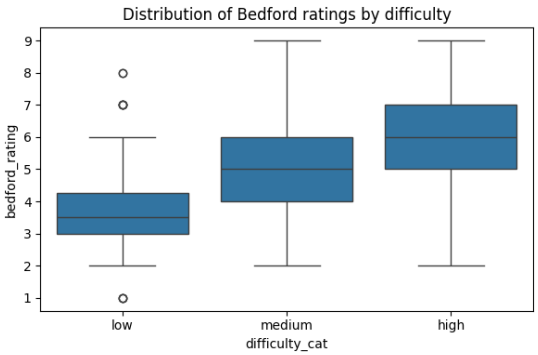


Figure 6.1: Distribution of Bedford ratings in the low, medium and high difficulty groups.

Finally, Figure 6.2 shows the coherence matrix between the imposed difficulty levels and the participants' self-evaluations, highlighting the degree of correspondence between the two measures. Figure 6.3 offers a comparison between the distribution of the imposed difficulty levels (Figure 6.3a) and that of the subjective ratings (Figure 6.3b), providing a direct visual indication of how closely the perceived workload reflects the experimental manipulation.

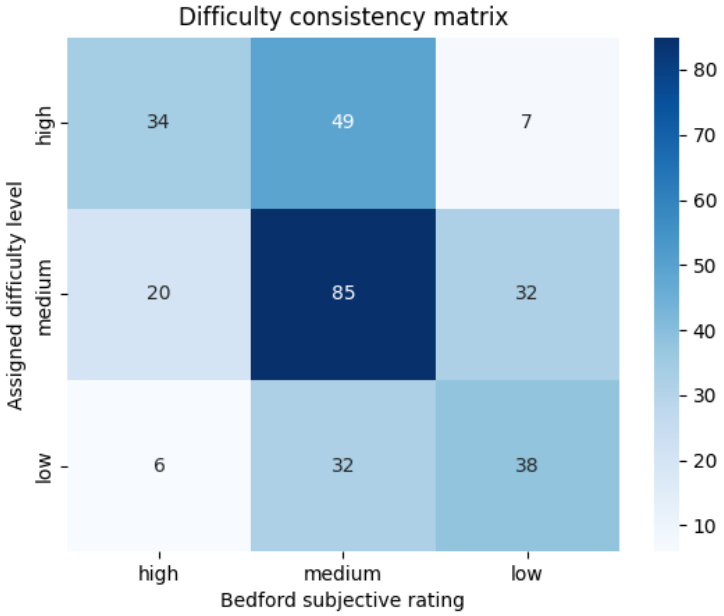
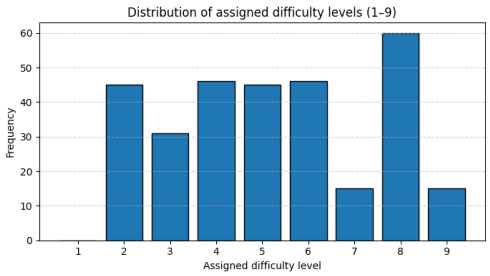
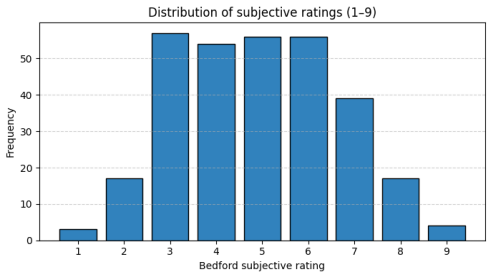


Figure 6.2: Coherence matrix between assigned and perceived difficulty.



(a) Distribution of assigned difficulty levels.



(b) Distribution of subjective workload ratings.

Figure 6.3: Comparison between the distribution of assigned difficulty levels (left) and the corresponding subjective workload ratings (right).

The mean subjective evaluation values across the three imposed difficulty groups (low, medium, high) are illustrated in Table 6.1. As can be observed, participants generally tended to overestimate workload in conditions classified as low difficulty, and underestimate it in conditions classified as high difficulty. This pattern results in a relative inflation of the medium-level group, effectively concentrating most ratings around the mid-range of the scale. Consequently, this behavior introduces an imbalance in the class distribution, which carries implications for the subsequent multiclass classification analysis.

Assigned Difficulty Level	**Mean Bedford Rating**	**SD**
Low	3.80	1.58
Medium	4.82	1.54
High	5.96	1.52

Table 6.1: Mean and standard deviation of Bedford subjective workload ratings for each assigned difficulty level.

6.2 Classification metrics

To evaluate the performance of the binary and multiclass problems, two metrics were considered. The first metric is **accuracy**, which measures the proportion of correctly classified samples out of the total number of samples. It provides an overall indication of how often the model's predictions match the true labels, independently of the class distribution. Accuracy is defined as:

$$Accuracy = \frac{\text{Number of correctly classified samples}}{\text{Total number of samples}}.$$

In terms of confusion matrix, it can be also expressed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where:

- **TP**: number of positive samples correctly classified.
- **TN**: number of negative samples correctly classified.
- **FP**: number of negative samples incorrectly classified as positive.
- **FN**: number of positive samples incorrectly classified as negative.

The second metric selected was the **F1-score**, a performance metric that represents the harmonic mean between precision and recall. It provides a single value that balances both false positives and false negatives, making it particularly useful when dealing with imbalanced datasets. It is formulated as:

$$F1_{\text{score}} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}};$$

where precision and recall are defined as:

$$\text{Precision} = \frac{TP}{TP + FP},$$

$$\text{Recall} = \frac{TP}{TP + FN}.$$

For the multiclass problem, in presence of more classes, the **macro F1-score** was selected instead of the F1-score. The macro F1-score is computed as the arithmetic mean of the F1 scores of all N classes:

$$F1_{\text{macro}} = \frac{1}{N} \sum_{i=1}^{N} F1_i = \frac{1}{N} \sum_{i=1}^{N} 2 \cdot \frac{\text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

Unlike the F1-score, the macro F1 treats all classes equally, giving the same weight to minority and majority classes. This makes it particularly useful when evaluating performance in imbalanced datasets.

6.3 Binary classification

For each dataset partition, the optimal combination of classifier, feature-selection method, and dimensionality-reduction technique was determined by examining the performance metrics on the test set together with the model's generalization ability, quantified as the difference between training and testing metrics. Particular attention was given to the **F1-score** on the test set, since this metric is less affected by class imbalance and offers a more reliable indication of the model's discriminative performance.

Table 6.2 describes the results selecting the best performances for each division of the dataset. These results were obtained with multiple combination of classifier - feature selection - dimensionality reduction, demonstrating the strength of the difference in mental workload and rest state through the use of frontal EEG data. The binary classification analysis yielded consistently high performance across all feature groups and time windows, with F1-scores exceeding 0.97 in every configuration.

No substantial differences were observed among the three segmentation windows (win2, win3, win4), suggesting that the temporal length of the analyzed segment does not significantly affect classification performance.

Overall, coherence features exhibited the best results, with an F1-score on the test equal to 99.21% on each window size.

Dataset	**AccTrain**	**F1Train**	**AccTest**	**F1Test**
'window2_coherence'	1	1	0.9851	0.9921
'window2_spectral'	0.9801	0.9895	0.9851	0.9921
'window3_coherence'	0.9801	0.9897	0.9851	0.9921
'window3_complexity'	0.9801	0.9896	0.9851	0.9921
'window4_coherence'	0.9960	0.9979	0.9851	0.9921
'window2_temporal'	1	1	0.9701	0.9844
'window3_spectral'	0.9761	0.9874	0.9701	0.9844
'window3_temporal'	0.9841	0.9917	0.9701	0.9844
'window4_spectral'	1	1	0.9701	0.9844
'window4_temporal'	0.9801	0.9897	0.9701	0.9844
'window2_complexity'	1	1	0.9552	0.9767
'window4_complexity'	1	1	0.9552	0.9767

Table 6.2: Results of the binary classification.

Tables 6.3, 6.4, 6.5, 6.6 and 6.7 shows the combination that produced the best results for each of the dataset division that produced the best F1-score on the test set.

Model	DR	FS	AccTrain	F1Train	AccTest	F1Test
SVC	None	None	0.9960	0.9979	0.9851	0.9921
SVC	PCA	None	0.9920	0.9959	0.9851	0.9921
GNB	PCA	None	0.9761	0.9876	0.9851	0.9921
KNN	None	ANOVA	1.0000	1.0000	0.9851	0.9921
KNN	None	KW	1.0000	1.0000	0.9851	0.9921
LR	PCA	RF	0.9801	0.9897	0.9851	0.9921
GNB	PCA	RF	0.9801	0.9897	0.9851	0.9921
GNB	PCA	LR	0.9761	0.9877	0.9851	0.9921
GNB	None	None	0.9641	0.9810	0.9851	0.9921

Table 6.3: Results for the binary classification with 'window2_coherence' dataset. DR = dimensionality reduction, FS = Feature Selection.

Model	DR	FS	AccTrain	F1Train	AccTest	F1Test
GNB	None	None	0.9641	0.9810	0.9851	0.9921
GNB	None	LR_coef	0.9801	0.9895	0.9851	0.9921

Table 6.4: Results for the binary classification with 'window2_spectral' dataset

Model	DR	FS	AccTrain	F1Train	AccTest	F1Test
LR	PCA	ANOVA_F	0.9801	0.9897	0.9851	0.9921

Table 6.5: Results for the binary classification with 'window3_coherence' dataset

Model	DR	FS	AccTrain	F1Train	AccTest	F1Test
LDA	None	CHI2	0.9801	0.9896	0.9851	0.9921
GNB	LDA_DR	CHI2	0.9801	0.9896	0.9851	0.9921
LDA	LDA_DR	CHI2	0.9801	0.9896	0.9851	0.9921

Table 6.6: Results for the binary classification with 'window3_complexity' dataset

Model	DR	FS	AccTrain	F1Train	AccTest	F1Test
LR	None	None	0.9960	0.9979	0.9851	0.9921
SVC	None	None	0.9920	0.9959	0.9851	0.9921
LR	PCA	None	0.9920	0.9959	0.9851	0.9921
SVC	PCA	None	0.9841	0.9917	0.9851	0.9921
LR	None	ANOVA_F	0.9880	0.9938	0.9851	0.9921
GNB	PCA	ANOVA_F	0.9801	0.9897	0.9851	0.9921
LDA	PCA	ANOVA_F	0.9880	0.9938	0.9851	0.9921
GNB	PCA	KW	0.9721	0.9855	0.9851	0.9921
LR	None	M_INFO	0.9960	0.9979	0.9851	0.9921
SVC	None	M_INFO	0.9841	0.9917	0.9851	0.9921
LR	PCA	M_INFO	0.9841	0.9917	0.9851	0.9921
SVC	PCA	M_INFO	0.9880	0.9938	0.9851	0.9921
LDA	PCA	M_INFO	0.9880	0.9938	0.9851	0.9921

Table 6.7: Results for the binary classification with 'window4_coherence' dataset

Figure 6.4 displays the confusion matrix for the test set, which is provided as output for each combination that yielded the best performance (99.21% F1-score on the test set)

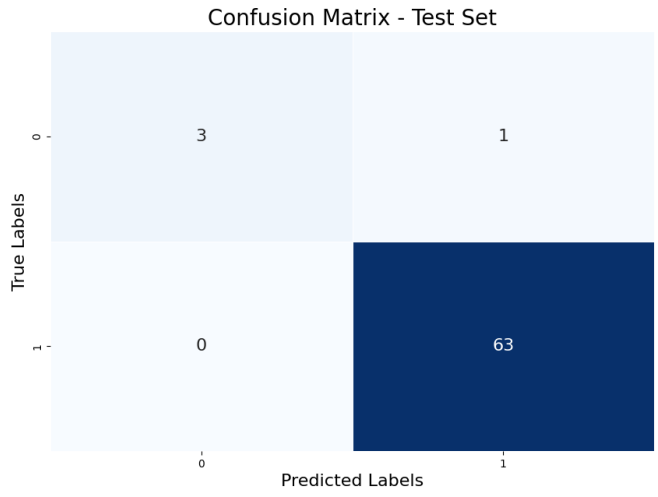


Figure 6.4: Confusion matrix for the test set of the best classifier combination.

6.4 Multiclass classification

The same considerations discussed for the binary classification case were extended to the multiclass classification setting. In this context, particular attention was devoted to the macro F1-score, previously described in Section 6.2.

Table 6.8 shows the results from selecting the best model setup for each dataset division. As expected, overall performance dropped compared to the binary classification scenario. The highest macro F1-score on the test set was 55.69%, achieved with spectral features using a 4-second window. The highest accuracy reached 56.52%, which came from using temporal features with the same window size.

Overall, the 4-second window produced the two best-performing models when combined with spectral and temporal features. This suggests that a longer window gives better access to informative EEG patterns. Among the feature groups, spectral features showed the best performance in the multiclass setting, highlighting the importance of frequency-domain information for evaluating different levels of cognitive load.

Dataset	**AccTrain**	**F1Train**	**AccTest**	**F1Test**
'window4_spectral'	0.9920	0.9933	0.5217	0.5569
'window4_temporal'	0.8153	0.7794	0.5652	0.5228
'window2_spectral'	0.9920	0.9933	0.4493	0.4645
'window2_coherence'	0.6827	0.7257	0.4493	0.4475
'window3_complexity'	1	1	0.4050	0.4392
'window2_complexity'	0.8755	0.8522	0.4928	0.4257
'window3_coherence'	0.3534	0.3280	0.4638	0.4177
'window4_coherence'	0.6145	0.6469	0.5362	0.4148
'window3_spectral'	0.7550	0.7592	0.4493	0.4010
'window4_complexity'	0.7791	0.7806	0.5072	0.4004
'window3_temporal'	0.8755	0.8955	0.4348	0.3964
'window2_temporal'	0.6747	0.6555	0.3333	0.3847

Table 6.8: Results of the multiclass classification.

Tables 6.9 and 6.10 illustrate the classifier–feature selector–dimensionality reduction combinations that yielded the best results for the two dataset divisions highlighted in green in Figure 6.8. The Mutual Information feature selection method retained 256 features from the original dataset, while the LDA_DR dimensionality reduction technique extracted two principal components. For the temporal dataset, the RF_importance method selected 117 features from the original dataset after applying SMOTE resampling.

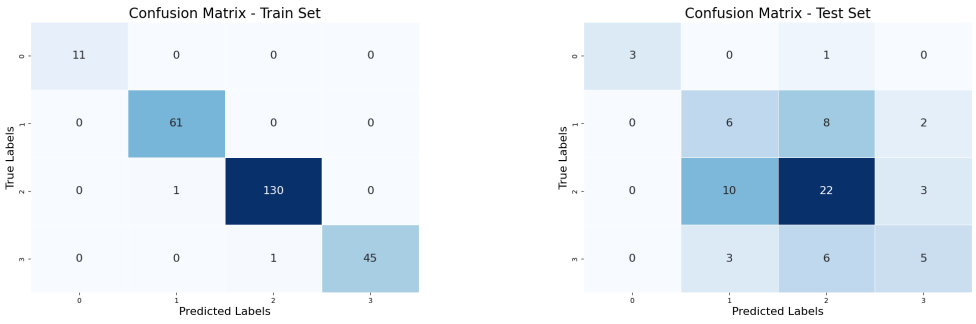
Model	DR	FS	AccTrain	F1Train	AccTest	F1Test
LR	LDA	MI	0.9920	0.9933	0.5217	0.5569

Table 6.9: Results for the multiclass classification with 'win4_spectral' dataset

Model	DR	FS	AccTrain	F1Train	AccTest	F1Test
DT	None	RF	0.8153	0.7794	0.5652	0.5228

Table 6.10: Results for the multiclass classification with 'win4_temporal' dataset

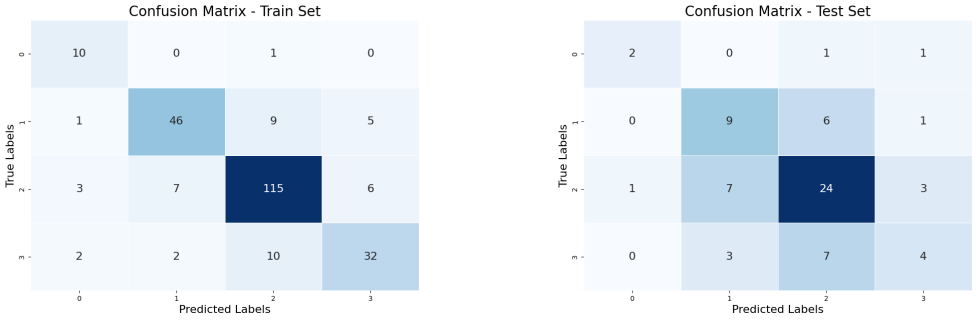
Although several strategies were implemented to mitigate class imbalance and overfitting, such as the use of cross-validation, overfitting still had a noticeable impact on the multiclass classification problem. This effect is particularly evident in Table 6.9, where a performance gap of approximately 45% between the training and test sets highlights the model's limited generalization capability. Figures 6.5 and 6.6 show the confusion matrices for the 'window4_spectral' and 'window4_temporal' feature sets. As can be observed, most misclassifications occurred between adjacent classes, suggesting that the model could effectively distinguish between low and high workload states but struggled with intermediate transitions.



(a) Confusion matrix for the train set of the 'window4_spectral' dataset.

(b) Confusion matrix for the test set of the 'window4_spectral' dataset.

Figure 6.5: Confusion matrices for the 'window4_spectral' dataset that produced the highest macro F1-score.



(a) Confusion matrix for the train set of the 'win4_temporal' dataset.

(b) Confusion matrix for the test set of the 'win4_temporal' dataset.

Figure 6.6: Confusion matrices for the 'win4_temporal' dataset that produced the highest accuracy.

Several factors may explain this behavior. First, the lack of an objective ground truth was important because the models used self-assessment ratings as class labels. This method depends on participants being familiar with the Bedford scale and being able to apply its definitions consistently and fairly. When these conditions are not met, differences occur between physiological measurements and personal evaluations. This leads to inconsistent model training and makes it harder to separate classes. Additionally, as shown in Figure 6.3b, many ratings were assigned to boundary values between workload groups, which further blurs the distinctions among classes.

Another factor that may have contributed to the results is the limited optimization of the machine learning algorithms. Since the goal of the study was to compare a wide range of classifiers and identify the most suitable modeling approach, the depth of hyperparameter tuning was necessarily constrained. Although a grid search was employed, the parameter ranges were intentionally kept narrow to reduce computational costs, which prevented a more exhaustive optimization of each model.

Finally, the small sample size may also have influenced results. Although each subject provided multiple signal segments, the overall number of segment associated to each class may have been insufficient for the models to extract EEG patterns across subjects. As a result, inter-subject variability limited the models' ability to capture physiological information associated with different workload states.

Chapter 7

Discussion

The results presented in the previous chapter provide meaningful insights into the relationship between frontal EEG activity and mental workload. The main objective of this thesis was to design a complete pipeline, starting from a solid experimental protocol for the acquisition of frontal EEG data under different workload conditions, to analyzing which window size and feature set produced the best predictions. EEG was selected as the primary source of information because it provides a direct measurement of brain activity and responds rapidly to changes in cognitive demand. These characteristics make it particularly suitable for developing models capable of real-time workload classification, which could enable human–machine systems to promptly intervene in critical situations. The exclusive use of frontal electrodes was motivated by both functional and practical considerations. The frontal and prefrontal cortices are known to be involved in higher-order cognitive processes, behavioral control, and action planning. Moreover, this region of the scalp is the most suitable for electrode placement, as it minimizes interference from hair and reduces electrode–skin impedance.

The experimental protocol proposed in this study proved effective in collecting a large number of self-evaluations within a relatively short time, allowing the construction of an extensive dataset for the subsequent machine learning analyses. The MATB-II software successfully simulated realistic operational conditions by placing participants in multitasking scenarios in which they had to manage several concurrent subtasks while also completing an additional arithmetic activity. Participants responded positively to the experimental procedure, noting that the overall duration of the test felt appropriate and that adding further sequences would likely have resulted in excessive fatigue.

Statistical analysis further confirmed the validity of the proposed protocol. Indeed, Bedford self-evaluations increased significantly as the imposed task difficulty rose, demonstrating that the progressive manipulation of task parameters effectively induced higher workload levels. This trend is consistent with findings reported

in the literature, supporting the reliability of the experimental design in eliciting measurable variations in cognitive load.

The classification results for the binary case were excellent. Several classifiers achieved an F1-score of 99.21% on the test set, showing a great ability of the EEG features to distinguish between resting and workload conditions. No clear pattern appeared among the different window sizes or feature groups. This suggests that neither the timing nor the specific kinds of features significantly affected performance for this task.

It is noteworthy that, between the best and the second-best classifier configurations (with F1-scores of 99.21% and 98.44%, respectively), the only misclassification involved a resting-state instance incorrectly assigned to the workload class. Considering the pronounced imbalance in the dataset (where the majority of examples belonged to the workload condition) this single improvement in correctly identifying a rest sample represents a substantial enhancement in the overall classification capability.

The multiclass classification results highlight the difficulty of differentiating between multiple levels of workload. Spectral features with a 4-second sliding window yielded the best performance, achieving the most macro F1 score (55.69%) on the test set. These macro F1 scores are slightly lower than values typically reported in the literature (typically 10% [51]).

However, it is noteworthy that most of the studies in the literature focused on three workload classes (specifically rest, low workload and high workload) and generally employed deep learning techniques; however, these techniques were not feasible for our study, as there were not enough available EEG recordings for our analysis. Also, due to the highly subjective nature of this task, there can be significant variability between self-reported workload levels and the measured physiological responses associated with workload. This limitation did not have any significative impact on the binary classification task, as there was a clear divide between the two classes of rest and workload. Despite these challenges, it is encouraging to note that the vast majority of the low workload and high workload misclassifications were very few, meaning the models clearly identified the two most distinct cognitive states (e.g. easy vs hard tasks).

Among all dataset configurations, the 4-second window generally produced better results, indicating that longer temporal segments allow more stable and informative representations of brain activity. Spectral features, in particular, proved to be the most effective, confirming the relevance of frequency-domain information in capturing variations in cognitive load, as also reported in previous studies on mental workload assessment.

Overall, these findings show that the experimental framework and processing pipeline developed in this thesis offer a solid foundation for EEG-based workload classification. Although additional refinements are needed to enhance generalization

and improve the discrimination of multiple workload levels, the results represent a promising step toward the development of adaptive human–machine systems capable of monitoring cognitive states in real time.

Chapter 8

Conclusion

Mental workload (MWL) plays a central role in shaping human performance in complex and multitasking environments. As discussed throughout this work, understanding how cognitive demand fluctuates during task execution is essential for designing systems that are safer, more efficient, and genuinely human-centered. Both excessive and insufficient workload can impair attention, decision-making, and situational awareness, particularly in safety-critical domains such as aviation, healthcare, and transportation. For these reasons, developing reliable techniques for monitoring MWL is a key step toward implementing adaptive human–machine interfaces that can adjust dynamically to the operator's cognitive state.

Electroencephalography (EEG) offers a powerful means of achieving this goal, as it provides a direct measurement of neural activity with excellent temporal resolution. Compared with behavioral indicators or self-report scales, EEG delivers objective and continuous insights into cognitive processes, allowing the detection of subtle variations in mental effort that may not be consciously perceived by the operator. In this thesis, frontal EEG was chosen deliberately due to its strong involvement in executive functions, attention, and working memory, as well as its practical advantages: electrodes placed over frontal areas cause minimal discomfort and avoid interference with the subject's hair, making this configuration well suited for real-world applications.

To induce controlled variations in workload, this study employed the Multi-Attribute Task Battery II (MATB-II), a multitasking simulation widely used in aviation research for assessing operator performance. By manipulating the frequency and complexity of events across subtasks, nine difficulty levels were defined and organized into structured sequences. Participants also completed a continuous secondary arithmetic task to further increase cognitive demand. Subjective workload ratings were collected after each task window using the Bedford Workload Scale, selected for its suitability for rapid, near real-time assessments.

A total of 18 volunteers participated in the experiment, with 16 subjects included in the final analysis after excluding recordings with corrupted data. This dataset supported a complete signal-processing pipeline that included artifact removal, segmentation into multiple window lengths (2, 3, and 4 seconds), and extraction of four families of features: temporal, spectral, complexity, and coherence features. Multiple machine learning algorithms, feature-selection strategies, and dimensionality-reduction methods were tested to identify the most effective configuration for workload classification.

The results showed that binary classification between rest and workload could be performed with very high reliability, reaching 98.51% accuracy and 99.21% F1-score on the test set for the best coherence-based models. These findings confirm a clear separability between resting-state EEG and cognitively demanding conditions within the frontal cortex. In contrast, multiclass classification (distinguishing rest, low, medium, and high workload levels) proved more challenging, with a maximum macro-F1 of 55.69% obtained using spectral features and 4-second windows. This reduction in performance reflects the intrinsic difficulty of predicting subjective workload levels, the class imbalance present in the dataset, and the more subtle physiological differences between adjacent workload states.

8.1 Limitations and future work

Despite the promising results obtained, several limitations of this study should be acknowledged. First, the relatively small number of participants limited the statistical power of the analyses and reduced the generalizability of the classification models. Increasing the sample size and gender distribution would help capture a broader range of inter-individual variability in EEG responses and improve the robustness of future models. Moreover, the use of self-assessment ratings as ground truth introduced a degree of subjectivity that may have led to inconsistencies between perceived and physiological workload levels. Incorporating objective performance indicators or additional behavioral measures could strengthen the reliability of class labeling. A future solution to this problem could be the implementation of performance-based correction, where ratings too inconsistent with the errors made and reaction time could be corrected, shifting the evaluation between labels.

Another limitation concerns the exclusive reliance on EEG data. Although EEG provides direct insight into neural activity, incorporating additional physiological modalities such as electrodermal activity, heart rate variability, or eye-tracking measures could yield a more comprehensive picture of mental workload. Moreover, only traditional machine learning algorithms were used in this study. Implementing deep learning architectures, such as convolutional or recurrent neural networks, could allow the automatic extraction of complex temporal and spatial EEG patterns,

potentially improving multiclass classification performance.

Finally, the short duration of each test phase (approximately two minutes) may have constrained the differences between the predefined difficulty levels. Because level 9 involved the maximum number of events that could reasonably occur within this time window, the temporal limitation reduced the range of task demands that could be imposed. Extending the duration of each phase would likely create clearer distinctions between workload conditions and allow additional features to emerge, which could enhance discrimination among difficulty levels.

Thus, although the selected length may not represent the optimal, it was determined to be an appropriate compromise of practical feasibility and the requirement to obtain adequate numbers of Bedford ratings from all subjects. Increasing the task durations would have resulted in overly long overall session durations that would likely lead to elevated levels of fatigue in many subjects.

In future research, improvements can be made through utilization of adaptive, person-specific models designed to address the cognitive processing differences of all subjects. Additionally, another avenue of future work would be to examine the incorporation of the proposed framework in real time systems so that human-to-machine interfaces may dynamically adapt both task and/or automation levels based upon the cognitive state or level of mental engagement of the operator.

Bibliography

[1] Richard S Lazarus. *Stress, appraisal, and coping*. Vol. 445. Springer, 1984 (cit. on p. 1).

[2] Yixiang Lim, Subramanian Ramasamy, Alessandro Gardi, Trevor Kistan, and Roberto Sabatini. «Cognitive human-machine interfaces and interactions for unmanned aircraft». In: *Journal of Intelligent & Robotic Systems* 91.3 (2018), pp. 755–774 (cit. on p. 1).

[3] National Cancer Institute. *SEER Training Modules: Nervous System*. Accessed: 2025-08-05. 2025. URL: https://training.seer.cancer.gov/anatomy/nervous/ (cit. on p. 4).

[4] Lauren Thau, Vamsi Reddy, and Paramvir Singh. «Anatomy, central nervous system». In: 2019 (cit. on p. 5).

[5] Kenia A Maldonado and Khalid Alsayouri. «Physiology, brain». In: *StatPearls [Internet]*. StatPearls Publishing, 2023 (cit. on pp. 5, 6).

[6] Cindy L. Stanfield. «Fisiologia». In: 5th ed. Milano: Pearson, 2020. Chap. 9, pp. 215–251. ISBN: 9788891910649 (cit. on pp. 5, 6).

[7] Alan Woodruff. *What is a neuron?* Accessed: 2025-08-06. Aug. 2019 (cit. on p. 8).

[8] Eric R. Kandel, James H. Schwartz, Thomas M. Jessell, Steven A. Siegelbaum, and John H. Byrne. *Principles of Neural Science*. 6th ed. New York: McGraw-Hill, 2021 (cit. on p. 8).

[9] Aeyal Raz and Misha Perouansky. «Chapter 7 - Central Nervous System Physiology: Neurophysiology». In: *Pharmacology and Physiology for Anesthesia*. Ed. by Hugh C. Hemmings and Talmage D. Egan. Philadelphia: W.B. Saunders, 2013, pp. 103–122. ISBN: 978-1-4377-1679-5. DOI: https://doi.org/10.1016/B978-1-4377-1679-5.00007-7. URL: https://www.sciencedirect.com/science/article/pii/B9781437716795000077 (cit. on p. 8).

[10] Shaochuan Chen, Teng Zhang, Stefan Tappertzhofen, Yuchao Yang, and Ilia Valov. «Electrochemical-Memristor-Based Artificial Neurons and Synapses—Fundamentals, Applications, and Challenges». In: *Advanced materials (Deerfield Beach, Fla.)* 35 (July 2023), e2301924. DOI: 10.1002/adma.202301924 (cit. on p. 9).

[11] Michael H. Grider, Rishita Jessu, and Rian Kabir. «Physiology, Action Potential». In: *StatPearls [Internet]*. Updated 2023 May 8. Treasure Island (FL): StatPearls Publishing, Jan. 2025. URL: https://www.ncbi.nlm.nih.gov/books/NBK538143/ (cit. on p. 10).

[12] Katarzyna Blinowska and Piotr Durka. «Electroencephalography (eeg)». In: *Wiley encyclopedia of biomedical engineering* 10 (2006), p. 9780471740360 (cit. on pp. 11, 14, 16).

[13] Jeffrey W. Britton, Lauren C. Frey, Jennifer L. Hopp, and et al. *Electroencephalography (EEG): An Introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children, and Infants*. Ed. by Erik K. St. Louis and Lauren C. Frey. Appendix 6. A Brief History of EEG. Chicago: American Epilepsy Society, 2016. URL: https://www.ncbi.nlm.nih.gov/books/NBK390348/ (cit. on p. 12).

[14] Pankaj Pandey, Richa Tripathi, and Krishna Miyapuram. «Classifying oscillatory brain activity associated with Indian Rasas using network metrics». In: *Brain Informatics* 9 (Dec. 2022). DOI: 10.1186/s40708-022-00163-7 (cit. on p. 13).

[15] Narayan P. Subramaniyam. *Wet and Dry Electrodes for EEG*. Accessed: 2025-09-03. June 2021. URL: https://sapienlabs.org/lab-talk/wet-and-dry-electrodes-for-eeg/ (cit. on pp. 14, 15).

[16] Daniel Silverman. «The Rationale and History of the 10-20 System of the International Federation». In: *American Journal of EEG Technology* 3.1 (1963), pp. 17–22. DOI: 10.1080/00029238.1963.11080602. eprint: https://doi.org/10.1080/00029238.1963.11080602. URL: https://doi.org/10.1080/00029238.1963.11080602 (cit. on p. 14).

[17] Junaid Ahmed. «Brain Machine Interface using EEG Sci-fi to Reality Neural Interface Engineering Brain Machine Interface using EEG 1 BRAIN MACHINE INTERFACE USING EEG». In: (Dec. 2016) (cit. on p. 16).

[18] Jose Antonio Urigüen and Begoña Garcia-Zapirain. «EEG artifact removal—state-of-the-art and guidelines». In: *Journal of neural engineering* 12.3 (2015), p. 031001 (cit. on p. 17).

[19] Kevin T Sweeney, Tomás E Ward, and Seán F McLoone. «Artifact removal in physiological signals—Practices and possibilities». In: *IEEE transactions on information technology in biomedicine* 16.3 (2012), pp. 488–500 (cit. on p. 17).

[20] Learning EEG. *Artifacts — Learning EEG*. `https://www.learningeeg.com/artifacts`. [accessed 3 Nov 2025]. n.d. (Cit. on pp. 18, 20).

[21] Ph.D. Cross Villasana Fernando. *Getting to know EEG artifacts and how to handle them in BrainVision Analyzer 2*. Press Release, Brain Products GmbH. Accessed: 2025-11-05. Dec. 2022. URL: `https://pressrelease.brainproducts.com/eeg-artifacts-handling-in-analyzer/` (cit. on p. 19).

[22] Peter A Hancock and Najmedin Meshkati. *Human mental workload*. Vol. 52. North-Holland Amsterdam, 1988 (cit. on p. 20).

[23] Alex Dan, Miriam Reiner, et al. «Real time EEG based measurements of cognitive load indicates mental states during learning». In: *Journal of Educational Data Mining* 9.2 (2017), pp. 31–44 (cit. on pp. 21, 27).

[24] John Sweller. «Cognitive load during problem solving: Effects on learning». In: *Cognitive science* 12.2 (1988), pp. 257–285 (cit. on p. 21).

[25] Peter Gerjets, Katharina Scheiter, and Richard Catrambone. «Designing instructional examples to reduce intrinsic cognitive load: Molar versus modular presentation of solution procedures». In: *Instructional Science* 32.1 (2004), pp. 33–58 (cit. on p. 21).

[26] John Sweller. «Element interactivity and intrinsic, extraneous, and germane cognitive load». In: *Educational psychology review* 22.2 (2010), pp. 123–138 (cit. on p. 21).

[27] Fred Paas, Alexander Renkl, and John Sweller. «Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture». In: *Instructional science* 32.1/2 (2004), pp. 1–8 (cit. on p. 21).

[28] Peter A Hancock. «The effect of performance failure and task demand on the perception of mental workload». In: *Applied Ergonomics* 20.3 (1989), pp. 197–205 (cit. on p. 22).

[29] {Mark S.} Young, {Karel A.} Brookhuis, {Christopher D.} Wickens, and {Peter A.} Hancock. «State of science: mental workload in ergonomics». English. In: *Ergonomics* 58.1 (Jan. 2015), pp. 1–17. ISSN: 0014-0139. DOI: `10.1080/00140139.2014.956151` (cit. on p. 22).

[30] Gerald Matthews and D Roy Davies. «Individual differences in energetic arousal and sustained attention: A dual-task study». In: *Personality and individual Differences* 31.4 (2001), pp. 575–589 (cit. on p. 22).

[31] MohammadReza Safari, Reza Shalbaf, Sara Bagherzadeh, and Ahmad Shalbaf. «Classification of mental workload using brain connectivity and machine learning on electroencephalogram data». In: *Scientific Reports* 14.1 (2024), p. 9153 (cit. on p. 23).

[32] Brad Cain. «A review of the mental workload literature». In: (2007) (cit. on pp. 23, 26).

[33] Peter Hoonakker, Pascale Carayon, Ayse P Gurses, Roger Brown, Adjhaporn Khunlertkit, Kerry McGuire, and James M Walker. «Measuring workload of ICU nurses with a questionnaire survey: the NASA Task Load Index (TLX)». In: *IIE transactions on healthcare systems engineering* 1.2 (2011), pp. 131–143 (cit. on p. 23).

[34] National Aeronautics and Space Administration. *Cognitive Workload Assessment Methods*. NASA-STD-3001 Technical Brief (OCHMO–TB–032), Rev. C. NASA Office of the Chief Health & Medical Officer (OCHMO), Dec. 2023. URL: https://www.nasa.gov/wp-content/uploads/2023/12/ochmo-tb-032-cognitive-workload.pdf (cit. on p. 24).

[35] Sarah Miller. «Workload measures». In: *National Advanced Driving Simulator. Iowa City, United States* (2001) (cit. on p. 24).

[36] A Ghanbary Sartang, M Ashnagar, E Habibi, and S Sadeghi. «Evaluation of Rating Scale Mental Effort (RSME) effectiveness for mental workload assessment in nurses». In: *Journal of Occupational Health and Epidemiology* 5.4 (2016), pp. 211–217 (cit. on p. 25).

[37] Mickaël Causse, Frédéric Dehais, Philippe-Olivier Faaland, and Fabrice Cauchard. «An analysis of mental workload and psychological stress in pilots during actual flight using heart rate and subjective measurements». In: *International Conference on Research in Air Transportation*. 2012 (cit. on p. 26).

[38] Da Tao, Haibo Tan, Hailiang Wang, Xu Zhang, Xingda Qu, and Tingru Zhang. «A Systematic Review of Physiological Measures of Mental Workload». In: *International Journal of Environmental Research and Public Health* 16.15 (2019). ISSN: 1660-4601. DOI: 10.3390/ijerph16152716. URL: https://www.mdpi.com/1660-4601/16/15/2716 (cit. on p. 26).

[39] Cornelia Setz, Bert Arnrich, Johannes Schumm, Roberto La Marca, Gerhard Tröster, and Ulrike Ehlert. «Discriminating stress from cognitive load using a wearable EDA device». In: *IEEE Transactions on information technology in biomedicine* 14.2 (2009), pp. 410–417 (cit. on p. 26).

[40] Haleh Aghajani, Marc Garbey, and Ahmet Omurtag. «Measuring mental workload with EEG+ fNIRS». In: *Frontiers in human neuroscience* 11 (2017), p. 359 (cit. on p. 26).

[41] Shabnam Samima and Monalisa Sarma. «EEG-based mental workload estimation». In: *2019 41st Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE. 2019, pp. 5605–5608 (cit. on p. 27).

[42] Chenglin Liu, Chenyang Zhang, Luohao Sun, Kun Liu, Haiyue Liu, Wenbing Zhu, and Chaozhe Jiang. «Detection of pilot's mental workload using a wireless EEG headset in airfield traffic pattern tasks». In: *Entropy* 25.7 (2023), p. 1035 (cit. on p. 27).

[43] Gerhard Marquart, Christopher Cabrall, and Joost De Winter. «Review of eye-related measures of drivers' mental workload». In: *Procedia Manufacturing* 3 (2015), pp. 2854–2861 (cit. on p. 27).

[44] Fares Al-Shargie, Tong Boon Tang, Nasreen Badruddin, and M. Kiguchi. «Mental Stress Quantification Using EEG Signals». In: Dec. 2016, pp. 15–19. ISBN: 978-981-10-0265-6. DOI: 10.1007/978-981-10-0266-3_4 (cit. on p. 28).

[45] Michael J Kane, Andrew RA Conway, Timothy K Miura, and Gregory JH Colflesh. «Working memory, attention control, and the N-back task: a question of construct validity.» In: *Journal of Experimental psychology: learning, memory, and cognition* 33.3 (2007), p. 615 (cit. on p. 29).

[46] Susanne M Jaeggi, Martin Buschkuehl, Walter J Perrig, and Beat Meier. «The concurrent validity of the N-back task as a working memory measure». In: *Memory* 18.4 (2010), pp. 394–412 (cit. on p. 29).

[47] Gary Gilmour et al. «Relating constructs of attention and working memory to social withdrawal in Alzheimer's disease and schizophrenia: issues regarding paradigm selection». In: *Neuroscience & Biobehavioral Reviews* 97 (2019), pp. 47–69 (cit. on p. 29).

[48] Yamira Santiago-Espada, Robert R. Myer, Kara A. Latorella, and James R. Comstock Jr. *The Multi-Attribute Task Battery II (MATB-II) Software for Human Performance and Workload Research: A User's Guide*. Tech. rep. NASA/TM–2011-217164. NASA Langley Research Center, 2011. URL: https://ntrs.nasa.gov/citations/20110014456 (cit. on p. 29).

[49] Jiapu Chai and Yan Li. «Examining mental workload based on multiple physiological signals: Review of the multi-attribute task battery (MATB) technique». In: *Medicine in Novel Technology and Devices* 24 (2024), p. 100340 (cit. on p. 29).

[50] A Bhatti et al. *CLARE: Cognitive Load Assessment in REaltime with Multimodal Data,(2024)* (cit. on pp. 30, 32).

[51] Miloš Pušica, Aneta Kartali, Luka Bojović, Ivan Gligorijević, Jelena Jovanović, Maria Chiara Leva, and Bogdan Mijović. «Mental workload classification and tasks detection in multitasking: Deep learning insights from EEG study». In: *Brain Sciences* 14.2 (2024), p. 149 (cit. on pp. 32, 79).

[52] *g.HIamp 256-Channel Biosignal Amplifier*. Product page on g.tec medical engineering website. Up to 256 channels, 24-bit resolution, CE certified and FDA cleared. Accessed 2025-10-29. 2025. URL: https://www.gtec.at/product/g-hiamp-256-channel-biosignal-amplifier/ (cit. on p. 37).

[53] *Kendall 24 mm Arbo ECG Electrodes H124SG – 500 pieces*. Product sheet available online. Ref. Code 31.1245.21; 24 mm round electrode, Ag/AgCl sensor, foam backing; www.medischevakhandel.nl item 1104037. Accessed 2025-10-29. 2008. URL: https://www.medischevakhandel.nl/en/kendall-24mm-arbo-ecg-electrodes-h124sg-500-pieces (cit. on p. 38).

[54] *biosignalsplux – Multi-sensor platform for raw biosignal acquisition*. Online. https://www.pluxbiosignals.com/pages/biosignalsplux?srsltid=AfmBOopXIwpzbrfPbJiSICXfGAI25R1pMjatztdLfeoHnicU1rppNQR5 (accessed Oct. 31, 2025). PLUX Biosignals, 2025 (cit. on p. 40).

[55] *g.GAMMAcap*. https://hooshmandfanavar.com/en/g-tec-medical-engineering-solutions/g-gammacap/. Accessed: 2025-10-30. Hooshmand Fanavar Tehran, 2025. URL: https://hooshmandfanavar.com/en/g-tec-medical-engineering-solutions/g-gammacap/ (cit. on p. 42).